ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

O. Univ.Prof. Dipl.-Ing. Dr. techn. Manfred Deistler

E119
Institut für Ökonometrie, Operations Research und Systemtheorie
1040 Wien, Argentinierstraße 8

eingereicht an der Technischen Universität Wien
Fakultät für Technische Naturwissenschaften und Informatik
von

Thomas Ribarits
Matr.Nr.: 9425026
Kollonitschtagasse 14/3/11
2700 Wiener Neustadt

Wien, am 30.10.2002
To whom it may concern ...
# Contents

Deutsche Kurzfassung ............................................... 11

Abstract ........................................................................ 13

Acknowledgements ...................................................... 15

1 Introduction ................................................................. 17
  1.1 General framework .................................................. 17
  1.2 Polynomial and rational matrices ............................... 20
  1.3 Spaces of transfer functions ...................................... 21
  1.4 Spaces of state-space realizations .............................. 22
  1.5 Spaces of ARMAX realizations ................................. 25

2 Parameter Estimation .................................................. 29
  2.1 The Kalman filter ...................................................... 29
    2.1.1 The time varying Kalman filter equations .............. 29
    2.1.2 Time invariance and stability of the Kalman filter .... 35
    2.1.3 Numerical simulations for the time varying Kalman filter 37
  2.2 Subspace estimation ................................................ 39
    2.2.1 Realization based methods: no exogenous inputs .... 43
    2.2.2 The state approach ............................................ 44
    2.2.3 MOESP-type algorithms .................................... 45
    2.2.4 Consistency .................................................... 46
    2.2.5 Asymptotic normality ........................................ 46
  2.3 Maximum likelihood estimation and related M-estimators .... 46
    2.3.1 Preliminaries .................................................. 46
    2.3.2 The finite sample likelihood function and its evaluation 48
    2.3.3 Separable least squares for approximate ML estimation 53
    2.3.4 Consistency .................................................. 56
    2.3.5 Asymptotic normality ........................................ 58
  2.4 Figures ....................................................................... 59
3 Manifolds in the space $S(n)$

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Non singular and orthogonal matrices</td>
<td>63</td>
</tr>
<tr>
<td>3.2</td>
<td>The manifold of observationally equivalent systems</td>
<td>64</td>
</tr>
<tr>
<td>3.3</td>
<td>The manifold of observationally equivalent MFN systems</td>
<td>67</td>
</tr>
<tr>
<td>3.4</td>
<td>The manifold of observationally equivalent BSA systems</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>The manifold of all BSA systems</td>
<td>76</td>
</tr>
<tr>
<td>3.6</td>
<td>The manifold of $L_2^c$-equivalent systems</td>
<td>77</td>
</tr>
<tr>
<td>3.7</td>
<td>Figures</td>
<td>81</td>
</tr>
</tbody>
</table>

4 Parametrizations of linear dynamic systems

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Full state-space parametrization</td>
<td>83</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Introduction</td>
<td>84</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Topological and geometrical properties</td>
<td>85</td>
</tr>
<tr>
<td>4.2</td>
<td>Echelon state-space forms</td>
<td>86</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Introduction</td>
<td>86</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Topological and geometrical properties</td>
<td>89</td>
</tr>
<tr>
<td>4.3</td>
<td>An overlapping state-space parametrization</td>
<td>90</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Introduction</td>
<td>90</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Topological and geometrical properties</td>
<td>92</td>
</tr>
<tr>
<td>4.4</td>
<td>Echelon ARMAX and reversed echelon ARMAX forms</td>
<td>93</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Introduction</td>
<td>93</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Topological and geometrical properties</td>
<td>98</td>
</tr>
<tr>
<td>4.5</td>
<td>Overlapping ARMAX parametrizations in $z^{-1}$ and $z$</td>
<td>98</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Introduction</td>
<td>98</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Topological and geometrical properties</td>
<td>100</td>
</tr>
<tr>
<td>4.6</td>
<td>Lyapunov balanced canonical form (Ober)</td>
<td>101</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Introduction</td>
<td>101</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Topological properties</td>
<td>111</td>
</tr>
<tr>
<td>4.7</td>
<td>Stochastically balanced canonical form (Ober)</td>
<td>113</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Introduction</td>
<td>113</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Topological properties</td>
<td>116</td>
</tr>
<tr>
<td>4.8</td>
<td>Minimum phase balanced canonical form (McGinnie)</td>
<td>117</td>
</tr>
<tr>
<td>4.8.1</td>
<td>Introduction</td>
<td>117</td>
</tr>
<tr>
<td>4.8.2</td>
<td>Topological properties</td>
<td>119</td>
</tr>
<tr>
<td>4.9</td>
<td>Data driven local coordinates (DDLC)</td>
<td>120</td>
</tr>
<tr>
<td>4.9.1</td>
<td>Data driven local parametrizations</td>
<td>120</td>
</tr>
<tr>
<td>4.9.2</td>
<td>Introduction</td>
<td>121</td>
</tr>
<tr>
<td>4.9.3</td>
<td>An illustrative example</td>
<td>121</td>
</tr>
<tr>
<td>4.9.4</td>
<td>Topological and geometrical properties</td>
<td>124</td>
</tr>
<tr>
<td>4.10</td>
<td>DDLC combined with separable least squares methods</td>
<td>130</td>
</tr>
<tr>
<td>4.10.1</td>
<td>Introduction</td>
<td>130</td>
</tr>
</tbody>
</table>
5 Numerical considerations

5.1 Preliminaries on extended prediction error filters ........................................... 147
5.2 The asymptotic likelihood \( L(\tau, \sigma) \) .................................................. 151
   5.2.1 Evaluating \( L(\tau, \sigma) \) for a particular parametrization ..................... 151
   5.2.2 Evaluating \( \nabla L(\tau, \sigma) \) for a particular parametrization .................. 151
   5.2.3 Evaluating an approximate Hessian of \( L(\tau, \sigma) \) for a particular parametrization ... 153
   5.2.4 Evaluations at the true parameter values ........................................... 155
5.3 The finite sample likelihood \( L_T(\tau, \sigma) \) ................................................ 156
   5.3.1 Evaluating \( L_T(\tau, \sigma), L_T^L(\tau) \) and \( L_T^{\sigma}(\tau^o) \) for a particular parametrization . 157
   5.3.2 Evaluating \( \nabla L_T(\tau, \sigma), \nabla L_T^L(\tau) \) and \( \nabla L_T^{\sigma}(\tau^o) \) for a particular parametrization ... 157
   5.3.3 Evaluating an approximate Hessian of \( L_T(\tau, \sigma), L_T^L(\tau) \) and \( L_T^{\sigma}(\tau^o) \) for a particular parametrization ... 159
5.4 Derivatives of state-space matrices .......................................................... 159
5.5 A few notes on optimization algorithms ....................................................... 164
5.6 Numerical considerations for various parametrizations ............................. 169
5.7 Future research ............................................................................................ 177
5.8 Figures ........................................................................................................ 178

A Appendix

A.1 Continuous time state-space systems ...................................................... 183
A.2 Spectral summands and spectral factors .................................................. 185
A.3 Balancing schemes ...................................................................................... 188
   A.3.1 Lyapunov balancing ................................................................. 189
   A.3.2 Positive real balancing ............................................................. 190
   A.3.3 Minimum phase balancing ........................................................ 191
   A.3.4 Stochastic balancing ................................................................. 192
   A.3.5 Summary of homeomorphisms relating balanced realizations ......... 194
A.4 Relation between discrete and continuous time state-space systems .......... 194
A.5 Partitioned Sylvester equations .................................................................. 196
A.6 Real analytic – and Riemannian manifolds .............................................. 196
A.7 Acronyms ..................................................................................................... 198
List of Figures

2.1 Input-output data and Kalman filter for stable and minimum phase models . . . . . . . 60
2.2 Kalman filter for stable and unstable non minimum phase models . . . . . . . . . . . . 61
3.1 Equivalence classes, MFN systems, BSA systems and $L^\infty_T$ equivalence classes . . . . 82
4.1 The boundary of the parameter space for Ober’s Lyapunov balanced canonical form . . 109
4.2 Canonical forms in the space of state-space systems . . . . . . . . . . . . . . . . . . . . . . 144
4.3 The DDLC construction in the space of state-space systems . . . . . . . . . . . . . . . . . . 145
4.4 The slsDDLC construction in the space of state-space systems . . . . . . . . . . . . . . . . 145
4.5 The orthoDDLC construction in the space of state-space systems . . . . . . . . . . . . . . 146
4.6 Image space of the mapping $F$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 146
5.1 Condition number of the Fisher information matrix for various parametrizations . . . 178
5.2 Graphical user interface for ML estimation using various state-space parametrizations . 179
5.3 Condition number and eigenvalues of the Fisher information matrix for DDLC . . . . . . 180
5.4 Gradient and Gauss-Newton search directions for the asymptotic likelihood function . . 181
Deutsche Kurzfassung


Die Dissertation ist wie folgt gegliedert:

In Kapitel 1 werden die wesentlichsten Begriffe aus der linearen Systemtheorie wiederrholt. Insbesondere werden drei mögliche Darstellungen von linearen dynamischen Systemen besprochen: Transferfunktionen, ARMAX-Modelle sowie Zustandsraum-Modelle.


Zuvor beschäftigt sich Kapitel 3 noch mit einer Reihe von speziellen Teilmengen im Raum aller Zustandsraum-Modelle der Form \((A,B,C,D)\). So wird zum Beispiel die Menge aller beobachtungäquivalenten Systeme (d.h. die Menge jener Zustandsraum-Modelle, welche die selbe Transferfunktion darstellen) ebenso analysiert wie die Menge aller balancierten Systeme \((A,B,C,D)\), welche sogenannte Allpass-Transferfunktionen repräsentieren. Für jede dieser Teilmengen im \((A,B,C,D)\)-Raum wird die Struktur einer differenzierbaren Untermannigfaltigkeit gezeigt und eine einfache Darstellung der Tangentialräume behandelt, was auch einige Einblicke in die Geometrie dieser speziellen Systemklassen erlaubt.


Der Anhang A beendet diese Arbeit und enthält naturgemäß Resultate und Definitionen, die an mehreren Stellen dieser Dissertation benötigt werden.

Diese Arbeit wurde finanziert durch den österreichischen Fonds zur Förderung der wissenschaftlichen Forschung, FWF, Projektnummer P 14438 sowie durch das TMR-Programm der Europäischen Union unter der Kontraktnummer ERB FMRX CT98 0206.
Abstract

This thesis is primarily concerned with the parametrization (and subsequent estimation) of linear, time invariant, finite dimensional dynamic systems. Hence, it may be assigned to the field of system identification which, during the last decades, has gradually become a subject of its own. The core question in system identification is how to obtain a good model from, in general, noisy data, and we indeed restrict ourselves to the model class of linear dynamic systems. Moreover, we will exclusively treat the case where the data is comprised of a number of vector valued quantitative observations which are ordered in time and measured with a fixed sampling rate.

The thesis is organized as follows:

In chapter 1, the most important concepts of linear system theory are repeated. In particular, we treat three different representations of linear dynamic systems: transfer functions, ARMAX models and state-space models.

Chapter 2 is concerned with different methods for the estimation of linear dynamic systems, both from a theoretical and practical point of view. An overview of different variants of so called subspace algorithms is given. These algorithms are used for the estimation of state-space models and do not require an explicit parametrization, but the corresponding estimates are in many cases inferior with respect to asymptotic properties to the estimates which are obtained by means of the maximum likelihood method. The latter estimation method is then discussed, and the practical evaluation of the likelihood function and its various approximations is treated. The Kalman filter is to be understood in this context, for it is not only an important tool for prediction but indeed also estimation and is therefore intensively discussed in a quite general setting. For the implementation of the maximum likelihood method, one has to use a certain parametrization for the model class considered. The topological and geometrical properties of the parametrization will then influence the statistical properties of the estimators as well as the numerical properties of the estimation algorithms, and this is the starting point for the investigation of different parametrizations which lies at the center of this thesis.

First, however, chapter 3 treats a number of special subsets in the space of all state-space systems of the form \((A, B, C, D)\). For instance, the set of observationally equivalent systems (i.e. the set of state-space systems corresponding to the same transfer function) is analyzed as well as the set of balanced systems \((A, B, C, D)\) representing so called stable allpass transfer functions. For each of these subsets in the \((A, B, C, D)\)-space the structure of a differentiable submanifold is shown and simple representations of the tangent spaces are given, yielding some insight into the geometry of these sets of systems.

The results of chapter 3 are used in the subsequent chapter 4 which contains the main part of this thesis and presents a number of different parametrizations of linear dynamic systems in a unified framework, comparing them with respect to topological and geometrical properties. Parametrizations for ARMAX models are treated as well as parametrizations for state-space models, the emphasis being on the state-space case. All of the parametrizations considered are for the case where subsets of the set of transfer functions of a given order \(n\) are described. The discussion includes the full state space parametrization, a number of canonical forms as well as overlapping parametrizations. Most of the results on topological properties of these parametrizations, which are summarized in each case in one theorem and discussed subsequently by means of illustrative examples, can be found at various places in the literature. The results on the class of so called data driven local parametrizations are new. The main idea of these approaches is to adapt the parametrization in the course of an iterative estimation algorithm in order to obtain desirable properties. This is motivated by the fact that the choice of the parametrization influences in particular the numerical properties of the identification algorithm, and these properties
depend on the true transfer function. But as the true transfer function is clearly unknown in advance, one has to adapt the parametrization around the current system estimate – note the local aspect at this point. Three different parametrizations of this class are treated: Data driven local coordinates (or, briefly, DDLC) as originally introduced in (McKelvey and Helmersson, 1999), are investigated in detail. Moreover, we suggest an extension of the DDLC concept leading to a new parametrization which we shall call \texttt{s1sDDLC} (Separable Least Squares combined with DDLC). In practice, an iterative search algorithm is most often used for the optimization of the likelihood function. Using the \texttt{s1sDDLC} parametrization slight modifications will yield a new class of estimation algorithms which we shall call \texttt{orthoDDLC}. This terminology is motivated by the fact that in each iteration step this approach guarantees that the parametrized system matrices \((A, B, C, D)\), if arranged appropriately, form an orthogonal matrix.

Finally, in chapter 5, issues are treated which are directly relevant for software implementations. In the first place, extended Kalman (or prediction error) filters are presented which are used for calculating the derivatives of the estimated innovations. Formulae for the evaluation of the asymptotic form of the likelihood function, its gradient and (approximate) Hessian are derived, where the evaluation is not necessarily at the parameter values corresponding to the true transfer function. The calculation of the analogous quantities in the finite sample case is also given. We continue with the computation of the derivatives of the state-space matrices with respect to the parameters for the parametrizations considered in this thesis and then present a short overview of some traditional optimization techniques. Finally, numerical properties of these algorithms are investigated in connection with the parametrizations considered. In particular, DDLC and \texttt{orthoDDLC} are compared in this respect.

The appendix A concludes this contribution and naturally contains results and definitions which are needed at various places throughout the thesis.

This work was funded by the Austrian science fund, FWF, project number P 14438 and the TMR program of the European Union under the contract number ERB FMRX CT98 0206.
Acknowledgements

Frankly speaking, I would have included a kind of traditional expression of thanks anyway. However, the nice thing is that I can really mean it without being too pathetic.

In the first place I want to thank my promotor Prof. Manfred Deistler. I had great support from him during the past years and took the pleasure to be confronted with his wide view on the scientific subject (and many other things). I cannot avoid to mention that travelling with him can lead to gorgeously bizarre situations, and this is truly stimulating my very personal sense of humor.

I started my PhD programme in Cambridge, being enrolled under the guidance of Dr. Jan M. Maciejowski. He brought me in touch with optimal control theory, a field of which I do not remember too much by now, I must confess. Nevertheless, he was (and hopefully still is) an excellent supervisor and, worth mentioning, a really nice person.

Dr. Bernard Hanzon was staying at our institute in Vienna for the first half of the year 2002, and it was during his stay that we confronted him with our work on data driven local coordinates. He had been working on balanced stable allpass systems recently, and thus actually came up with the idea of combining this field of research with what we had done until that point in time. This is how orthDDLC came into existence. Strange enough, I spent some time during his stay in Vienna in Brussels with Prof. Schoukens’ group at the Vrije Universiteit, and I hereby want to thank him too for granting me the total freedom of working on the PhD project without having to worry about anything else (because everything had actually been perfectly arranged).

As far as all the time in Vienna is concerned, it is almost unavoidable to mention Prof. Wolfgang Scherrer. In particular, he took the pain to do a lot of proofreading. Needless to say that his suggestions are always more than only generically useful!

Dr. Dietmar Bauer had to cope with my personal shortcomings for a long time, because we shared a room at the institute (and, strange enough, also at one or the other (conference) hotel – no worries, he is married). It seems that this at least did not keep him from answering my stupid questions.

Of course, there could be others to mention at this point. But, c’est la vie, I can only thank them as an anonymous crowd here. Immanuel Kant considered the human evil to be the price of freedom, and although I really appreciated the company of all (former) colleagues, some might consider this paragraph to be my private evil version (in case they ever come to read it, which will most probably not be the case anyway). This is, however, not intended and presumably just illustrates some inherent problems of a universal term like justice.

To come to private matters, I want to express my gratitude to my parents for their patience and love. I once read a nice dedication in a thesis stating that without this love, the thesis would never have come into existence (literally) …

Finally, let me deliberately chose my freedom to be a bit more serious again: I want to thank my family for granting me an incredible amount of support and freedom. Thanks also to my girlfriend who knows me for a relatively short time (at least compared to my parents), but nevertheless seems to understand me for a surprisingly large part. I hope that both are a bit proud that they indirectly also contributed to this thesis. Eventually, I want to mention my friends who will hopefully celebrate with me soon.
Chapter 1

Introduction

The main field to which this thesis may be assigned to is the field of system identification which, during the last decades, has gradually become a subject of its own. System identification deals with the question of how to find a good model from, in general, noisy data, i.e. with data driven modeling. Of course, this is a rather vague formulation, and we will try to make it more precise in section (1.1) below. In particular, we will explain what we mean by noisy data and what kind of models we take into account. A few facts about polynomial and rational matrices will be repeated in section (1.2), and the rest of this introductory chapter is concerned with different representations of linear dynamic systems: Section (1.3) deals with transfer functions, section (1.4) treats different classes of state-space realizations and section (1.5) is concerned with ARMAX realizations.

1.1 General framework

In the case considered in this thesis, the data will be comprised of observed vector valued sequences \( (y_t) \) and \( (u_t), t = 1, \ldots, T \). Note that the user is assumed to be able to a priori distinguish between two groups of observations: \( y_t \) is the \( s \)-dimensional observed output of a system and \( u_t \) denotes the \( m \)-dimensional measured external force, or input to the system.

Given a set of (feasible) data, the question arises which class of models, or systems, one wants to consider. In this thesis, the class of all a priori feasible candidate systems comprises the set of finite dimensional, linear (and dynamic), time invariant, causal, discrete-time stochastic systems; these terms will be explained in the sequel.

For us, a system is always given by a concrete representation: for example, in terms of an input-output, ARMAX or state-space representation as discussed below in sections (1.3), (1.4) and (1.5), respectively.

We start by briefly introducing the state-space representation of a finite dimensional, linear (and dynamic), time invariant, causal, discrete-time stochastic system which is given by

\[
\begin{align*}
x_{t+1} &= Ax_t + Bu_t + K\varepsilon_t \\
y_t &=Cx_t + Du_t + \varepsilon_t
\end{align*}
\]

Here, \( x_t \) is the \( n \)-dimensional state vector which is not directly observed in general, and \( A \in \mathbb{R}^{n \times n} \), \( B \in \mathbb{R}^{n \times m} \), \( C \in \mathbb{R}^{s \times n} \), \( D \in \mathbb{R}^{s \times m} \) and \( K \in \mathbb{R}^{n \times s} \) are parameter matrices; as mentioned above, \( y_t \) and \( u_t \) are the observed outputs and exogenous inputs, respectively. In addition to the inputs \( u_t \), we consider other unobserved influences causing variation in \( y_t \). These influences may be due to measurement errors (making it impossible to determine the output exactly) or simply due to the fact that certain aspects of the output sequence cannot be described accurately by the model in hand. In any case, these unobserved additional influences are represented by the stochastic process \( (\varepsilon_t) \) which is assumed to be a white noise process, i.e. \( E\varepsilon_t = 0 \) and \( E\varepsilon_s\varepsilon_t' = \delta_{s,t} \Sigma \) for all \( s, t \in \mathbb{Z} \), where \( \Sigma > 0 \). We also assume that \( E u_s \varepsilon'_t = 0 \) for
all \( s, t \in \mathbb{Z} \), which is trivially satisfied if \((u_t)\) is treated as being a deterministic sequence, for instance. More precise assumptions on \((u_t)\) and \((\varepsilon_t)\) can be found in section (2.3.4).

For obvious reasons, the first equation in (1.1) is sometimes called the (state) transition equation, and the second equation is called the measurement equation.

Most often, the stability property

\[
|\lambda_{\text{max}}(A)| < 1 \tag{1.2}
\]

and the minimum phase property

\[
|\lambda_{\text{max}}(A - KC)| \leq 1 \tag{1.3}
\]

will be imposed. Here, \(\lambda_{\text{max}}(X)\) denotes the eigenvalue(s) of the square matrix \(X\) of maximum modulus. If the inequality in (1.3) is strict, we speak of the strict minimum phase property.

State-space systems satisfying (1.2) or (1.3) are called stable or minimum phase state-space systems. If both properties are satisfied, the state-space representation is said to be in innovations (or prediction error) form. In this case, \((\varepsilon_t)\) is not only white noise, but in addition the prediction error of the best, linear, least squares prediction of \(y_t\) given \(y_s, s < t\) and \(u_t, t \in \mathbb{Z}\), i.e. \(\varepsilon_t = y_t - y_{t-1}\). The quantities \(\varepsilon_t\) are then also called (linear) innovations; see chapter 1 in (Hannan and Deistler, 1988).

Using the symbol \(z\) to denote the backward shift operator on \(\mathbb{Z}\) (as will be done throughout this thesis), i.e. \(z(x_t|t \in \mathbb{Z}) = (x_{t-1}|t \in \mathbb{Z})\), the first equation of (1.1) can be written as

\[
(I - zA)x_t = zBu_t + zK\varepsilon_t \tag{1.4}
\]

Evidently, (1.4) is a vector difference equation (VDE). A solution to (1.4) is a process \((x_t|t \in \mathbb{Z})\) satisfying (1.4) for given \((u_t|t \in \mathbb{Z}), (\varepsilon_t|t \in \mathbb{Z})\) and given \(A, B\) and \(K\). As is well known, solutions always exist and the set of solutions is in fact given as the sum of one particular solution and the set of all solutions to the so-called homogeneous equation \((I - zA)x_t = 0\). Imposing the stability condition (1.2), one solution in input-output representation is given by

\[
x_t = (I - zA)^{-1}zBu_t + (I - zA)^{-1}zK\varepsilon_t \\
= \sum_{j=1}^{\infty} A^{j-1}Bu_{t-j} + \sum_{j=1}^{\infty} A^{j-1}K\varepsilon_{t-j} \tag{1.5}
\]

if \((u_t)\) and \((\varepsilon_t)\) are such that the infinite sums, which are to be understood as limits in the mean square sense, exist. Under the stability assumption, this can easily be guaranteed for any stationary process with finite second moments, implying that \(\sum_{j=1}^{\infty} A^{j-1}K\varepsilon_{t-j}\) is well defined and even stationary. As far as the exogenous input is concerned, we may, e.g., assume that \((u_t)\) is deterministic, satisfies \(u_s = 0\) for all \(s \leq 0\) and \(\sup_{t \in \mathbb{Z}} ||u_t|| = M < \infty\), where \(||.||\) denotes an arbitrary vector norm. Then, clearly, \(\sum_{j=1}^{\infty} A^{j-1}Bu_{t-j}\) exists and represents the mean of \(x_t\). Hence, the corresponding solution to (1.1) is given by

\[
y_t = \sum_{j=1}^{t-1} CA^{j-1}Bu_{t-j} + Du_t + \sum_{j=1}^{\infty} CA^{j-1}K\varepsilon_{t-j} + \varepsilon_t \tag{1.6}
\]

Clearly, (1.6) is a special form of a linear system in input-output representation

\[
y_t = \sum_{j=-\infty}^{\infty} L_j u_{t-j} + \sum_{j=-\infty}^{\infty} K_j \varepsilon_{t-j} \tag{1.7}
\]
It is worth noting that under the stability assumption (1.2) every solution to the homogeneous VDE $(I - zA)x_t = 0$ satisfies $x_t \rightarrow 0$ for $t \rightarrow \infty$, and, therefore, every solution to the VDE (1.4) converges to (1.5) for $t \rightarrow \infty$, implying of course, that every solution to (1.1) converges to (1.6) for $t \rightarrow \infty$. This is the reason for calling (1.6) the steady state solution. It is the unique stable solution to (1.1), where stability really means that if the input (trajectory) $((u'_t, \varepsilon'_t)^t | t \in \mathbb{Z})$ is bounded, then the output (trajectory) $((y_t | t \in \mathbb{Z})$ is also bounded. In particular, if the input process is stationary, so will be the output process.

It is obvious that the input-output representation of the stable solution in (1.6) is uniquely determined by the impulse response (or the sequence of so called Markov parameters) corresponding to the state-space representation in (1.1). The impulse response is given by

$$(L_j, K_j)_{j \in \mathbb{N}} = ((D, I), C(B, K), CA(B, K), CA^2(B, K), \ldots)$$

Under the stability assumption (1.2) and by setting $(L_j, K_j) = 0$ for $j < 0$, this sequence can be embedded into the Hilbert space $l_2((\mathbb{R}^{m+s})^\mathbb{Z})$, the space of square summable matrix sequences which is often called the time domain. The corresponding inner product is given by $(f_j, g_j) = \sum_{j=-\infty}^{\infty} \text{tr}(f_j g_j^*)$. The corresponding transfer function (from $(u'_t, \varepsilon'_t)^t$ to $y_t$) is given by

$$(l(z), k(z)) = \sum_{j=1}^{\infty} C A^{j-1}(B, K) z^j + (D, I)$$

$$= C(z^{-1} I - A)^{-1}(B, K) + (D, I)$$

$$=: \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

(1.8)

Here, $z = e^{j\lambda}$ denotes a complex variable and the last expression is just a convenient shorthand notation. Under the stability assumption, this transfer function is an element of the Hilbert space $L_2^2((\mathbb{Z}^{m+s})) [0, 2\pi]$, the space of complex, matrix valued and square integrable Lebesgue measurable functions on the unit circle which is often called the frequency domain. The corresponding inner product is given by $(f(z), g(z)) = \frac{1}{2\pi} \int_0^{2\pi} \text{tr}(f(e^{i\lambda})) g(e^{-i\lambda})^* d\lambda$.

We are now ready to clarify a number of terms used in the discussion above: In discrete time systems, the time index is running over the integers $\mathbb{Z}$ or, if explicitly stated, over the natural numbers $\mathbb{N}$. Systems are called linear (dynamic) if the current output depends linearly on present (and past and future) inputs $u_{t-j}$ and $\varepsilon_{t-j}$, $j \in \mathbb{Z}$; see (1.7). The system is said to be time invariant if the coefficients $(L_j, K_j)$ in the input-output representation in (1.7) do not depend on $t$, and it is called causal if the output $y_t$ depends only on present and past inputs $u_k$ and $\varepsilon_k$, $k \leq t$. For the input-output representation in (1.7), this is equivalent to $(L_j, K_j) = 0$ for $j < 0$ as it is the case in (1.6). Finally, a system is called finite dimensional if the corresponding transfer function is a rational function of the complex variable $z$; this is clearly the case in (1.8).

We conclude this introductory section by computing the covariance function of the state and the output process corresponding to the state-space representation (1.1). Note that the means of the state and output are given by

$$\mathbb{E}x_t = \sum_{j=1}^{t-1} A^{j-1} Bu_{t-j}$$

and that these quantities only depend on the transfer function $l(z)$. Clearly, then

$$x_t - \mathbb{E}x_t = \sum_{j=1}^{\infty} A^{j-1} K \varepsilon_{t-j}$$

and

$$y_t - \mathbb{E}y_t = \sum_{j=1}^{\infty} CA^{j-1} K \varepsilon_{t-j} + \varepsilon_t$$
depend solely on \( k(z) \) and \( \Sigma \), implying that the corresponding covariance functions

\[
\gamma_x(k) = \mathbb{E}(x_{t+k} - \mathbb{E}x_{t+k})(x_t - \mathbb{E}x_t)' \quad \gamma(k) = \mathbb{E}(y_{t+k} - \mathbb{E}y_{t+k})(y_t - \mathbb{E}y_t)'
\]

are also only depending on \( k(z) \) and \( \Sigma \). As limits in the mean square sense and expectations are interchangeable, we get

\[
\gamma_x(k) = \sum_{i,j=1}^{\infty} A^{j-1}_i K \mathbb{E} \varepsilon_{t+k-j} \varepsilon_{t-i} K' (A^{i-1})' \\
= \sum_{j=k+1}^{\infty} A^{j-1}_i K \Sigma K' (A^{j-k-1})' \quad k \geq 0 \quad (1.9)
\]

\[
\mathbb{E}(x_{t+k} - \mathbb{E}x_{t+k}) \varepsilon_t' = \sum_{j=1}^{\infty} A^{j-1}_i K \mathbb{E} \varepsilon_{t+k-j} \varepsilon_t' \\
= A^{k-1}_i K \Sigma, \quad k \geq 0 \quad (1.10)
\]

Setting the unconditional state covariance \( \gamma_x(0) = P \), it follows from (1.9) with \( k = 0 \) that \( P \) satisfies the discrete time Lyapunov equation

\[
P = APA' + K \Sigma K' \quad (1.11)
\]

the solution of which is unique under the stability condition; in fact, it is unique under a less stringent condition. The autocovariance function of \( (y_t) \) is given by

\[
\gamma(0) = CPC' + \Sigma \quad (1.12)
\]

and

\[
\gamma(k) = C \gamma_x(k)' C' + C \mathbb{E}(x_{t+k} - \mathbb{E}x_{t+k}) \varepsilon_t' \\
= C \sum_{j=k+1}^{\infty} A^{j-1}_i K \Sigma K' (A^{j-k-1})' C' + CA^{k-1}_i K \Sigma \\
= CA^{k-1}_i \sum_{j=k+1}^{\infty} A^{j-k-1}_i K \Sigma K' (A^{j-k-1})' C' + CA^{k-1}_i K \Sigma \\
= CA^{k-1}_i \left[ \frac{APC' + K \Sigma}{M} \right] \quad k > 0 \quad (1.13)
\]

1.2 Polynomial and rational matrices

This section presents a few facts about polynomial and rational matrices which will be needed frequently.

A square polynomial matrix \( u(z) = \sum_{j=0}^{p} U(j) z^j \), \( U(j) \in \mathbb{R}^{n \times n} \), is called unimodular if it is invertible and its inverse \( u^{-1}(z) \) is also a polynomial matrix. It is easy to see that a square polynomial matrix is unimodular if and only if its determinant is a nonzero constant.

The (normal) rank of a polynomial matrix \( p(z) \) is the maximally possible rank of \( p(z) \) for at least one (and thus all but a finite number of) \( z \in \mathbb{C} \). The (normal) rank of a rational matrix \( h(z) \) can be defined analogously.

The following theorem is of great importance:
1.3. SPACES OF TRANSFER FUNCTIONS

Theorem 1.2.1 (Smith McMillan form). Every \( p \times q \) rational matrix \( h(z) \) of rank \( r \) can be written as

\[
h(z) = u(z) \Lambda(z) v(z) \tag{1.14}
\]

where \( u(z) \) and \( v(z) \) are unimodular \( p \times p \) and \( q \times q \) polynomial matrices, respectively, and \( \Lambda(z) \) is diagonal of the form

\[
\Lambda(z) = \begin{pmatrix}
\varepsilon_1(z) & \cdot & \cdot & \cdot \\
\cdot & \ddots & \cdot & \cdot \\
\cdot & \cdot & \varepsilon_{i+1}(z) & \cdot \\
\cdot & \cdot & \cdot & \cdot
\end{pmatrix}
\]

where \( \varepsilon_i(z) \) and \( \psi_i(z) \) are relatively prime monic polynomials, \( \varepsilon_i(z) \) divides \( \varepsilon_{i+1}(z) \) and \( \psi_i(z) \) divides \( \psi_{i+1}(z) \) for \( i = 1, \ldots, r - 1 \). \( \Lambda(z) \) is called the Smith McMillan form of \( h(z) \) and is unique for given \( h(z) \), whereas \( u(z) \) and \( v(z) \) are not unique in general.

Proof. This theorem can be found in any standard textbook of linear system theory; see, e.g. section 6.3.2 in (Kailath, 1980) or lemma 2.4.3 in (Hannan and Deistler, 1988).

There are several different ways to define poles and zeros for a rational \( p \times q \) matrix \( h(z) \). Following (Kailath, 1980), we give the following

Definition 1.2.1 (Poles and Zeros of a rational matrix). The poles of a rational \( p \times q \) matrix \( h(z) \) are the zeros of the denominator polynomials \( \psi_i(z) \) in the Smith McMillan form of \( h(z) \). The zeros of \( h(z) \) are the zeros of the (nonzero) numerator polynomials \( \varepsilon_i(z) \) in the Smith McMillan form of \( h(z) \).

Note that \( h(z) \) may have a pole and a zero at the same \( z \).

1.3 Spaces of transfer functions

Let us start with the following

Definition 1.3.1 (Causality). A rational \( s \times (m + s) \) discrete time transfer function \((l(z), k(z))\) is called causal if it has a power series expansion in a certain neighborhood of zero, i.e. if \( z = 0 \) is no pole.

The terms McMillan degree, degree and order will be used interchangeably. Note that a transfer function \((\bar{l}(z), \bar{k}(z))\) is strictly proper if \( \lim_{|z| \to \infty}(\bar{l}(z), \bar{k}(z)) = (0, 0) \):

Definition 1.3.2 (McMillan degree of a discrete time transfer function). A rational and causal \( s \times (m + s) \) discrete time transfer function \((l, k)\) is said to be of McMillan degree or order \( n \) if the strictly proper transfer function \((\bar{l}(z), \bar{k}(z)) = (l(z^{-1}), k(z^{-1})) - (l(0), k(0))\) has McMillan degree \( n \), i.e. the sum of the denominator degrees in the Smith McMillan form of \((\bar{l}(z), \bar{k}(z))\) is equal to \( n \).

Remark 1.3.1. Note that one could also define the order of \((l, k)\) in a direct way by considering the Smith McMillan form of \((l, k)\). A scalar (MA) - transfer function \( k(z) = 1 + b z = 1 \cdot (1 + b z) \cdot 1 \) would then have order zero. However, according to definition (1.3.2), we get \( \bar{k}(z) = b z^{-1} = 1 \cdot z^{-1} \cdot b \) and therefore \( k(z) \) has order one.

Let \( U_A \) be the set of all rational and causal \( s \times (m + s) \) transfer functions of the form

\[
(l(z), k(z)) = \sum_{j=0}^{\infty} (L_j, K_j) z^j \quad \text{where} \quad (l(0), k(0)) = (L_0, K_0) = (D, I) \tag{1.15}
\]
for an arbitrary $D \in \mathbb{R}^{s \times m}$. Clearly, we can identify the causal transfer function $(l, k)$ with the sequence $((L_j, K_j) | j \in \mathbb{N})$ of its power series coefficients. Clearly, $((L_j, K_j) | j \in \mathbb{N}) \in (\mathbb{R}^{s \times (m+s)})^\mathbb{N}$. Unless explicitly stated otherwise, let us endow the latter space with the product topology of the Euclidean space $\mathbb{R}^{s \times (m+s)}$. The set $U_A$ is endowed with the corresponding relative topology which is called the pointwise topology, $T_{pt}$. $U_A$ is "infinite dimensional" and may be broken into finite dimensional bits $U_\alpha$, $\alpha \in I$, say. Usually these bits are described by a subset of a Euclidean space (to be more precise, by the set of free parameters). We will consider the case where the sets $U_\alpha$ are subsets of the class $\mathcal{M}(n)$ of rational and causal $s \times (m+s)$ transfer functions of fixed McMillan degree $n$. Note that

$$\mathcal{M}(n) \subset U_A = \bigcup_{i \in \mathbb{N}} \mathcal{M}(i)$$

The subset of $\mathcal{M}(n)$ corresponding to stable transfer functions is denoted by $\mathcal{M}_s(n)$:

**Definition 1.3.3 (Stability).** A rational and causal $s \times (m+s)$ discrete time transfer function $(l(z), k(z))$ is said to be **stable** if all the poles of $(l(z), k(z))$ are outside the closed unit disk in the complex plane.

**Remark 1.3.2.** Note that the poles of $(l, k)$ are not defined as the poles of $(\hat{l}, \hat{k})$ as it is the case for the definition of the order.

Finally, the subset of $\mathcal{M}_s(n)$, where additionally the square transfer function $k(z)$ is (strictly) minimum phase, is denoted by $\mathcal{M}_{smp}(n)$:

**Definition 1.3.4 (Minimum Phase Property).** A rational and causal $s \times s$ discrete time transfer function $k(z)$ is said to be **minimum phase** if all the zeros of $k(z)$ are outside the open unit disk in the complex plane. It is said to be **strictly minimum phase** if $k(z)^{-1}$ is stable.

**Remark 1.3.3.** Note that the zeros of $k$ are not defined as the zeros of $\hat{k}$ as it is the case for the definition of the order.

In the rest of this thesis we will treat the case where the original model class $U_A$ is broken into the finite dimensional subsets $\mathcal{M}(n)$, whereafter pieces of $\mathcal{M}(n)$ are considered and have to be parametrized:

$$U_\alpha \subseteq \mathcal{M}(n) \subset U_A$$

However, it is important to note that there are many parametrizations for rational transfer functions that do not commence from further breaking down $\mathcal{M}(n)$ into pieces $U_\alpha$ but rather consider a different finite dimensional piece of $U_A$ in the first step: Examples are ARMAX parametrizations\(^1\) according to the column degrees of $(a(z), d(z), b(z))$ or parametrizations of ARMAX systems using the coefficients of $c(z)$ and $p(z)$ where $(l(z), k(z)) = c^{-1}(z)p(z)$ and $c(z)$ is the least common multiple of the denominator polynomials of the entries $(l_{ij}(z), k_{ij}(z))$ of $(l(z), k(z))$ and the degrees of $c(z)$ and $p(z)$ are prescribed. Another important case is when the model class is highly structured by a priori information coming e.g. from physical or economic theory.

For an overview of the main classes $U_A, \mathcal{M}(n), \mathcal{M}_s(n), \mathcal{M}_{smp}(n)$ and $\mathcal{M}_{smp}(n)$ of (discrete time) transfer functions, see the appendix (A.7). Note that this notation also includes the case where $m = 0$ and $s$ and $n$ are arbitrary, i.e. the case without exogenous inputs. In this setting, we only consider the $s \times s$ transfer function $k(z)$ with $k(0) = I$. Of course, we could also deal with transfer functions corresponding to output error models, i.e. the case where only $\hat{l}(z)$ is considered and $k(z)$ reduces to the identity matrix.

### 1.4 Spaces of state-space realizations

The following lemma is a standard result:

**Lemma 1.4.1.** Every transfer function $(l, k) \in U_A$ has a state-space realization (1.1), i.e. can be represented by

\(^1\)For a brief treatment of ARMAX systems see section (1.5) below.
\[ (l(z), k(z)) = C(z^{-1}I - A)^{-1}(B, K) + (D, I) = \begin{bmatrix} A & B \\ C & D & K \end{bmatrix} \]

where \( A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{s \times n}, D \in \mathbb{R}^{s \times m} \) and \( K \in \mathbb{R}^{n \times s} \) for some \( n \in \mathbb{N} \).

**Proof.** See, e.g. the proof of lemma 1.2.1 in (Hannan and Deistler, 1988).

In the sequel we consider state-space systems of the form (1.1) which will be abbreviated by using the matrix quintuple \((A, B, C, D, K)\). The set of all state-space systems \((A, B, C, D, K)\) for fixed \( m \) and \( s \), but variable \( n \) is denoted by \( S_A \). If \( n \) is fixed too, we denote the corresponding set of state-space systems \((A, B, C, D, K)\) by \( S(n) \subset S_A \). Of course, one can also embed the matrices \((A, B, C, D, K)\) in \( \mathbb{R}^{n^2 + 2ns + m[n + s]} \) and identify \( S(n) \) with \( \mathbb{R}^{n^2 + 2ns + m[n + s]} \). We always identify \((A, B, C, D, K)\) with

\[
\begin{pmatrix}
\text{vec}(A) \\
\text{vec}(\bar{B}) \\
\text{vec}(C) \\
\text{vec}(D)
\end{pmatrix}
\]

where \( \bar{B} = (B, K) \) throughout this thesis and \( \text{vec}(X) \) stacks the first, second, etc. column of the matrix \( X \) on top of each other. We will therefore also use the same symbol for a mapping that is defined for matrix tuples \((A, B, C, D, K)\) or the corresponding vectorization of these matrices. Note that for \( m = 0 \), i.e. in the case where no exogenous inputs are present and the state-space system is given by \((A, K, C, I)\), the embedding described above simplifies to \( \text{vec}(A)’, \text{vec}(K)’, \text{vec}(C)’ \).

\( S(n) \) is endowed with the Euclidean norm for \((A, B, C, D, K)\).

Finally, we can introduce the set

\[
S_m(n) = \{(A, B, C, D, K) \in S(n)|\text{(A, B, C, D, K) is minimal}\}
\]

This necessitates the following definition:

**Definition 1.4.1 (Minimality).** A state-space realization \((A, B, C, D, K)\) \( \in S(n) \) of a transfer function \((l, k) = C(z^{-1}I - A)^{-1}(B, K) + (D, I) \) is called minimal if for all \((\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K}) \in S(\bar{n})\) with \( \bar{C}(z^{-1}I - \bar{A})^{-1}(\bar{B}, \bar{K}) + (\bar{D}, I) = (l, k) \) we have: \( n \leq \bar{n} \).

**Remark 1.4.1.** The state dimension \( n \) of a minimal state-space realization \((A, B, C, D, K)\) of \((l, k)\) can be shown to coincide with the order of the transfer function \((l, k)\) as given in definition (1.3.2) above.

By what is known as **Kalman canonical decomposition** in linear system theory, one can show that lemma (1.4.1) can be strengthened to

**Lemma 1.4.2.** Every transfer function \((l, k) \in U_A\) has a minimal state-space realization \((A, B, C, D, K)\).

**Proof.** For a detailed proof, see, e.g. the proof of lemma 2.3.1 in (Hannan and Deistler, 1988).

Other important concepts are given in the next definition:

**Definition 1.4.2 (Controllability and Observability).** Let \((A, B, C, D, K) \in S(n)\) be a state-space realization of a transfer function \((l, k)\). The matrices

\[
\mathcal{C}_n = \begin{pmatrix} \bar{B}, AB, A^2 \bar{B}, \ldots, A^{n-1} \bar{B} \end{pmatrix} \in \mathbb{R}^{n \times (n \times m + s)}
\]

\[
\mathcal{O}_n = \begin{pmatrix} (C, CA), (CA^2), \ldots, (CA^{n-1}) \end{pmatrix} \in \mathbb{R}^{n \times m}
\]

are called **controllability matrix** and **observability matrix**, respectively. The pair \((A, \bar{B})\) is called **controllable** if \( \mathcal{C}_n \) is of full row rank. The pair \((C, A)\) is called **observable** if \( \mathcal{O}_n \) is of full column rank.
Remark 1.4.2. A state-space realization is minimal if and only if \((C,A)\) is observable and \((A,B)\) is controllable.

Remark 1.4.3. By the Cayley-Hamilton theorem, the rank of \(C_n\) and \(O_n\) cannot be increased by adding terms of the form \(A^j B\) or \(C A^j\) for \(j \geq n\), i.e. the (infinite) controllability matrix \(C\) and the (infinite) observability matrix \(O\) which are given by

\[
C = (\tilde{B}, AB, A^2 B, \ldots) \in \mathbb{R}^{n \times \infty}
\]
\[
O = (C', (CA')_1, (CA^2')_1, \ldots)' \in \mathbb{R}^{\infty \times n}
\]

have the same rank as \(C_n\) and \(O_n\).

The product of the (infinite) controllability matrix \(C\) and the (infinite) observability matrix \(O\) is a (block) Hankel matrix and is called the Hankel matrix corresponding to the transfer function \((l, k)\):

\[
H = OC = \begin{pmatrix}
C \tilde{B} & C \beta B & CA \beta B & \cdots \\
C \beta B & C \beta ^2 B & CA \beta ^2 B & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix} = \begin{pmatrix}
(L_1, K_1) & (L_2, K_2) & (L_3, K_3) & \cdots \\
(L_2, K_2) & (L_3, K_3) & (L_4, K_4) & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\]

(1.16)

Remark 1.4.4. Clearly, from remark (1.4.3), it follows that the rank of \(H\) equals the rank of \(H_n^n\) where \(H_n^n\) is the northwest corner of \(H\) containing the first \(n\) block rows and the first \(n\) block columns.

Remark 1.4.5. It can be shown that the rank of the Hankel matrix \(H\) is equal to the order of the corresponding transfer function \((l, k)\).

Definition 1.4.3 (Controllability and Observability Gramian). Let \((A, B, C, D, K)\) \(\in S(n)\) be given and let \(A\) satisfy the stability condition (1.2). The matrices \(P = CC' = \sum_{j=0}^{\infty} A^j B B' A'^j\) and \(Q = O'O = \sum_{j=0}^{\infty} A^j C' C A'^j\) are called controllability Gramian and observability Gramian, respectively.

Remark 1.4.6. It follows immediately from definition (1.4.2) that \(P > 0\) if and only if \((A, \tilde{B})\) is controllable and that \(Q > 0\) if and only if \((C, A)\) is observable.

Finally, let us introduce the mapping \(\pi\) attaching transfer functions to the system matrices of a state-space representation:

\[
\pi: \quad S_{A} (A, B, C, D, K) \rightarrow \quad C (z^{-1} I - A)^{-1} (B, K) + (D, I)
\]

Often we will restrict the domain of definition to \(S(n)\) for some (arbitrary) \(n \in \mathbb{N}\) and use the same symbol without making this restriction explicit. It is evident that such a mapping \(\pi\) is continuous; recall that \(U_A\) is endowed with the pointwise topology \(T_p\).

The following definition is of central importance:

Definition 1.4.4 (The \((l, k)\)-Equivalence Class \(E(A, \tilde{B}, C, D)\)). Let \((l, k) \in \mathbb{M}(n)\) be given and let \((A, B, C, D, K)\) be a minimal state-space realization of \((l, k)\). The set \(E(A, \tilde{B}, C, D) = \pi^{-1}(l, k) \cap S(n)\) is called the class of observationally equivalent minimal systems \(A\), briefly, the \((l, k)\)-equivalence class. Every point in the \((l, k)\)-equivalence class corresponds to a point in \(S_m(n)\). Two minimal state-space systems \((A, B, C, D, K)\) and \((\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K})\) in \(E(A, B, C, D)\) are called observationally equivalent, which is denoted by \((A, B, C, D, K) \sim (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K})\).

Remark 1.4.7. The structure of the set \(E(A, \tilde{B}, C, D)\) will be clarified in lemma (3.2.1) in chapter 3.

Remark 1.4.8. Clearly, one need not restrict the definition of observational equivalence to \(S_m(n)\). A very general possibility would be to consider the whole set \(S_{A}\) and call the set \(\pi^{-1}(l, k)\) an \((l, k)\)-equivalence class. This, of course, means that non minimal state-space systems are included in \(\pi^{-1}(l, k)\). Assume that the transfer function \((l, k)\) is stable. Then, all state-space systems in \(\pi^{-1}(l, k)\) will correspond to
one stable solution as has been discussed in section (1.1) above. However, this solution will not be the steady state solution to all systems in $\pi^{-1}(l, k)$. In fact, there will be state-space systems in $\pi^{-1}(l, k)$ for which not all solutions converge to the stable solution for $t \to \infty$; some of the solutions may even diverge to infinity.

Finally, canonical forms can be defined as follows:

**Definition 1.4.5 (Canonical Form for Minimal State-Space Systems).** Let $S_m(n)$ be given. A **canonical form** (for system equivalence) is a map $\Gamma : S_m(n) \to S_m(n)$ such that

$$
\Gamma(A, B, C, D, K) \sim (A, B, C, D, K) \quad \forall (A, B, C, D, K) \in S_m(n)
$$

$$(A, B, C, D, K) \sim (\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K}) \quad \Rightarrow \quad \Gamma(A, B, C, D, K) = \Gamma(\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K})$$

if $(A, B, C, D, K)$ and $(\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K})$ are elements of $S_m(n)$.

**Remark 1.4.9.** It is obvious that canonical forms pick a unique representative out of each $(l, k)$-equivalence class $E(A, B, C, D)$.

**Remark 1.4.10.** Of course, the same definition could also be considered for subsets of $S_m(n)$, e.g. one could only be interested in canonical forms for state-space systems corresponding to transfer functions in $M_0(n)$ or $M_{amp}(n)$. Such canonical forms will be treated in sections (4.6), (4.7) and (4.8) in chapter 4.

**Remark 1.4.11.** On the other hand, one could also use larger sets for the definition of a canonical form $\Gamma$. A very general possibility would be to consider $\Gamma : S_A \to S_A$ selecting from every equivalence class $\pi^{-1}(l, k)$ a unique representative.

### 1.5 Spaces of ARMAX realizations

In section (1.1) we have encountered the state-space and input-output representations of linear systems. Here, we will briefly discuss the so called **ARMAX representation**:

$$
\sum_{j=0}^{p} A(j)y_{t-j} = \sum_{j=0}^{r} D(j)u_{t-j} + \sum_{j=0}^{q} B(j)\varepsilon_{t-j} 
$$

(1.17)

Here, $y_t$, $u_t$ and $\varepsilon_t$ again denote the outputs, the observed inputs and the stochastic errors and are assumed to have the same properties as discussed in section (1.1) above. In particular, $E u_t \varepsilon_t = 0$ for all $s, t \in \mathbb{Z}$ is part of the definition of an ARMAX system. The matrices $A(j) \in \mathbb{R}^{s \times s}$, $D(j) \in \mathbb{R}^{s \times m}$ and $B(j) \in \mathbb{R}^{s \times s}$ are the associated parameter matrices. Again using the symbol $z$ to denote the backward shift operator on $\mathbb{Z}$, the VDE (1.17) can be written as

$$
a(z)y_t = d(z)u_t + b(z)\varepsilon_t
$$

(1.18)

where $a(z) = \sum_{j=0}^{p} A(j)z^j$, $d(z) = \sum_{j=0}^{r} D(j)z^j$ and $b(z) = \sum_{j=0}^{q} B(j)z^j$ are the associated polynomial matrices. The assumption $\det a(z) \neq 0$ is always imposed. Hence, we can alternatively define an ARMAX system to be a triple $(a(z), d(z), b(z))$ where $a(z)$, $d(z)$ and $b(z)$ are $s \times s$, $s \times m$ and $s \times s$ polynomial matrices, respectively, and where $\det a(z) \neq 0$.

Considering the VDE in (1.18), the stability condition becomes

$$
\det a(z) \neq 0 \quad \text{for } |z| \leq 1
$$

(1.19)

In analogy to the discussion below (1.4), one solution to (1.18) is given by

$$
\sum_{j=0}^{p} A(j)y_{t+j} = \sum_{j=0}^{r} D(j)u_{t+j} + \sum_{j=0}^{q} B(j)\varepsilon_{t+j}
$$

$^2$An alternative description is in terms of the forward shift $z^{-1}$, i.e. $a(z^{-1})y_t = d(z^{-1})u_t + b(z^{-1})\varepsilon_t$ corresponding to $\sum_{j=0}^{p} A(j)y_{t+j} = \sum_{j=0}^{r} D(j)u_{t+j} + \sum_{j=0}^{q} B(j)\varepsilon_{t+j}$. 


where the infinite sums are well defined under the same conditions as discussed below (1.5); (1.20) is the (unique) steady state solution to (1.18).

The following lemma is a standard result:

**Lemma 1.5.1.** Every transfer function \((l, k) \in U_A\) has an ARMAX realization \((a(z), d(z), b(z))\) with \(\det A(0) \neq 0\), i.e. \((l, k)\) can be represented in terms of a (left) matrix fraction description (MFD) of the form

\[(l(z), k(z)) = a^{-1}(z)(d(z), b(z))\]

where \(a(z) = \sum_{j=0}^{p} A(j)z^j\), \(d(z) = \sum_{j=0}^{r} D(j)z^j\) and \(b(z) = \sum_{j=0}^{q} B(j)z^j\) are polynomial matrices of dimension \(s \times s\), \(s \times m\) and \(s \times s\), respectively. The natural numbers \(p, r\) and \(q\) are chosen such that they correspond to the maximum degrees of all entries in \(a(z), d(z)\) and \(b(z)\), respectively.

**Proof.** See, e.g. the proof of lemma 1.2.1 in (Hannan and Deistler, 1988). \(\square\)

**Remark 1.5.1.** The assumption \(\det A(0) \neq 0\) guarantees that \((l(z), k(z)) = \frac{1}{\det(a(z))} \text{adj}(a(z))(d(z), b(z))\) has a convergent power series expansion at \(z = 0\), i.e. \((l, k)\) is causal.

**Remark 1.5.2.** The assumption \(\det A(0) \neq 0\) also implies that \(a(z)\) will be nonsingular for all \(z \in \mathbb{C}\) except for a finite number of points \(z_0\). Hence, the (normal) rank of \(a(z)\) is \(s\) and \(a(z)\) is nonsingular.

**Remark 1.5.3.** Clearly, \(A^{-1}(0)(D(0), B(0)) = (D, I)\) must hold true.

In the sequel we consider the set of all ARMAX systems \((a, d, b)\) for fixed \(m\) and \(s\), but *variable* maximum degrees \(p, r\) and \(q\). This set of polynomial matrices is denoted by \(T_A\). Clearly, all elements of \(T_A\) must satisfy \(\det A(0) \neq 0\) and \(A^{-1}(0)(D(0), B(0)) = (D, I)\) for an arbitrary \(D \in \mathbb{R}^{s \times m}\).

With slight abuse of notation, the mapping attaching transfer functions to polynomial matrices in \(T_A\) will again be denoted by \(\pi\):

\[\pi : T_A \rightarrow U_A \quad (a(z), d(z), b(z)) \mapsto a^{-1}(z)(d(z), b(z))\]

In the sequel, it will always be clear whether \(\pi\) refers to the state-space or the ARMAX case. It is straightforward to see that the image set \(\pi(T_A)\) is indeed \(U_A\).

Next, we consider the set \(T(n)\) comprising all ARMAX systems \((a, d, b)\) in \(T_A\) such that \(\pi(a, d, b) = (l, k)\) has McMillan degree \(n\). According to lemma (1.5.1), we obtain \(\pi(T(n)) = \mathfrak{M}(n)\). It is important to note that if \((a, d, b) \in T(n)\) then, clearly, also \((\tilde{a}, \tilde{d}, \tilde{b}) = p(z)(a, d, b) \in T(n)\) where \(p(z)\) denotes any polynomial \(s \times s\) matrix with nonsingular \(p(0)\). In particular, the maximum degrees \(p, r\) and \(q\) cannot be bounded in \(T(n)\). Note, moreover, that \(\pi(T(n))\) does not contain any transfer functions of McMillan degree other than \(n\) by definition.

By \(\nu(\psi(z))\) we denote the degree of the polynomial \(\psi(z)\). We can then introduce the set

\[T_l(n) = \{(a, d, b) \in T(n) | a^{-1}(d, b) \text{ is irreducible } \}\]

This necessitates the following definition:

**Definition 1.5.1 (Irreducibility of an MFD).** A matrix fraction description \(a^{-1}(d, b)\) of \((l, k) = \pi(a, d, b)\) is called irreducible if for all \((\tilde{a}, \tilde{d}, \tilde{b})\) with \(\pi(\tilde{a}, \tilde{d}, \tilde{b}) = (l, k)\) we have: \(\nu(\text{det}(a)) \leq \nu(\text{det}(\tilde{a}))\).

**Remark 1.5.4.** The corresponding ARMAX system \((a, d, b)\) is then called (relatively) left prime or left coprime.
1.5. SPACES OF ARMAX REALIZATIONS

Remark 1.5.5. There are a couple of equivalent characterizations of irreducible MFDs; see lemma 2.2.1 in (Hannan and Deistler, 1988), for instance. One alternative is to call \(a^{-1}(d, b)\) irreducible if \((a, d, b)\) has full rank \(s\) for all \(z \in \mathbb{C}\).

Remark 1.5.6. It is clear from definition (1.5.1) that \(\nu(\det a)\) is minimal among all MFDs and thus invariant for all irreducible MFDs of the same \((l, k)\). Note that if \((a, d, b)\) and \((\tilde{a}, \tilde{d}, \tilde{b})\) are both contained in \(T_0(\mathbb{C})\) but represent different transfer functions, then the degrees \(\nu(\det a)\) and \(\nu(\det (\tilde{a}))\) need not be the same. However, if we consider irreducible MFDs \(\tilde{a}^{-1}(\tilde{d}, \tilde{b})\) of \((\tilde{l}(z), \tilde{k}(z)) = (\tilde{l}(z^{-1}), \tilde{k}(z^{-1})) - (\tilde{l}(0), \tilde{k}(0))\) – see definition (1.3.2) –, then the irreducible MFDs of \((l, k)\) corresponding to any \((l, k) \in \mathbb{M}(n)\) satisfy \(\nu(\det (\tilde{a})) = n\); see section (4.4) for more details. This in turn means that

- the degree \(\nu(\det (\tilde{a}))\), where \(\tilde{a}\) corresponds to any irreducible MFD \(\tilde{a}^{-1}(\tilde{d}, \tilde{b}) = (\tilde{l}, \tilde{k})\), coincides with
- the order of \((l, k)\) as given in definition (1.3.2),
- the rank of the Hankel matrix corresponding to \((l, k)\) according to remark (1.4.5) and
- the state dimension of a minimal state-space realization \((A, B, C, D, K)\) of \((l, k)\) according to remark (1.4.1).

The following definition is of central importance:

Definition 1.5.2 (Observational equivalence in \(T_0(\mathbb{C})\)). Let \((l, k) \in \mathbb{M}(n)\) be given. The set \(\pi^{-1}(l, k) \subseteq T_0(\mathbb{C})\) is called the class of observationally equivalent left prime ARMAX systems or, briefly, the left prime \((l, k)\)-equivalence class. Two left prime ARMAX systems in this equivalence class are called observationally equivalent, which is denoted by \((a, d, b) \sim (\tilde{a}, \tilde{d}, \tilde{b})\).

The following lemma clarifies the structure of the left prime \((l, k)\)-equivalence classes. It is the ARMAX analogue to lemma (3.2.1) in section (3.2):

Lemma 1.5.2 (Structure of the left prime \((l, k)\)-equivalence classes). Two left prime ARMAX systems \((a, d, b)\) and \((\tilde{a}, \tilde{d}, \tilde{b})\) are observationally equivalent if and only if there exists a unimodular polynomial matrix \(u\) such that \((\tilde{a}, \tilde{d}, \tilde{b}) = u(a, d, b)\).

Remark 1.5.7. Clearly, one need not restrict the definition of observational equivalence to \(T_0(\mathbb{C})\). A very general possibility would be to consider the whole set \(T_a\) and call the set \(\pi^{-1}(l, k)\) an \((l, k)\)-equivalence class; see remark (1.4.8).

Remark 1.5.8. Starting from a left prime ARMAX system \((a, d, b)\) and a general \((\tilde{a}, \tilde{d}, \tilde{b})\), \((a, d, b) \sim (\tilde{a}, \tilde{d}, \tilde{b})\) holds true if and only if there exists a nonsingular polynomial matrix \(u\) such that \((\tilde{a}, \tilde{d}, \tilde{b}) = u(a, d, b)\).

Remark 1.5.9. It is important to note that the left prime \((l, k)\)-equivalence classes contain ARMAX systems \((a, d, b)\) of arbitrarily high maximum degrees \(p, q\) and \(r\) in the multi output case. This is because unimodular \(s \times s\) matrices can have polynomial entries of arbitrarily high degree if \(s > 1\).

Finally, canonical forms can be defined:

Definition 1.5.3 (Canonical form for left prime ARMAX systems). Let \(T_0(\mathbb{C})\) be given. A canonical form (for system equivalence) is a map \(\Gamma : T_0(\mathbb{C}) \rightarrow T_0(\mathbb{C})\) such that

\[
\Gamma(a, d, b) \sim (a, d, b) \quad \forall (a, d, b) \in T_0(\mathbb{C})
\]

\[
(a, d, b) \sim (\tilde{a}, \tilde{d}, \tilde{b}) \quad \Rightarrow \quad \Gamma(a, d, b) = \Gamma(\tilde{a}, \tilde{d}, \tilde{b})
\]

if \((a, d, b)\) and \((\tilde{a}, \tilde{d}, \tilde{b})\) are elements of \(T_0(\mathbb{C})\).

Remarks analogous to those below definition (1.4.5) can also be made at this point.
Chapter 2

Parameter Estimation

This chapter deals with different methodologies of identification of linear systems, both from a theoretical and practical point of view. We start with an introduction of the Kalman filter in section (2.1). This section is included for the following reasons: First, it will become clear in section (2.3) that the Kalman filter plays a vital role for the practical computation of the likelihood function and thus has become an essential tool for the estimation of linear systems which is, of course, a central topic of this thesis. Second, we will provide a rather general version of the Kalman filter which seems to be hard to find in the literature (at least if one does not want to combine simpler versions from various places). A simple proof is also presented.

The remaining two sections are intended as a short overview of two identification techniques: Estimators obtained by optimization of some type of criterion function over an a priori specified parameter space comprise the class of so called M-estimators. Probably the most important criterion functions are of the Gaussian maximum likelihood type as discussed in section (2.3). The computation of the (exact) maximum likelihood function using the Kalman filter with a particular initialization is treated as well as asymptotic equivalence of maximum likelihood estimation with other methods using slightly modified criterion functions. Assumptions on data and model class required for the main results on (coordinate free) consistency and asymptotic normality are briefly dealt with.

A second approach to identification of linear systems is the class of so called direct procedures. Here, the parameter estimator is not defined implicitly as an optimizing argument of some criterion function, but rather given explicitly as a function of data. Subspace identification techniques are examples for such direct procedures and section (2.2) is intended to provide the reader with a general view of these techniques, including some remarks on suitable choices of design parameters.

Finally, section (2.4) presents some figures showing the performance of the Kalman filter in various situations.

2.1 The Kalman filter

Useless saying that the Kalman filter is an important tool in connection with estimation in all fields where state-space systems turn up. There is, of course, an extensive literature on the subject, and besides Kalman’s original papers we just mention (Anderson and Moore, 1979), (Hannan and Deistler, 1988) and (Caines, 1988). Emphasis on economic applications is in the book (Harvey, 1994).

2.1.1 The time varying Kalman filter equations

In the sequel we will introduce the important equations for the general case where

- the noise terms appearing in the transition equation and the measurement equation of the state-space model are correlated as it is also the case for the innovations representation.

\[\text{Of course, there is some ambiguity in the distinction between these two different approaches . . .}\]
different noise terms $\xi_t$ and $\eta_t$ of (possibly) different dimensions appear in the state transition and the measurement equation.

- exogenous inputs $(u_t)$ are incorporated and the present input $u_t$ is assumed to influence the present output $y_t$.

- The exogenous variables $u_{t+n}, n \geq 1$ can be predicted by some model for $(u_t)$. This assumption includes cases where the inputs are considered to be perfectly predictable (consider, for instance, engineering applications where $(u_t)$ is some user-defined – and thus deterministic – excitation signal) as well as the case where the prediction of future values of $(u_t)$ is only possible with nonzero prediction errors.

- the coefficient and covariance matrices can be time varying, i.e. instead of $A,B,K,C,D,Q,R$ and $S$ one could have $A_t,B_t,K_t,C_t,D_t,Q_t,R_t$ and $S_t$ in theorem (2.1.1) below. For notational simplicity, however, we have adopted the time invariant approach here, but the proof remains essentially unchanged in case of time varying matrices.

- the system considered below need not be stable\(^2\) and clearly not minimum phase.

- the error terms in the transition and measurement equation need not be Gaussian. The proof only uses Hilbert space arguments for linear projections, and the filter delivers optimal linear least squares estimates. This is independent of the distribution of the error terms (as long as the second moments are finite, of course).

The model considered is of the form\(^3\)

\[
\begin{align*}
x_{t+1} &= Ax_t + Bu_t + K\xi_t \\
y_t &= Cx_t + Du_t + \eta_t
\end{align*}
\]

We make the following assumptions:

(a) The process $(\xi'_t, \eta'_t)$ is a white noise process with

\[
E \begin{pmatrix} \xi_t \\ \eta_t \end{pmatrix} \begin{pmatrix} \xi_t' \\ \eta_t' \end{pmatrix}' = \begin{pmatrix} Q & S \\ S' & R \end{pmatrix} \geq 0, R > 0
\]

(b) The exogenous input process $(u_t)$ satisfies $u_s = 0$ for $s \leq 0$. Moreover, it has finite second (and therefore also finite first) moments: $E u_t < \infty$, $E u'_t u_t < \infty$ for all $t \in \mathbb{Z}$. Finally, $(u_t)$ is orthogonal to both processes $(\xi_t)$ and $(\eta_t)$, i.e. $E u_t \xi'_t = 0$, $E u_t \eta'_t = 0$ for all $s,t \in \mathbb{Z}$. Note that this formulation also includes the case of bounded deterministic input sequences $(u_t)$.

(c) The system (2.1) is started at time $t = 0$ with $x_0$ having mean $\mu_0$ and variance $\Omega_0$. Additionally, $x_0$ is orthogonal to the processes $(\xi_t)$, $(\eta_t)$ and $(u_t)$, i.e. $E x_0 \xi'_t = 0$, $E x_0 \eta'_t = 0$ and $E (x_0 - \mu_0)(u_t - E u_t)' = 0$ for all $t \in \mathbb{Z}$.

We make a few remarks:

**Remark 2.1.1.** The third assumption means that we only consider the solution to the difference equation in (2.1) for $t \in \mathbb{N}$ which is obtained by choosing a particular initial $x_0$. Therefore,

- $x_t$ is a linear function of $x_0, \xi_0, \ldots, \xi_{t-1}, u_1, \ldots, u_{t-1}, t \geq 1$ only
- $y_t$ is a linear function of $x_0, \xi_0, \ldots, \xi_{t-1}, \eta_t, u_1, \ldots, u_t, t \geq 1$ only

\(^2\)Note that all projections in the proof of theorem (2.1.1) are only onto spaces spanned be a finite number of past random variables.

\(^3\)Note that omitting the $K$ matrix in the equation below would not mean a restriction of generality. However, the matrix is included for an easier application of the following results to the case where an innovations representation (1.1) is considered instead.
2.1. THE KALMAN FILTER

Remark 2.1.2. Note that if $A$ is stable, we may consider another solution to the state transition equation in (2.1), namely the steady state solution (on the negative integers) where $x_0 = \sum_{j=-\infty}^{\infty} A^{j-1} K \xi_{0-j}$; note that $u_s = 0$ for $s \leq 0$. This case is, however, also covered by our exposition, as we would only have to choose $\mu_0 = 0$ and $\Omega_0 = P$ where $P$ is the unique solution to the Lyapunov equation $P = APA' + KQK'$. This implies that

- $x_t$ is a linear function of $\xi_{-1}, \xi_{-2}, \ldots \xi_0, \xi_{-1}, \ldots$ and $u_1, \ldots, u_{t-1}, t \geq 1$
- $y_t$ is a linear function of $\xi_{-1}, \xi_{-2}, \ldots \xi_0, \xi_{-1}, \ldots$ and $\eta, u_1, \ldots, u_t, t \geq 1$

Remark 2.1.3. Finally, also note that if (2.1) is a stable innovations model of the form (1.1), $Q = R = S = \Sigma$ holds true.

We can put

$$\zeta_t = \sum_{j=1}^{t-1} A^{j-1} Bu_{t-j}, \quad t \geq 2 \quad \zeta_1 = 0$$

and it is clear that

$$
\zeta_{t+1} = Bu_t + ABu_{t-1} + \cdots + A^{t-1} Bu_1 \\
= A\zeta_t + Bu_t
$$

(2.3)

Let us extract the influence of $u_1, \ldots, u_t$ from the state and the output:

$$
x_t^u = x_t - \zeta_t \quad t \geq 1
$$

(2.4)

$$
y_t^u = y_t - C\zeta_t - Du_t
$$

(2.5)

Then, $(x_t^u, y_t^u)$ satisfy the following system equations

$$
x_{t+1}^u = x_{t+1} - \zeta_{t+1}
$$

$$
= Ax_t + Bu_t + K\xi_t - A\zeta_t - Bu_t
$$

$$
= Ax_t^u + K\xi_t
$$

(2.6)

$$
y_t^u = Cx_t + Du_t + \eta_t - C\zeta_t - Du_t
$$

$$
= Cx_t^u + \eta_t
$$

(2.7)

Clearly, $y_t^u = CA^t x_0 + CA^{t-1} K\xi_0 + CA^{t-2} K\xi_1 + \cdots + CA K\xi_{t-2} + CK\xi_{t-1} + \eta_t$.

Let us now introduce the following notation:\footnote{All predictions below are clearly based on the finite past only, and thus are different from the quantities $x_{t|s}$, etc. which depend on the infinite past of $y_k, u_k$ with $k \leq s$ whenever this infinite past has meaning (for instance, in case of stationarity).}

$$
\hat{Z}_t|s = (Z_t | u_1, \ldots, u_s, y_1, \ldots, y_s) \quad \text{for } t > s \geq 1
$$

Here, $(Z_t | u_1, \ldots, u_s, y_1, \ldots, y_s)$ or, equivalently, $(Z_t | sp\{u_1, \ldots, u_s, y_1, \ldots, y_s\})$, are shorthand notations for the orthogonal projection of the square integrable random variable $Z_t$ onto the Hilbert space $sp\{u_1, \ldots, u_s, y_1, \ldots, y_s\}$ where $u_{i}^{(i)}$ and $y_{i}^{(j)}$ denote the $i.$ and $j.$ component of the random variables $u_i$ and $y_i$, respectively and $sp\{\}$ denotes the set of all linear combinations of the arguments. In the sequel, $Z_t$ will be $x_t, y_t, u_t, \zeta_t, x_t^u, \text{or } y_t^u$. We have the following results:
\begin{equation}
\hat{u}_{t|s} = (u_t | u_s, \ldots, u_s) \quad t > s \geq 1
\end{equation}

\begin{equation}
\hat{\zeta}_{t|s} = (\zeta_t | u_s, \ldots, u_s) \quad t > s \geq 1
\end{equation}

\begin{equation}
\hat{x}_{t|s} = (x_t | y_t^u, \ldots, y_s^u) \quad t > s \geq 1
\end{equation}

\begin{equation}
\hat{y}_{t|s} = (y_t^u | y_t^{u}, \ldots, y_s^{u}) \quad t > s \geq 1
\end{equation}

\begin{equation}
\hat{\eta}_{t|s} = 0 \quad t > s \geq 1
\end{equation}

Equation (2.12) is evident as \( \eta_t \) is orthogonal to \( x_0, \xi_0, \ldots, \xi_{s-1}, \eta_1, \ldots, \eta_s \) and \( u_1, \ldots, u_s \) for \( s < t \). To see (2.8)-(2.11), note that

\[
sp\{u_1, \ldots, u_s, y_1, \ldots, y_s\} = sp\{u_1, \ldots, u_s, y_1^u, \ldots, y_s^u\}
\]

\[
= sp\{u_1, \ldots, u_s, CAx_0 + CK\xi_0 + \eta_1, \ldots, CA^s x_0 + CA^{s-1}K\xi_0 + \ldots + CK\xi_{s-1} + \eta_s\}
\]

\[
= sp\{u_1, \ldots, u_s\} \oplus sp\{y_1^u, \ldots, y_s^u\}
\]

because \( (u_t) \) is orthogonal to \( x_0, (\zeta_t) \) and \( (\eta_t) \). Hence, all projections in (2.8)-(2.11) can be done onto \( sp\{u_1, \ldots, u_s\} \) and \( sp\{y_1^u, \ldots, y_s^u\} \) individually. As \( u_t \) and \( \zeta_t \) are orthogonal to \( y_1^u, \ldots, y_s^u \) and \( x_t^u \) and \( y_t^u \) are orthogonal to \( u_1, \ldots, u_s \), (2.8)-(2.11) follow immediately.

As \( \zeta_t \) depends only on \( u_{t-1}, \ldots, u_1 \), and the past and the present of the observed input is trivially known at time \( t \), the first of the following two equations is obvious; the second one is a consequence of the first one and equation (2.3):

\begin{equation}
\hat{\zeta}_{t+1|t} = \zeta_{t+1}
\end{equation}

\begin{equation}
\hat{\zeta}_{t+1|t} = A\hat{\zeta}_{t|t-1} + Bu_t
\end{equation}

Finally, from (2.4) and (2.5), we get

\begin{equation}
\hat{x}_{t+1|t} = \hat{x}_{t|t+1} + \hat{\eta}_{t+1|t}
\end{equation}

\begin{equation}
\hat{y}_{t+1|t} = \hat{y}_{t+1|t} + C\hat{\zeta}_{t+1|t} + D\hat{u}_{t+1|t}
\end{equation}

We are now ready to state a quite general formulation of the Kalman filter:

**Theorem 2.1.1.** Let the system (2.1) together with assumptions (a), (b) and (c) be given. Let \( e_t = y_t - \hat{y}_{t|t-1} \). Then the following equations are satisfied:

\begin{equation}
\hat{x}_{t+1|t} = A\hat{x}_{t|t-1} + K_t e_t + Bu_t - K_t D(u_t - \hat{u}_{t|t-1})
\end{equation}

\begin{equation}
y_t = C\hat{x}_{t|t-1} + D\hat{u}_{t|t-1} + e_t
\end{equation}

\begin{equation}
P_{t+1} = AP_t A' + KQK' - K_t \Sigma_t K_t' \quad \text{where} \quad P_t = \mathbb{E}(x_t - \hat{x}_{t|t-1})(x_t - \hat{x}_{t|t-1})'
\end{equation}

\begin{equation}
\Sigma_t = CP_tC' + R \quad \text{where} \quad \Sigma_t \neq \mathbb{E} e_t e_t' \text{ if } DDE(u_t - \hat{u}_{t|t-1})(u_t - \hat{u}_{t|t-1})'D' \neq 0
\end{equation}

\begin{equation}
K_t = (AP_tC' + KS)\Sigma_t^{-1}
\end{equation}

**Remark 2.1.4.** Evidently, the first two equations of theorem (2.1.1) can be rewritten such that the output is \( e_t \) and the input consists of \( (u_t, y_t) \) and the predictions \( (\hat{u}_{t|t-1}) \) obtained from some model for \( (u_t) \):

\begin{equation}
\hat{x}_{t+1|t} = A\hat{x}_{t|t-1} + K_t(y_t - C\hat{x}_{t|t-1} - D\hat{u}_{t|t-1}) + Bu_t - K_t D(u_t - \hat{u}_{t|t-1})
\end{equation}

\begin{equation}
e_t = -C\hat{x}_{t|t-1} - D\hat{u}_{t|t-1} + y_t
\end{equation}

\begin{equation}
\hat{\eta}_{t+1|t} = \hat{\eta}_{t+1|t}
\end{equation}

\begin{equation}
\hat{\zeta}_{t+1|t} = \hat{\zeta}_{t+1|t}
\end{equation}

\begin{equation}
\hat{\eta}_{t+1|t} = \hat{\eta}_{t+1|t}
\end{equation}

\begin{equation}
\hat{\zeta}_{t+1|t} = \hat{\zeta}_{t+1|t}
\end{equation}
2.1. THE KALMAN FILTER

Remark 2.1.5. In view of the second equation in (2.19), it would perhaps be more precise to write $P_{t|t-1}$ instead of $P_t$. Analogously, one could write $\Sigma_{t|t-1}$ instead of $\Sigma_t$ and $K_{t|t-1}$ instead of $K_t$.

Remark 2.1.6. Clearly, for the recursions to yield the true best linear least squares estimates, the filter must be started off with the “correct” initial values $\tilde{u}_{1|0}$, $\tilde{x}_{1|0}$ and $P_1$ (note that the best linear least squares estimate of a random variable $Z_1$ based on no information is the (unconditional) expected value of $Z_1$): $\tilde{u}_{1|0}$ should be chosen to be the unconditional expected value of $u_1$ and, accordingly, one should choose $\tilde{x}_{1|0} = A\mu_0$ and $P_1 = A\Omega_0 A^T + KQK^T$. As $\mu_0$ and $\Omega_0$ will be unknown in practice, however, $\tilde{u}_{1|0}$, $\tilde{x}_{1|0}$ and $P_1$ will be chosen according to some initial guess. Clearly, $P_1$ then prescribes the confidence of the choice of the initial value for $\tilde{x}_{1|0}$. As will be discussed below, the influence of the initial values decreases exponentially fast in many cases and is thus not critical; see section (2.1.2) for details.

Remark 2.1.7. The matrix difference equation (2.19) is called Riccati equation (RE). The evolution of the covariance matrix $P_t$ is independent of the data and can therefore be computed off-line.

Remark 2.1.8. From equations (2.17) and (2.18) it follows that the Kalman filter yields a linear system with time varying parameter matrices even though the original system (2.1) is time invariant.

Remark 2.1.9. For a given system of the form (2.1), the question whether the filter (2.17) and (2.18) converges to a time invariant filter (i.e. $K_t \rightarrow K$) is treated in the next subsection. In case this is true, another question is whether this time invariant filter (obtained by replacing $K_t$ by $K$) is stable. This is also addressed there.

We can now prove the theorem. Note that the proof is only using Hilbert space arguments:

Proof. Let us denote the Hilbert spaces $sp\{y^u_1, \ldots, y^u_n\}$ and $sp\{e^u_t\}$ (where $e^u_t = y^u_t - \tilde{y}^u_{t|t-1}$) by $H^u_t$ and $I^u_t$, respectively. As $e^u_t \perp H^u_{t-1}$ by definition, we have

$$H^u_t = H^u_{t-1} \oplus I^u_t$$

Moreover, put $P^u_t = \mathbb{E}(x^u_t - \tilde{x}^u_{t|t-1})(x^u_t - \tilde{x}^u_{t|t-1})'$. By (2.4), (2.13) and (2.15), we have $(x^u_t - \tilde{x}^u_{t|t-1}) = (x_t - \tilde{x}_{t|t-1} + \tilde{\xi}_t) = (x_t - \tilde{x}_{t|t-1})$, and it follows that $P^u_t = P_t$ with $P_t$ as in (2.19).

We will now prove the theorem step by step:

- (2.18): This is directly obtained from (2.1) and the linearity of projections:

$$e_t = y_t - \tilde{y}_{t|t-1} = y_t - C\tilde{x}_{t|t-1} - D\tilde{u}_{t|t-1} = y_t - C\tilde{x}_{t|t-1} - D\tilde{u}_{t|t-1}$$

This shows (2.18). Moreover, note that from equation (2.7) it follows that

$$e^u_t = y^u_t - C\tilde{x}^u_{t|t-1} - \tilde{\eta}_{t|t-1} = y^u_t - C\tilde{x}^u_{t|t-1}.$$

Therefore, from (2.5) and (2.15), we get

$$e_t - e^u_t = y_t - C\tilde{x}_{t|t-1} - D\tilde{u}_{t|t-1} - y^u_t + C\tilde{x}^u_{t|t-1} = (y_t - y^u_t) - C(\tilde{x}_{t|t-1} - \tilde{x}^u_{t|t-1}) - D\tilde{u}_{t|t-1} = (C\tilde{\zeta} + Du) - C\tilde{\zeta}_{t|t-1} - D\tilde{u}_{t|t-1} = D(u_t - \tilde{u}_{t|t-1}) \quad (2.24)$$

- (2.20): Put $\Sigma^u_t = \mathbb{E}e^u_t e^u_t'$. Then, $\Sigma^u_t = \mathbb{E}(y^u_t - C\tilde{x}^u_{t|t-1})(y^u_t - C\tilde{x}^u_{t|t-1})' = \mathbb{E}[C(x^u_t - \tilde{x}^u_{t|t-1}) + \eta^u_t]C(x^u_t - \tilde{x}^u_{t|t-1}) + \eta^u_t' = C\mathbb{E}(x^u_t - \tilde{x}^u_{t|t-1})(x^u_t - \tilde{x}^u_{t|t-1})' + R = CP^u_t C' + R$ because $\eta^u_t \perp x^u_t$ and $\eta^u_t \perp \tilde{x}^u_{t|t-1}$. As $P^u_t = P_t$, we get $\Sigma^u_t = CP_t C' + R$ and putting $\Sigma_t = \Sigma^u_t$ yields (2.20). The additional remark in line (2.20) is easily seen from (2.24) because $e_t = e^u_t + D(u_t - \tilde{u}_{t|t-1})$ and $e^u_t \perp \{u_1, \ldots, u_t\}$ implies

$$\Sigma_t = \mathbb{E}e_t e^t_t' = \Sigma_t + DE(u_t - \tilde{u}_{t|t-1})(u_t - \tilde{u}_{t|t-1})'D'$$
\[(2.21) \text{ and } (2.17): \text{ From } (2.10) \text{ we know that } \hat{x}_{t+1|t}^u = (x_{t+1}^u|sp\{y_t^u, \ldots, y_t^u\}). \text{ Hence, we get}
\]

\[
\hat{x}_{t+1|t}^u = (Ax_t^u + K\xi_t|H_{t-1}^u \oplus I_t^u)
\]

\[
= A\hat{x}_{t|t-1}^u + 0 + A\xi_t^ve_t^u[\xi_t^ve_t^u]^{-1}e_t^u + K\xi_t_ee_t^u_2[\xi_t^ve_t^u]^{-1}e_t^u
\]

This is because projections are linear, the regressors are orthogonal and we can therefore regress on each regressor individually. Note that the projection of \(\xi_t\) onto \(H_{t-1}^u\) is zero. Now

\[
\begin{align*}
E_x^u e_t^u &= E_x^{u'}(y_t^u - C\hat{x}_{t|t-1}^u)' = E_x^{u'}[C(x_t^u - \hat{x}_{t|t-1}^u) + \eta_t]' \\
&= E(x_t^u - \hat{x}_{t|t-1}^u)(x_t^u - \hat{x}_{t|t-1}^u)'C' = P_t^u C' \\
E_\xi e_t^u &= E_\xi(y_t^u - C\hat{x}_{t|t-1}^u)' = E_\xi[C(x_t^u - \hat{x}_{t|t-1}^u) + \eta_t]' = S
\end{align*}
\]

such that

\[
\begin{align*}
\hat{x}_{t+1|t}^u &= A\hat{x}_{t|t-1}^u + AP_t^u C'S_t^u e_t^u + KS_t^u e_t^u \\
&= A\hat{x}_{t|t-1}^u + (AP_t^u C' + KS_t^u)S_t^{-1} e_t^u
\end{align*}
\]

Put

\[
K_t^u = (AP_t^u C' + KS_t^u)S_t^{-1} = (AP_t C' + KS)S_t^{-1} = K_t
\]

yielding (2.21). Finally, by applying (2.15), we get \(\hat{x}_{t+1}^u = \hat{x}_{t+1|t}^u + \hat{\xi}_{t+1|t} = A\hat{x}_{t|t-1}^u + K_t e_t^u + \hat{\xi}_{t+1|t}\).

Using (2.14), we can continue to write \(\hat{x}_{t+1|t} = A\hat{x}_{t|t-1} + K_t e_t^u + A\hat{\xi}_{t|t-1} + Bu_t = A\hat{x}_{t|t-1} + K_t e_t^u + Bu_t\),

and by plugging in from (2.24), we get

\[
\hat{x}_{t+1|t} = A\hat{x}_{t|t-1} + K_t[e_t - D(u_t - \hat{u}_{t|t-1})] + Bu_t
\]

verifying (2.17).

\(\Box\) (2.19): From (2.27) and (2.6), one gets

\[
P_{t+1} = P_{t+1}^u = E(x_{t+1}^u - \hat{x}_{t+1|t}^u|x_{t+1|t}^u - \hat{x}_{t+1|t}^u)' \\
= E(Ax_t^u + K\xi_t - A\hat{x}_{t|t-1}^u - K_t e_t^u)(Ax_t^u + K\xi_t - A\hat{x}_{t|t-1}^u - K_t e_t^u)' \\
= E[A(x_t^u - \hat{x}_{t|t-1}^u + K\xi_t - K_t e_t^u)(x_t^u - \hat{x}_{t|t-1}^u + K\xi_t - K_t e_t^u)'
\]

As \((x_t^u - \hat{x}_{t|t-1}^u) \perp \xi_t\), we can continue

\[
P_{t+1} = AP_tA' + KQK' + K_t\Sigma_t K_t' - AE(x_t^u - \hat{x}_{t|t-1}^u)e_t^u K_t' \\
- K\xi_t e_t^u K_t' - K_t\xi_t e_t^u_2(x_t^u - \hat{x}_{t|t-1}^u)'A' - K_t\xi_t e_t^u_2 \xi_t' K'
\]

and from (2.25) and (2.26) we get

\[
P_{t+1} = AP_tA' + KQK' + K_t\Sigma_t K_t' - (AP_t C' K_t' + KSK_t') - ((AP_t C' + KS)K_t')' \\
= AP_tA' + KQK' + K_t\Sigma_t K_t' - (AP_t C' + KS)K_t' \\
= AP_tA' + KQK' - K_t\Sigma_t K_t'
\]

This completes the proof. \(\Box\)
2.1. THE KALMAN FILTER

Remark 2.1.10. We consider two special cases for the exogenous input:

1. If the process \( (u_t) \) is such that its past does not contain any covariance information which can be used in least squares prediction (e.g., if all \( u_s \) and \( u_t, s, t \in \mathbb{N} \) are uncorrelated for \( s \neq t \)) and if, additionally, we have removed its mean, then

\[
\hat{u}_{t|s} = 0 \text{ for } t > s \geq 0
\]

Equations (2.17) and (2.18) simplify to

\[
\begin{align*}
\hat{x}_{t+1|t} &= A\hat{x}_{t|t-1} + K_t e_t + (B - K_tD)u_t \\
y_t &= C\hat{x}_{t|t-1} + e_t
\end{align*}
\]

If, in addition, the present \( u_t \) does not influence the present output, i.e. \( D = 0 \), then the first equation further simplifies to \( \hat{x}_{t+1|t} = A\hat{x}_{t|t-1} + K_t e_t + Bu_t \). It is this case which is considered in (Hannan and Deistler, 1988), for instance. Note that \( D = 0 \) also implies \( e_t = e_t^u \) and thus \( \Sigma^e_t = \Sigma^y_t = \Sigma_t \).

2. Let us assume that the input is exactly predictable ((\( u_t \)) might even be deterministic), i.e.

\[
\hat{u}_{t|s} = u_t \text{ for } t > s \geq 0
\]

In particular, this amounts to choosing \( \hat{u}_{1|0} = u_1 \). Note, however, that for \( \hat{u}_{t|s} \) to be in accordance with its previous definition as a linear least squares prediction for \( t > s \geq 1 \), we have to further assume (without restriction of generality) that \( u_t \neq 0 \). This is because if \( u_1 = 0 \) and \( u_2 \neq 0 \), for instance, then we could not find a linear function \( \hat{u}_{2|1} = K_1 u_1 \) such that \( u_{2|1} = u_2 \). The first two equations in the theorem now become

\[
\begin{align*}
\hat{x}_{t+1|t} &= A\hat{x}_{t|t-1} + K_t e_t + Bu_t \\
y_t &= C\hat{x}_{t|t-1} + e_t + Du_t
\end{align*}
\]

Of course, exact predictability of \( (u_t) \) also implies \( e_t = e_t^u \) and thus \( \Sigma^e_t = \Sigma^y_t = \Sigma_t \).

2.1.2 Time invariance and stability of the Kalman filter

As already mentioned in remark (2.1.9), two questions are treated now:

(i) Under what conditions is the filter (2.17) and (2.18) asymptotically time invariant?

(ii) When is the resulting time invariant filter stable? And does that mean that the time variant filter is stable as well?

Before we address these questions, let us first consider the following definition:

Definition 2.1.1 (Stabilizability and Detectability). Let \( (A, B, C, D, K) \in \mathcal{S}(n) \) be a state-space realization of \( (\ell, k) = \pi(A, B, C, D, K) \) and let \( C_n \) and \( O_n \) be the controllability and observability matrix as given in definition (1.4.2). As usual, put \( \bar{B} = (B, K) \). The pair \( (A, \bar{B}) \) is called stabilizable if there exists a matrix \( F \) such that \( A + BF \) is stable. The pair \( (C, A) \) is called detectable if there exists a matrix \( L \) such that \( A + LC \) is stable.

Remark 2.1.11. There are many equivalent formulations for stabilizability and detectability as given in theorems 3.2 and 3.4 of (Zhou et al., 1996), for instance. Clearly, stabilizability means that there exists a state feedback \( (\hat{u}_t, \hat{z}_t)' = Fx_t \) such that the system becomes stable. The pair \( (C, A) \) is detectable if and only if \( (A', C') \) is stabilizable.
Let us now recall the first two equations of theorem (2.1.1) in the form given in (2.22) and (2.23):

\[ \hat{x}_{t+1|t} = (A - K_tC)\hat{x}_{t|t-1} + (B - K_tD)u_t + K_t y_t \]
\[ e_t = C\hat{x}_{t|t-1} - Du_{t|t-1} + y_t \]

Clearly, for asymptotic time invariance, \( K_t \) must converge to a fixed matrix \( \hat{K} \), which is satisfied if \( P_{t+1} \) converges to some (covariance matrix) \( \hat{P} \). In view of (2.19) this matrix must satisfy

\[ \hat{P} = A\hat{P}A' + KQK' - (A\hat{P}C' + KS)(C\hat{P}C' + R)^{-1}(A\hat{P}C' + KS)' \]

(2.30)

which is known as the discrete time algebraic Riccati Equation (ARE). Clearly, \( \hat{\Sigma} = C\hat{P}C' + R \) and \( \hat{K} = (A\hat{P}C' + KS)\hat{\Sigma}^{-1} \).

The following results give sufficient conditions for convergence and asymptotic stability for time invariant models with correlated state and measurement equation errors and can be found in (Caines, 1988) where a separate section is devoted to the properties of the Riccati equation:

**Theorem 2.1.2 (Asymptotic time invariance of the Kalman filter).** Let the state-space system (2.1) be given and let \( (C, A) \) be detectable and \( (A - KSR^{-1}C, K(Q - SR^{-1}S')^{1/2}) \) be stabilizable where \((.)^{1/2}\) denotes any square root matrix. Then, for any given symmetric, positive semidefinite \( P_1 \geq 0 \), the sequence of solutions to the RE (2.19) converges to the unique symmetric, positive semidefinite solution \( \hat{P} \) of the ARE (2.30) and the convergence is exponentially fast: \( \|P_t - \hat{P}\| < c_p^t \) for some \( c > 0 \) and \( 0 < p < 1 \). If \( (A - KSR^{-1}C, K(Q - SR^{-1}S')^{1/2}) \) is controllable, then additionally \( \hat{P} > 0 \).

**Proof.** See theorem 5.4 on page 176 in (Caines, 1988) or page 112 in (Anderson and Moore, 1979).

**Remark 2.1.12.** Starting with \( P_1 = 0 \), detectability of \( (C, A) \) is sufficient to guarantee the convergence of \( P_t \) to a steady state matrix \( \hat{P} \) satisfying the ARE (2.30).

**Theorem 2.1.3 (Stability of the time invariant Kalman filter).** Let a symmetric, positive semidefinite solution \( \hat{P} \) to the ARE (2.30) be given and let \( (A - KSR^{-1}C, K(Q - SR^{-1}S')^{1/2}) \) be stabilizable where \((.)^{1/2}\) denotes any square root matrix. Then \((A - \hat{K}C)\) is stable, i.e. \( |\text{max}(A - \hat{K}C)| < 1 \) and \( \hat{P} \) is the unique symmetric, positive semidefinite solution \( P \) to the ARE (2.30). If \( (A - KSR^{-1}C, K(Q - SR^{-1}S')^{1/2}) \) is controllable, then additionally \( \hat{P} > 0 \).

**Proof.** See theorem 5.3 on page 175 in (Caines, 1988) or pages 112 and 113 of (Anderson and Moore, 1979).

In summary, under the detectability and stabilizability assumptions, we see that the time invariant steady state Kalman filter for the model (2.1) is stable (as well as the time varying but asymptotically time invariant filter (2.17) and (2.18); this is due to the exponential convergence of the filter matrix \( K_t \).

**Remark 2.1.13.** It is worth stressing that these results do obviously not rely on the stability of \( A \), i.e. the stability of the original system. In fact, stability of \( A \) will automatically imply detectability of \( (C, A) \) and stabilizability of \((A, Q^{1/2})\) but not necessarily the stabilizability of \( (A - KSR^{-1}C, K(Q - SR^{-1}S')^{1/2}) \) except for the case when the noise terms are uncorrelated, i.e. \( S = 0 \).

**Remark 2.1.14.** Consider an innovations representation of the form (1.1). Then \( Q = R = S = \Sigma \), and the model is both stable and strictly minimum phase. Thus \((C, A)\) is always detectable and \( (A - KSR^{-1}C, K(Q - SR^{-1}S')^{1/2}) \) is stable and never controllable. From theorem (2.1.2) it follows that \( \hat{P} = 0 \) is the unique solution to the ARE and that convergence takes place exponentially fast for any \( P_1 \). Clearly, \( \hat{P} = 0 \) implies \( \hat{\Sigma} = \Sigma \) and \( \hat{K} = K \). Theorem (2.1.3) evidently holds true now.

**Remark 2.1.15.** Consider now a stable, but not strictly minimum phase model without zeros of unit modulus, of the same form as (1.1), i.e. with \( Q = R = S = \Sigma \). Detectability of \((C, A)\) holds true, but \((A - KSR^{-1}C, K(Q - SR^{-1}S')^{1/2}) = (A - K, C, 0) \) is stabilizable, but never controllable. From theorem (2.1.2) it cannot be applied. Evidently, starting with \( P_1 = 0 \) yields a solution \( \hat{P} = 0 \) to the ARE (2.30), but it is not the only one in this case. Our experiences with simulations lead us to the conjecture that there always
exists a largest \( P^+ \) among the set of solutions to the ARE (2.30) such that the corresponding \( \tilde{K}^+ \) satisfies
\[
|\lambda_{\text{max}}(A - \tilde{K}^+ C)| < 1.
\]
In that case, the Kalman filter equations (2.22) and (2.23) become
\[
\begin{align*}
\hat{x}_{t+1|t} &= (A - \tilde{K}^+ C)\hat{x}_{t|t-1} + (B - \tilde{K}^+ D)u_t + \tilde{K}^+ y_t \\
e_t &= -C\hat{x}_{t|t-1} + D\tilde{u}_{t|t-1} + y_t
\end{align*}
\]
and the generated sequence \((e_t)\) indeed still converges to the true innovations exponentially fast. A proof of this statement cannot be given at this point, however.

**Remark 2.1.6.** If, in addition, the case without exogenous inputs is considered in remark (2.1.15) above, the steady state Kalman filter (2.17) and (2.18) becomes
\[
\begin{align*}
\hat{x}_{t+1|t} &= A\hat{x}_{t|t-1} + \tilde{K}^+ e_t \\
y_t &= C\hat{x}_{t|t-1} + e_t
\end{align*}
\]
which is the innovations representation corresponding to the stable and strictly minimum phase spectral factor of the spectral density of the observed stationary process, i.e. \([C(z^{-1}I - A)^{-1}K + I]\Sigma[C(zI - A)^{-1}K + I]' = [C(z^{-1}I - A)^{-1}\tilde{K}^+ + I]\Sigma^*[C(zI - A)^{-1}\tilde{K}^+ + I]'\). Note that this spectral factor corresponds, after normalization, to the Wold decomposition of the process \((y_t)\).

### 2.1.3 Numerical simulations for the time varying Kalman filter

We conclude this section by presenting a few simulation examples. We consider systems with exogenous input of the form
\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}
\]
where \((u_t)\) is chosen to be a binary signal with amplitude 1.5 and a passband of \((0, \frac{\pi}{2})\) generated in MATLAB by \(u = \text{idinput}([\text{max}, 1], \text{Ars', } [0, 0.5], \text{L-1.5, 1.5])\). We assume that \(u_{t|s} = u_t\) for \(t > s \geq 0\) (perfect predictability). Typical input and output signals can be seen in plot (1) of figure (2.1). We only considered models with identical noise terms \(\xi\) in both the transition and the measurement equation. Hence, \(Q = R = S = \Sigma\) holds true.

In the sequel, three different systems are discussed, namely systems that are

(a) stable and strictly minimum phase,

(b) stable, but not strictly minimum phase without zeros of unit modulus,

(c) stable and minimum phase with zeros of unit modulus

\[
(a) : \begin{bmatrix}
0 & -0.5 & 1 & 0 \\
1 & 0 & 1 & 0.5 \\
0 & 1 & 0.5 & 1
\end{bmatrix}, \quad (b) : \begin{bmatrix}
0 & -0.5 & 1 & 1 \\
1 & 0 & 1 & 0 \\
1 & 1 & 0.5 & 1
\end{bmatrix}, \quad (c) : \begin{bmatrix}
0 & 1 & -1 \\
1 & 0.5 & 1
\end{bmatrix}
\]

We designed a graphical user interface (GUI) that should be pretty much self explanatory; see figures (2.1) and (2.2) in section (2.4) below. For each of the three cases we considered two different situations (e.g. initialization with different \(\hat{x}_{t|0}\) and \(P_1\) and different noise covariances \(Q = R = S = \Sigma\).

First, a few general remarks are in place:
(i) The size of the generated sample is $T = 1000$; $\xi_t$ is the noise term in both the transition and the measurement equation of the original model, $e_t$ is the prediction error obtained from the Kalman filter and $e_{t,\text{pe}}$ denotes the corresponding values obtained by the built-in function \texttt{pe}. The empirical variances of $\xi_t$, $e_t$ and $e_{t,\text{pe}}$ are denoted by "EmpVar" in figures (2.1) and (2.2).

(ii) Note that in any of the following examples the choice $P_1 = 0$ corresponds to a formal "inversion" of the transfer function $(l, k)$; see all plots with a straight line for $\Sigma_{t|t-1}$. The corresponding filter then generates the true innovations only in the case when the original system is in innovations representation and the strict minimum phase property is satisfied.

(iii) Additionally, the built-in function \texttt{pe} seems to require a system in innovations representation in order to work properly, i.e. it is not applicable for the calculation of the true one-step ahead prediction error if the system is not minimum phase. This can be seen by inspection of the empirical variance of the generated error sequence $e_{t,\text{pe}}$ in the cases when the original system is not minimum phase.

We now comment on each individual example separately:

(a) The model is given in terms of an innovations representation without zeros of unit modulus, $(\xi_t)$ is the innovations sequence. Theorem (2.1.2) applies, $P_{t+1}$ in (2.19) converges to $P = 0$ exponentially fast, and $P = 0$ is the unique solution to the ARE (2.30). The generated sequence $(e_t)$ converges to the true innovations exponentially fast: The effects of the choice of $P_1$ (see plot (a,1) in figure (2.1)) and $x_{1|0}$ (see plot (a,2) in figure (2.1)) become irrelevant after a few time instants.

(b) The model is stable, does not have zeros of unit modulus, but is not given in terms of an innovations representation, $(\xi_t)$ is not the innovations sequence. Theorem (2.1.2) does not apply, however $P_{t+1}$ in (2.19) still converges to $P^+ > 0$ exponentially fast if $P_1 = P_{ss}$ (where $P_{ss}$ is the unique solution to the Lyapunov equation $P = AP' + KQK'$; see plot (b,1) in figure (2.2)) or if $P_1 = I_2$ (see plot (b,2) in figure (2.2)). The corresponding innovations variances are given by $\Sigma^+ = 22.25$ (see plot (b,1)) and $\Sigma^+ = 11.25$ (see plot (b,2)), respectively. The generated sequence $(e_t)$ converges to the true innovations exponentially fast in both cases, whereas the built-in function \texttt{pe} fails to deliver the true innovations sequence. This is because \texttt{pe} simply inverts $(l, k)$, i.e. chooses $P_1 = 0$.

(c) This model is given in terms of an innovations representation with a zero of unit modulus: $(l, k) = (0.5 + z, 1 - z)$, i.e. $y_t = 0.5u_t + u_{t-1} + \xi_t - \xi_{t-1}$ and $(\xi_t)$ is the innovations sequence. Theorem (2.1.2) does not apply. For $Q = R = S = \Sigma = 1$ and any initial $P_1 = P_{ss} = 1$ is the unique solution to the Lyapunov equation $P = AP' + KQK'$. As $K_{t+1} = -\frac{1}{1 + \frac{1}{1+tp}} \to -1$ and $(B - K_tD) = 1 + \frac{1}{2(1+tp)} \to \frac{3}{2}$, the time varying and the time invariant filter are given by

\[
\hat{x}_{t+1|t} = \frac{1+(t-1)p}{1+tp} \hat{x}_{t|t-1} + \left(1 + \frac{1 + (t-1)p}{2(1+tp)}\right) u_t + \left(-\frac{1+(t-1)p}{1+tp}\right) y_t
\]

and

\[
e_t = -1 \cdot \hat{x}_{t|t-1} - \frac{1}{2} \cdot u_t + y_t
\]

and

\[
\hat{x}_{t+1|t} = 1 \cdot \hat{x}_{t|t-1} + \frac{3}{2} u_t - y_t
\]

\[
e_t = -1 \cdot \hat{x}_{t|t-1} - \frac{1}{2} \cdot u_t + y_t
\]
respectively. Clearly, the time invariant filter (2.32) corresponds to the initialization \( P_1 = 0 \); see plot (c,2) in figure (2.2). The sequence \( (e_t) \) generated by (2.32) is given by

\[
e_{t+1} = -\hat{x}_1y_1 - \frac{3}{2}y_1 - \frac{3}{2}y_1 + \frac{1}{2}y_{t+1} + \sum_{j=1}^{t+1} y_j
\]

\[
= -\hat{x}_1y_1 + \sum_{j=1}^{t+1} (\xi_j - \xi_{j-1})
\]

\[
= -\hat{x}_1y_1 + \xi_{t+1} - \xi_0
\]

where we have used the fact that \( y_j = \frac{1}{t}u_j + u_{j-1} + \xi_j - \xi_{j-1} \) and that \( u_0 = 0 \). Hence, the output sequence \( (e_t) \) of (2.32) does not converge (a.s.) to the true innovations sequence \( (\xi_t) \) and the effects of initialization of \( \hat{x}_1 \) do not vanish; see plot (c,2) in figure (2.2).

However, for the time varying filter in (2.31) we obtain

\[
e_{t+1} = - \prod_{j=1}^{t} A_j \hat{x}_{1|0} - \prod_{j=1}^{t} A_j [B_1, K_1] \left( \frac{u_1}{y_1} \right) - \cdots - \prod_{j=1}^{t} A_j [B_{t-1}, K_{t-1}] \left( \frac{u_{t-1}}{y_{t-1}} \right) - [B_t, K_t] \left( \frac{u_t}{y_t} \right) + \frac{1}{2} \left( \frac{u_{t+1}}{y_{t+1}} \right)
\]

\[
= - \frac{1}{1 + tp} \hat{x}_{1|0} - \frac{1 + p}{1 + tp} [B_1, K_1] \left( \frac{u_1}{y_1} \right) - \cdots - \frac{1 + (t-1)p}{1 + tp} [B_{t-1}, K_{t-1}] \left( \frac{u_{t-1}}{y_{t-1}} \right) - [B_t, K_t] \left( \frac{u_t}{y_t} \right) + \frac{1}{2} \left( \frac{u_{t+1}}{y_{t+1}} \right)
\]

and after some calculations one gets

\[
e_{t+1} = -\frac{1}{1 + tp} \hat{x}_{1|0} + \frac{1}{1 + tp} y_t + \frac{1 + p}{1 + tp} y_2 + \cdots + \frac{1 + (t-1)p}{1 + tp} y_t + y_{t+1} (t - 1)p
\]

where \( y_t = y_t - \frac{1}{t}u_t - u_{t-1} \), and the quantities \( (e_t) \) in (2.33) can be shown to converge to the true innovations in the mean square sense if \( p > 0 \). Hence, the sequence \( (e_t) \) generated by (2.31) converges to the true innovations and the effects of initialization of \( \hat{x}_{1|0} \) trivially become negligible.

**Remark 2.1.17.** Note that in (c) the output sequence \( (e_t) \) of the time varying filter converged for any initial \( \hat{x}_{1|0} \) to the true innovations, whereas the output sequence \( (e_t) \) of the time invariant filter did not converge to the true innovations for any initial \( \hat{x}_{1|0} \). Despite this fact, this sequence \( (e_t) \) remained bounded for any initial \( \hat{x}_{1|0} \). It is possible to consider examples of systems with zeros of unit modulus where the output sequence \( (e_t) \) of the time invariant filter also does not converge to the true innovations for any initial \( \hat{x}_{1|0} \), but even diverges to infinity for nonzero initial \( \hat{x}_{1|0} \). An example would be the MA(2) model \( y_t = \xi_t - 2\xi_{t-1} + \xi_{t-2} \).

### 2.2 Subspace estimation

Subspace identification can be seen as a means to identify linear systems via a succession of regressions and singular value decompositions. The main property that is used is the fact that the state of a linear system acts as an interface between the past and the future of the process \( (y_t) \).

Let us commence from the system (1.1) in innovations representation corresponding to

\[
(I(z), k(z)) = \begin{bmatrix} A & B & K \\ C & D & I \end{bmatrix}
\]
with $\Sigma \varepsilon_i e_i' = \Sigma$. A state-space realization of the inverse system, i.e., the realization corresponding to the transfer function from $(u'_t, y'_t)$ to $(\varepsilon_t)$, is then given by

$$
(\tilde{I}(z), \tilde{k}(z)) = \begin{bmatrix}
A - KC & B - KD & K \\
-C & -D & 0 \\
0 & 0 & I
\end{bmatrix}
$$

From (2.34) it is clear that

$$
y_t = C x_t + D u_t + \varepsilon_t \\
y_{t+1} = C A x_t + C B u_t + C K \varepsilon_t + D u_{t+1} + \varepsilon_{t+1} \\
y_{t+2} = C A x_{t+1} + C B u_{t+1} + C K \varepsilon_{t+1} + D u_{t+2} + \varepsilon_{t+2}
$$

This can be rewritten as

$$
Y_t^+ = \begin{pmatrix} C \\ CA \\ CA^2 \\
\vdots \\
0
\end{pmatrix} x_t + \begin{pmatrix} D \\ CB \\ CAB \\ CB \\
\vdots \\
0
\end{pmatrix} U_t^+ + \begin{pmatrix} I \\ CK \\ CK \\ CK \\
\vdots \\
0
\end{pmatrix} E_t^+ (2.36)
$$

where $Y_{t}^+ = (y'_t, y'_{t+1}, \ldots)'$ contains the present $y_t$ and the whole future of $(y_t)$; $U_{t}^+$ and $E_{t}^+$ are defined analogously. From (2.35), one can obtain the present state $x_t = \sum_{j=1}^{\infty} (A - KC)^{j-1} (B - KD, K) \begin{pmatrix} u_{t-j} \\ y_{t-j}
\end{pmatrix}$ which can be rewritten in matrix form as

$$
x_t = \begin{pmatrix} u_{t-1} \\ y_{t-1} \\ u_{t-2} \\ y_{t-2} \\ u_{t-3} \\ y_{t-3} \\ \vdots \\
0
\end{pmatrix}
$$

Here, $K \in \mathbb{R}^{n \times \infty}$, and $Z_t^-$ contains the whole past of both $(u_t)$ and $(y_t)$. It follows that

$$
Y_t^+ = \mathcal{O} K Z_t^- + \mathcal{U} U_t^+ + \mathcal{E} E_t^+, \quad \text{where } E_t^+ \perp U_t^+ \text{ and } E_t^+ \perp Z_t^-
$$

under standard assumptions; see, e.g., the assumptions in theorem (2.3.4) in section (2.3.4) below. Thus, we have an infinite regression model. If we restrict ourselves to a finite part, i.e., by considering $Y_{t,f}^+ = (y'_t, y'_{t+1}, \ldots, y'_{t+f-1})'$ only, we obtain

$$
Y_{t,f}^+ = \mathcal{O}_f x_t + \mathcal{U}_f U_{t,f}^+ + \mathcal{E}_f E_{t,f}^+ = \mathcal{O}_f K_p Z_t^- + \mathcal{O}_f (A - KC)^p K Z_t^- + \mathcal{U}_f U_{t,f}^+ + \mathcal{E}_f E_{t,f}^+ (2.39)
$$

where $\mathcal{O}_f \in \mathbb{R}^{s \times n}$, $\mathcal{U}_f \in \mathbb{R}^{s \times f}$ and $\mathcal{E}_f \in \mathbb{R}^{s \times f}$ consist of the first $f$ block rows of $\mathcal{O}$, $\mathcal{U}$ and $\mathcal{E}$, respectively; $K_p \in \mathbb{R}^{n \times p \times m+\delta}$ consists of the first $p$ block columns of $K$. Note that $O_f K Z_t^-$ contains the whole past of $(u_t)$ and $(y_t)$—which will not be available in practice—and that this term has been decomposed into the sum $O_f K_p Z_t^- + O_f (A - KC)^p K Z_t^-$ where the first summand contains only the past $p$ observed inputs and outputs and the second one contains the rest. Note, moreover, that (2.39) is
2.2. SUBSPACE ESTIMATION

a regression model, and the term $O_f (A - KC) K Z^-_{i,p}$ will be small if $p$ is only large enough; see remark (2.2.1) below for a more precise discussion. For given data and after neglecting the term just mentioned, the above regression (2.39) would be of the form

$$Y^+_i = O_f K P Z^-_{i,p} + \mathcal{U}_f U^+_i + \mathcal{E}_f E^+_i$$

In terms of the given observations, this equation can be rewritten in matrix form:

$$Y^+_i = O_f K P Z^-_{i,p} + \mathcal{U}_f U^+_i + \mathcal{E}_f E^+_i$$

Here, the expressions for the data matrices are shorthand notations neglecting $f$ and $p$. The main focus in the first step will be on the estimation of

$$\hat{\beta}_z = O_f K P \in \mathbb{R}^{f \times (m+s)}$$

This can be done either by regressing $Y^+_i$ onto both $(Z^-_{i,p}, U^+_i)$ to obtain the estimate $(\hat{\beta}_z, \hat{\beta}_u)$ or by performing two regressions: first regress both $Y^+_i$ and $Z^-_{i,p}$ onto $U^+_i$, i.e. take the residuals

$$Y^+_{res} = Y^+ - (Y^+ U^+')(U^+ U'^+)^{-1} U^+$$

Then,

$$\hat{\beta}_z = (Y^+_{res} Z^-_{res}') (Z^-_{res} Z^-_{res})^{-1}$$

Clearly, this approach neglects the lower triangular block Toeplitz structure of $\mathcal{U}_f$ (see (2.36)). A regression where this structure is explicitly considered (yielding different estimates for $\beta_z$) can be found in (Peternell et al., 1996).

The second step of any of the following procedures contains a singular value decomposition (SVD) of a weighted version of the estimate $\hat{\beta}_z$:

$$\hat{\Sigma}_n \hat{V}_n = \hat{U}_n \Sigma_n \hat{V}_n + \hat{R}_n$$

Here, $\hat{W}_f \in \mathbb{R}^{f \times f}$ and $\hat{W}_p \in \mathbb{R}^{(m+s) \times (m+s)}$ are to be chosen by the user; see remark (2.2.3) below. The matrix $\hat{\Sigma}_n$ contains only the $n$ largest singular values of $\hat{W}_f \hat{\beta}_z \hat{W}_p$, where the integer $n$ also has to be chosen; see remark (2.2.2) below. Clearly, $\hat{U}_n$ and $\hat{V}_n$ consist of the corresponding left and right singular vectors and the matrix $\hat{R}_n$ is the residual matrix containing the contribution of the neglected singular values.

From (2.44), the corresponding estimates for $O_f$ and $K_p$ in (2.41) are then obtained via

$$\hat{O}_f = \hat{W}_f^{-1} \hat{U}_n \Sigma_n^\frac{1}{2}$$

$$\hat{K}_p = \Sigma_n^\frac{1}{2} \hat{V}_n \hat{W}_p^{-1}$$
Remark 2.2.1 (Choice of \(f\) and \(p\)). Of course, the choice of \(p\) and \(f\) is an important and, at the same time, delicate issue. Note that in the case \(n = 0\), a fixed \(f, p \geq n_0\) – the true order \(n_0\) is usually unknown – guarantees that the true \(\beta_z\) has rank \(n_0\) (this is not a necessary condition) whereas in the case where exogenous inputs are present the situation is much more involved ((Chui, 1997) showed that \(f, p \geq 3n_0\) is sufficient for \(\beta_z\) to have rank \(n_0\)). As the order estimation is based on the SVD of \(\hat{\beta}_z\) (which itself depends on \(f\) and \(p\)) one should choose \(f, p\) in any case such that the true \(\beta_z\) has rank \(n_0\) and, of course, such that \(\hat{\beta}_z\) is a consistent estimate of \(\beta_z\). The second objective can only be obtained by \(p \to \infty\) because otherwise one would obtain an asymptotic bias\(^5\) due to the neglected term \(O_f(A - K)KZ_{T-1-p}^\dagger\). Thus one approach is to take \(f = p\) and to choose \(p\) via the following procedure: Estimate an ARX model\(^6\) of the following form

\[
y_t = \sum_{j=1}^{p} A(j)y_{t-j} + \sum_{j=0}^{p} D(j)u_{t-j} + \varepsilon_t
\]

where \(p\) is determined by minimization of the Bayesian information criterion BIC yielding \(\hat{p}_{\text{BIC}}\), for a discussion of BIC and other information criteria, see, e.g., chapter 5 in (Hannan and Deistler, 1988). Under the assumption that \((y_t)\) was generated by an ARMAX model with non-trivial MA-part (i.e. \(k_0(z)\) has zeros not only at \(z = 1\)), \(k_0(z)\) is strictly minimum phase and some stationarity assumptions on the input process \((u_t)\), this BIC estimate can be shown to satisfy

\[
\frac{\hat{p}_{\text{BIC}} - p_{\text{ARX}}}{\log(T)} \to 1 \quad \text{a.s.}
\]

where \(\rho_0\) denotes the inverse of the modulus of the zero of \(k_0(z)\) which is closest to the unit circle\(^7\); see (Hannan and Deistler, 1988), page 261, theorem 6.6.3. The fact that \(\hat{p}_{\text{BIC}}\) tends to infinity at a speed given by the equation above can be used to show that the term \(O_f(A - K)KZ_{T-1-p}^\dagger\) which is neglected by our regression below equation (2.39) – will be of order \(o(T^{-1/4})\) if we choose \(p = [d \cdot \hat{p}_{\text{BIC}}]\), \(d > 1\) ([\(\cdot\]) denotes the greatest integer less than or equal to the value inside the square brackets). This in turn means that neglecting the term in actual computations will not introduce any asymptotic bias.

Remark 2.2.2 (Choice of \(n\)). It is at this stage where a model order selection may take place: Only the first \(n\) singular values are taken into account in the first term on the right hand side of (2.44). The choice of \(n\) is usually based on the minimization of some type of criterion function (similar to information criteria), for example:

\[
\begin{align*}
\arg\min_n \text{NIC}(n) &= \arg\min_n \|\hat{R}_n\|_F^2 + \frac{c(T)}{T}d(n) \\
\arg\min_n \text{SVC}(n) &= \arg\min_n \|\hat{R}_n\|_2^2 + \frac{c(T)}{T}d(n)
\end{align*}
\]

where \(c(T)\) is some function satisfying \(\frac{c(T)}{T} \to 0\), \(c(T) > 0\), \(\frac{c(T)}{T \log \log |T|} \to \infty\), \(\|X\|_2\) and \(\|X\|_F\) denote the spectral norm and the Frobenius norm of the matrix \(X\), respectively, and \(d(n) = 2ns + m(n + s)\) is the number of free parameters needed to cover a generic subset of \(\mathbb{M}(n)\); see chapter 4. As usual, the first term can be viewed as some goodness of fit criterion whereas the second term penalizes models of high complexity\(^8\). NIC was analyzed in (Petersen, 1995), and both criteria (together with a third one) are treated in (Bauer, 2001), for instance.

---

\(^5\)The MOESP algorithm described below estimate \(C_f\) without asymptotic bias for finite \(f\) and \(p\).

\(^6\)Vector ARX systems are identifiable if \(\varepsilon_t\) and \(A(0) = I\). Therefore, all entries in \(A(1), \ldots, A(p), D(0), \ldots, D(p)\) can be taken to be free.

\(^7\)Note that for a true ARX model \(\log \rho_0\) is not defined, but \(\hat{p}_{\text{BIC}}\) will be a consistent estimate of the true \(p_0\) (not tending to infinity). In fact, \(O_f(A - K)KZ_{T-1-p}^\dagger\) will vanish for the true \(p_0\) and therefore, the choice \(p = [d \cdot \hat{p}_{\text{BIC}}]\), \(d > 1\) will be reasonable in this case, too.

\(^8\)Higher complexity here means higher McMillan degree. Of course, there could be other reasonable measures of complexity; see (Deistler, 2001) for a short discussion.
Remark 2.2.3 (Choice of $\hat{W}_f$ and $\hat{W}_p$). Another important issue relates to the choice of the weighting matrices $\hat{W}_f$ and $\hat{W}_p$ in (2.44). Both matrices should be nonsingular (at least from a certain $T_0$ onwards), but there is no clear answer as to how to choose them to get good "asymptotic performance". In fact, the weightings influence the asymptotic bias in case of undermodeling ($n$ is chosen too small) and the asymptotic variance in case of a correct specification. We will give some common choices for $\hat{W}_f$ and $\hat{W}_p$ in the sequel.

Having obtained a rank $n$ approximation $\hat{O}_f\hat{K}_p$ of $\hat{\beta}_z$ from (2.45) and (2.46), the system matrices $(A,B,C,D,K)$ and the innovations covariance $\Sigma$ are finally estimated in the third step. This estimation can be done in several ways:

### 2.2.1 Realization based methods: no exogenous inputs

In the case where no exogenous inputs are present, i.e. $m = 0$, equation (2.39) simplifies to

\[
Y_{t,f}^+ = \hat{O}_f K_p Y_{t,p}^+ + \hat{O}_f (A - KC)^p K Y_{t-p}^- + \epsilon_f E_{t,f}^+ \\
E_{t,f}^+ \perp Y_{t,p}^-, \quad E_{t,f}^+ \perp Y_{t-p}^-(2.47)
\]

Estimation of $\beta_z$ in the first regression step therefore yields

\[
\hat{\beta}_z = (Y^+Y^-')(Y^-Y'^-)^{-1} \\
= \hat{H}_p \hat{\Gamma}_p
\]

where $\hat{H}_p$ is an estimate of the first $f$ block rows and the first $p$ block columns of the Hankel matrix $H_p$ in (A.16) and $\hat{\Gamma}_p$ is the sample covariance matrix of $Y_{t,p}^-$. Note that $\hat{H}_p$ will itself not be a Hankel matrix.

Consider, e.g. the $(1,2)$ element $\hat{H}_{p,12} = y_{1+p}y_{p-1} + y_{2+p}y_{p-2} + \cdots + y_{r_f-1}y_{r_f-1}$ and the $(2,1)$ element $\hat{H}_{p,21} = y_{2+p}y_{p} + y_{3+p}y_{p+1} + \cdots + y_{r_f+1}y_{r_f}$, but include different summands.

The weightings for the SVD in the second step can be chosen to be

\[
\hat{W}_f = I \quad \hat{W}_p = (Y^-Y'^-)^{-\frac{1}{2}} \\
\hat{W}_f = (Y^+Y'^+)^{-\frac{1}{2}} \quad \hat{W}_p = (Y^-Y'^-)^{\frac{1}{2}} (2.48)
\]

with $\hat{O}_f$ and $\hat{K}_p$ as in (2.45) and (2.46); the matrix square roots in (2.49) are taken to be the Cholesky factors. Finally, estimation of $(A,K,C)$ and $\Sigma$ in the third step is done as follows. Put

\[
\hat{C} = \hat{O}_f \hat{A} = (\hat{O}_{f-1}\hat{O}_{f-1})^{-\frac{1}{2}} \hat{O}_{f-1}\hat{O}_f^+ (2.50)
\]

where $\hat{O}_f^+$ denotes the first $s$ rows (i.e. the first block row) of $\hat{O}_f$; $\hat{O}_{f-1}$ denotes the first $f-1$ block rows of $\hat{O}_f$ and $\hat{O}_f^+$ is obtained from $\hat{O}_f$ by deleting the first $s$ rows. Note that the second equation is a least squares solution to $\hat{O}_{f-1}\hat{A} = \hat{O}_f^+$. Clearly, the shift invariance property of the observability matrix is used here, and by introducing an additional zero block in $\hat{O}_f$, stability of $\hat{A}$ can be guaranteed at the expense of an asymptotic bias if $f$ does not tend to infinity: see (Chui and Maciejowski, 1996a) or (Chui and Maciejowski, 1996a). However, stability of $\hat{A}$ is vital for the rest of the procedure. Estimation of $K$ and $\Sigma$ is now done via the Positive Real Riccati Equation: Taking $\hat{D} = \frac{1}{2}\hat{\gamma}(0)$ and $\hat{B}$ to be the first $s$ columns of $\hat{K}_p\hat{\Gamma}_p$ (note that $\hat{O}_f\hat{K}_p\hat{\Gamma}_p = \hat{H}_p$ and see (A.16)), an estimated realization of the spectral summand $\hat{\Phi}(z)$ is given by
\[
\hat{\Phi}(z) = \begin{bmatrix} \hat{A} \\ \hat{C} \\ \hat{B} \\ \hat{D} \end{bmatrix}
\]

and a solution to the \textit{Positive Real Riccati Equation} (A.18) finally yields estimates \( \hat{\Sigma} \) and \( \hat{K} \) via (A.20) and (A.19). Note that due to theorem (A.2.2) the resulting system is guaranteed to be stable and strictly minimum phase if \( \hat{\Phi}(z) \) represents a strictly positive real system. The problem is, however, that there is no guarantee that the estimated \( \hat{\Phi}(z) \) is (strictly) positive real in which case the procedure totally breaks down: no estimates can be provided at all. In (Peternell, 1995) simulation studies revealed that for the second choice of weightings in (2.49), this was even less often the case than for (2.48).

These methods have their origin in the works of (Faurre, 1976) and (Akaike, 1975).

\section{2.2.2 The state approach}

Exogenous inputs may be present now. Therefore, the first and second steps are as described at the beginning of this section. In the third step, \( \hat{K}_p \) from (2.46) is used to reconstruct the state in analogy to (2.37):

\[
\dot{x}_t = \hat{K}_p Z_{t,p}, \quad t = p + 1, \ldots, T
\]

Once the state is known, the measurement equation \( y_t = C x_t + D u_t + \epsilon_t \) amounts to a usual regression model (\( \epsilon_t \) is uncorrelated both with \( x_t \) and \( u_t \)). Regressing \( y_t \) onto \( (x_t, u_t) \) therefore yields estimates \( \hat{C} \) and \( \hat{D} \), and the residuals \( \tilde{\epsilon}_t \):

\[
y_t = \hat{C} \hat{x}_t + \hat{D} u_t + \tilde{\epsilon}_t
\]

Clearly, \( \Sigma_T = \frac{1}{T} \sum_{t=p+1}^{T} \tilde{\epsilon}_t \tilde{\epsilon}_t' \). In order to use the state transition equation as a second regression, an estimate of \( \hat{x}_{t+1} \) is needed: It can either be obtained via shifting \( \hat{x}_t \) one time instant or by considering another SVD and defining \( \hat{x}_{t+1} = \hat{K}_{p+1} Z_{t+1,p+1} \), where \( \hat{K}_{p+1} \) is obtained as in (2.46), considering \( p + 1 \) past inputs and outputs in the computation of \( \hat{\beta}_z \) in the first regression step. The final step then results in a linear least squares problem of the form

\[
\hat{x}_{t+1} = \hat{A} \hat{x}_t + \hat{B} u_t + \hat{K} \tilde{\epsilon}_t + \hat{\rho}_t
\]

yielding \( (\hat{A}, \hat{B}, \hat{K}) \). Note that we always obtain estimates using this approach, but there is no guarantee that the estimated system is stable and minimum phase. Also note that one might consider one single regression comprising both the state transition and the measurement equation:

\[
\begin{pmatrix} \hat{x}_{t+1} \\ y_t \end{pmatrix} = \begin{pmatrix} \hat{A} \\ \hat{C} \\ \hat{B} \\ \hat{D} \end{pmatrix} \begin{pmatrix} \hat{x}_t \\ u_t \end{pmatrix} + \hat{\rho}_t, \quad \tilde{\Omega} = \begin{pmatrix} \hat{Q} \\ \hat{S}' \\ \hat{S} \end{pmatrix}
\]

where \( \tilde{\Omega} \) is the sample covariance matrix of the residuals \( \tilde{\rho}_t \in \mathbb{R}^{n+s} \). The estimation of \( K \) and \( \Sigma \) can be done by an SVD of \( \tilde{\Omega} = U A^{1/2} A^{1/2} U' \), taking the first \( s \) columns of \( U A^{1/2} \) to be \( (K \Sigma^{1/2})', \Sigma^{1/2} \)' to get estimates\footnote{Note that \( \hat{K} \) and \( \hat{\Sigma} \) will be different for the single regression approach.} \( \hat{K} \) and \( \hat{\Sigma} \).

The state approach is used in the following well known algorithms:
2.2. SUBSPACE ESTIMATION

Canonical correlation analysis (CCA, Larimore)

This procedure usually uses the weightings

$$\hat{W}_f = (Y_{res}^+ Y_{res}^{+\prime})^{-1/2}$$
$$\hat{W}_p = (Z_{res}^- Z_{res}^{-\prime})^{1/2}$$

such that the SVD of $(Y_{res}^+ Y_{res}^{+\prime})^{-1/2}(Y_{res}^+ Z_{res}^{-\prime})(Z_{res}^- Z_{res}^{-\prime})^{-1/2}$ is to be calculated. The estimate $\hat{x}_{t+1}$ is obtained by shifting the reconstructed state.

N4SID (Van Overschee, De Moor)

There is more than just one version of N4SID, thus making it difficult to provide an overview here. One algorithm chooses the weightings as follows:

$$\hat{W}_f = I$$
$$\hat{W}_p = (Z^- Z^{-\prime})^{1/2}$$

Note that $\hat{W}_p$ contains the covariance of $Z^-$ and not of $Z_{res}^-$. Finally, N4SID uses a new SVD in order to obtain $\hat{x}_{t+1}$.

2.2.3 MOESP-type algorithms

Again, this is a whole class of algorithms. The most general one – see the P6 scheme in (Verhaegen, 1994) – was originally introduced by M. Verhaegen with the weightings

$$\hat{W}_f = I$$
$$\hat{W}_p = (Z^- Z^{-\prime})^{1/2}$$

and was originally intended to be used for estimation of $l(z)$ only. However, the extension to the estimation of $(l(z), k(z))$ is not difficult. MOESP does not reconstruct the state in order to obtain the system matrices in the third step. Instead, $A$ and $C$ are estimated using the shift invariance property of the observability matrix as in (2.50) and in (2.51). Then the structure of $U_f = U_f(A, B, C, D)$ is used. It follows from (2.36) that for $\breve{C}$ and $\breve{A}$ given, $U_f(\breve{A}, B, \breve{C}, D)$ depends linearly on $B$ and $D$ and can be written as

$$\text{vec } U_f(\breve{A}, B, \breve{C}, D) = L(\breve{A}, \breve{C}) \cdot \begin{pmatrix} \text{vec}(B) \\ \text{vec}(D) \end{pmatrix}$$

(2.53)

From (2.40) we see that

$$O_\hat{f}^{-1/2} Y^+ \approx O_\hat{f}^{-1/2} U_f U^+ + O_\hat{f}^{-1/2} \mathcal{E}_f E^+$$

(2.54)

where $O_\hat{f}^{-1/2} \in \mathbb{R}^{s \times (s-n)}$ is of full column rank and $O_\hat{f}^{-1/2} O_\hat{f} = 0$. Hence, given $O_f$, one can obtain estimates for $B$ and $D$ by considering the following linear least squares problem:

$$\hat{B}, \hat{D} = \arg \min_{B,D} \sum_{i=1}^{T-\hat{f}} \| O_\hat{f}^{-1/2} Y_{i,\hat{f}}^+ - O_\hat{f}^{-1/2} U_{i,\hat{f}}^+ U_{i,\hat{f}}^+ \|_F$$

For $\hat{f} = 1$, $\hat{B}$ corresponds to the one-step prediction error.
where $\|X\|_F$ again denotes the Frobenius norm of the matrix $X$. Similarly, after postmultiplying (2.54) by the appropriate factor, one obtains

$$
O_f(Y+U')(U+U')^{-1} \approx O_f(U+U')(U+U')^{-1} + O_f(E+U')(U+U')^{-1}
\rightarrow O_f(U_f)
(2.55)
$$

where $(U+U')$ has to be invertible for $T$ large enough and $(E+U')$ converges to zero under standard assumptions.

Finally, if $O_f$ is replaced by $\hat{O}_f$ satisfying $\hat{O}_f \hat{O}_f = 0$ and if $\hat{O}_f$ (and therefore $\hat{O}_f$) is a consistent estimate, then (2.55) motivates the following estimation procedure for $B$ and $D$:

$$(B, \hat{D}) = \arg\min_{B, D} \| \hat{O}_f(Y+U')(U+U')^{-1} - \hat{O}_f(U_f(A, B, \hat{C}, D)) \|_F$$

Here, $\text{vec}(\hat{O}_f \cdot U_f \cdot I) = (I \otimes \hat{O}_f) \cdot \text{vec}(U_f)$ was used together with (2.53). Note that the observations $y_1, \ldots, y_p$ and $u_1, \ldots, u_p$ can also be used in (2.54) and (2.55) because the present state $x_t$ and the past $p$ inputs and outputs – have been eliminated by premultiplication of $O_f$.

As already mentioned, MUESP usually only aims at estimating the input-to-output transfer function $l(z)$. However, one could now also obtain estimates for $K$ and $\Sigma$ by again computing the state $\hat{x}_t$ (or its sample covariance matrix) to obtain $\Omega$ in (2.52) and $\hat{K}$ and $\hat{\Sigma}$ as outlined below (2.52).

### 2.2.4 Consistency

As this thesis primarily deals with maximum likelihood estimation of linear systems, we will only briefly mention that consistency results are available for all subspace algorithms discussed above. Of course, the integers $f$ and $p$ and the weighting matrices have to be chosen suitably and assumptions for the exogenous inputs and the white noise process have to be made in analogy to what will be presented for the maximum likelihood case in the sequel; see, e.g., theorem (2.3.4) in section (2.3.4).

### 2.2.5 Asymptotic normality

Again, we will only mention a few results at this point: First, asymptotic normality can be shown for all subspace algorithms discussed above. Of course, results of this kind rely upon conditions for the input and the noise process which are similar in nature to the conditions needed to prove asymptotic normality in the maximum likelihood case; see, e.g., theorem (2.3.5) in section (2.3.5) below. However, the expressions for the asymptotic variances are mostly too complex to allow much insight; see, e.g., (Bauer and Ljung, 2002) for formulae in the case of CCA.

It is important to note that estimates obtained from most subspace algorithms do not seem to be asymptotically efficient, i.e., their asymptotic variance does not reach the (asymptotic) Cramer-Rao lower bound as it is the case for maximum likelihood estimates. It is only the CCA algorithm where asymptotic efficiency could be proved up to now; see (Bauer, 2000).

### 2.3 Maximum likelihood estimation and related M-estimators

#### 2.3.1 Preliminaries

In the sequel we consider a process generated by an innovations model in state-space form (1.1). However, the definition of the maximum likelihood estimates (MLEs) in section (2.3.2) and consistency results in
section (2.3.4) will be given in a coordinate free way, i.e. in terms of the transfer functions and the innovations covariance directly (there is no need to use a particular representation or parametrization). Hence, we first introduce the following notation: Put \( \Sigma = \{ \sigma(\Sigma) | \Sigma > 0, \Sigma' = \Sigma \} \) where \( \sigma(\Sigma) \in \mathbb{R}^{(s+1)/2} \) is the vector of on- and above-diagonal elements of \( \Sigma \). Next, we consider some set \( U \subset U_A \) of stable and minimum phase transfer functions \( (l, k) \) where \( (l(0), k(0)) = (D, I) \) for an arbitrary \( D \in \mathbb{R}^{n \times m} \). Clearly, the set \( U \) may contain transfer functions of different McMillan degrees. Finally,

\[
\Theta \subset U \times \Sigma
\]

(2.56)

denotes the underlying parameter space for our coordinate free approach. The fact that a "\( \rightarrow \)" appears in (2.56) shows that cross restrictions between \( (l, k) \) and \( \Sigma \) could be incorporated. As usual, sets of transfer functions are endowed with the pointwise topology \( T_p \), and sets like \( \Theta \) are endowed with the relative topology in the product space \( U_A \times \mathbb{R}^{(s+1)/2} \). Consider the following two sets:

\[
\hat{\Theta} = \{(l, k, \sigma(\Sigma)) \in \Theta | (l, k) \text{ has no pole for } |z| = 1 \text{ and } \Sigma > 0\}
\]

\[
\Theta^* = \{(l, k, \sigma(\Sigma)) \in \Theta | k \text{ has no zero for } |z| = 1\}
\]

Note that we will always use \( \theta \) for the parameter \( (l, k, \sigma(\Sigma)) \) and \( \theta_0 \) for \( (l_0, k_0, \sigma(\Sigma_0)) \) for the true parameter value in the following subsections.

For the consistency of the MLEs to hold, certain assumptions on the nature of the exogenous input sequence \( (u_t) \), the noise sequence \( (\xi_t) \) and (implicitly) the model class (or even the identification method) have to be made. We will briefly discuss such assumptions now. For a more detailed discussion and references to original works see, e.g. section 1 of chapter 4 in (Hannan and Deistler, 1988).

The exogenous input process \( (u_t) \) is asymptotically stationary, i.e. \( (u_t) \) is non stochastic and satisfies

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T-s} u_{t+s} u_t' = \gamma_u(s) \quad \forall s \in \mathbb{Z}
\]

(2.57)

where (2.57) holds in the sense that the limit exists. Alternatively, we may also call a stochastic process \( (u_t) \) asymptotically stationary if (2.57) is satisfied almost surely. Clearly, in the case when \( (u_t) \) is considered to be non stochastic, it is wholly independent of \( (\xi_t) \). Otherwise, independence amounts to an additional assumption which will be needed below.

The process \( (\xi_t) \) is a white noise process and additionally satisfies (2.59):

\[
\mathbb{E} \xi_t = 0 \quad \mathbb{E} \xi_t \xi_t' = \delta_{s, l} \Sigma
\]

(2.58)

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T-s} \xi_{t+s} \xi_t' = \delta_{0, s} \Sigma \quad \text{a.s.}
\]

(2.59)

Note that the existence of the limits in (2.57) implies that there exists a nonnegative definite Hermitian matrix-valued function \( F_u(\omega) \) such that

\[
\gamma_u(s) = \int_{-\pi}^{\pi} e^{i \omega s} dF_u(\omega)
\]

(2.60)

and \( F_u(\omega_2) - F_u(\omega_1) \geq 0 \) for \( \omega_2 \geq \omega_1 \) and \( F_u(-\omega) = \gamma_u(0) - F_u(\omega)' \) at all points of continuity of \( F_u(\omega) \). This spectral distribution function need not define an absolutely continuous matrix measure on \([-\pi, \pi]\), it will, in fact, jump at some frequencies \( \omega \) in a number of practical applications\(^\text{10}\). Moreover, for asymptotically stationary \( (u_t) \), the following additional relations hold true for the second moments:

\(^{10}\text{Consider, for instance, the incorporation of a sine-wave of period 4 for correction of seasonal impacts on } y_t \text{ in case of quarterly data. Then, } F_u(\omega) \text{ would jump at the corresponding frequencies } \pm \pi/2.\)
\[\gamma_{y}(s) = \mathbb{E}(y_{t+s}y_{t}^{\prime}) = \sum_{j=0}^{\infty} K_{j+s} \sum_{k=0}^{\infty} L_{j} \gamma_{u}(s+j-k)L_{k}^{\prime}\]
\[= \int_{-\pi}^{\pi} e^{i\omega s} l(e^{i\omega}) dF_u(\omega) l(e^{i\omega})^* + \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\omega s} k(e^{i\omega}) \Sigma k(e^{i\omega})^* d\omega \quad (2.61)\]

\[\gamma_{yu}(s) = \mathbb{E}(y_{t+s}u_{t}^{\prime}) = \sum_{j=0}^{\infty} L_{j} \gamma_{u}(s-j)\]
\[= \int_{-\pi}^{\pi} e^{i\omega s} l(e^{i\omega}) dF_u(\omega) \quad (2.62)\]

Here, \(k(e^{i\omega})^* = k(e^{-i\omega})^*\) and \(l(e^{i\omega})^* = l(e^{-i\omega})^*\). It is important to note that \(\gamma_{y}(s) \neq \gamma(s)\) (see (1.13) in chapter 1) because \(\gamma_{y}(s)\) denotes the (non-centered) second moments whereas \(\gamma(s) = \mathbb{E}(y_{t+s} - \mathbb{E}y_{t+s})(y_{t} - \mathbb{E}y_{t})^\prime\). In theory, the infinite covariance sequences (2.60), (2.61) and (2.62) are given. It is then necessary to impose some persistent exciting condition in order to ensure that

Equations (2.60), (2.61) and (2.62) uniquely determine \(k\), \(l\) and \(\Sigma\) from the sequences \(\gamma_{y}(s)\), \(\gamma_{u}(s)\) and \(\gamma_{yu}(s)\) \(\forall s \in \mathbb{Z}\).  

(2.63)

We remark that such a condition means that \((u_{t})\) excites sufficiently many frequencies in a persistent fashion, i.e. (2.60) determines \(F_{u}(\omega)\) as a function increasing at sufficiently many \(\omega\) to determine \(l(z)\) uniquely from (2.61) and (2.62). Note that stronger conditions have to be used for more complex model classes: Consider, for instance, the case \(u_{t} = 1\) for all \(t \in \mathbb{Z}\). From (2.60) we see that \(F_{u}(\omega)\) only jumps at frequency zero by unity, such that from (2.62) we would only get \(l(e^{i0}) = l(1)\) and not \(l(z)\). Let us now distinguish between two ARMAX model classes: \(M_{1} = \{(a(z), d(z), b(z)) = (1 + a_{1} z, d_{0}, 1 + b_{1} z), |a_{1}| < 1, |b_{1}| < 1, a_{1} \neq b_{1} d_{0} \in \mathbb{R}\}\) and \(M_{2} = \{(a(z), d(z), b(z)) = (1 + a_{1} z, d_{0} + d_{1} z, 1 + b_{1} z), |a_{1}| < 1, |b_{1}| < 1, a_{1} \neq b_{1} d_{0} \in \mathbb{R}\}\) where \((l(z), k(z)) = a^{-1}(z)(d(z), b(z))\) and, evidently, \(m = s = 1\). In any case, the first term on the right hand side of (2.61) is \(l(1)l(1)\), which is known, so that the second term is known for all \(s \in \mathbb{Z}\) and hence \(k_{\Sigma} k^{*}\) and thus \(k\) and \(\Sigma\) are known because \(k(0) = 1\) and \(k(z)\) has to be stable and minimum phase by assumption. Therefore, \(a(z)\) and \(b(z)\) are known and in case of model class \(M_{1}\) we can obtain \(d_{0}\) from \(a(1)l(1) = d(1) = d_{0}\), ensuring that \(l(z) = a^{-1}(z)d_{0}\) is known. For model class \(M_{2}\), however, the knowledge of \(l(1)\) is not sufficient to determine \(l(z)\) uniquely as we can only know \(d_{0} + d_{1}\) from \(a(1)l(1) = d_{0} + d_{1}\): The input sequence \(u_{t} \equiv 1\) is not persistently exciting for model class \(M_{2}\). To sum it up, persistent exciting conditions are not restricted to inputs only, but also take into account the model class and (by considering (2.60), (2.61) and (2.62) in this case) the identification method.

### 2.3.2 The finite sample likelihood function and its evaluation

In this section we will define the maximum likelihood estimates without using coordinates, i.e. in terms of the “parameters” \(\theta = (l, k, \sigma(\Sigma))\) as introduced in the section above. For practical reasons, however, the evaluation of the formulae in terms of the matrices \((A, B, C, D, K)\) and \(\Sigma\) in (1.1) will also be given.

We start from the relation \(y_{t} = k(z)\varepsilon_{t} + l(z)u_{t}\). First, we put \(y_{t} = y_{t}^{u} + v_{t}\) where

\[y_{t}^{u} = k(z)\varepsilon_{t}\]
\[v_{t} = l(z)u_{t} \quad \text{with} \quad u_{s} = 0, s \leq 0\]

Thus, \(y_{t}^{u}\) is assumed to be already stationary (see remark (2.1.2) in section (2.1) and \(v_{t} = \sum_{j=1}^{s} L_{j}u_{t-j}\). Hence, using the notation of section (2.1), \(v_{t} = C\varepsilon_{t} + D\varepsilon_{t}\); see (2.5).

Now we can write down the likelihood function as if the innovations \(\varepsilon_{t}\) were Gaussian: As \(A\) is assumed to be stable, the random variable \(y_{t}\), given the input \(u(s), 1 \leq s \leq T\), is Gaussian distributed with mean
and covariance

\[ \gamma(0; \theta) = \mathbb{E}(y_i - \bar{y}_t)(y_i - \bar{y}_t) = e^{i\omega _t \mu} \int_{-\pi}^{\pi} e^{i\omega k(e^{i\omega})^*} \omega d\omega \]

Note that the mean depends on \( \theta \) only through \( l(z) \) and the covariance depends on \( \theta \) only through \( k(z) \) and \( \Sigma \). To be more precise, the latter dependence is only through the spectrum \( k \Sigma k^* \).

Finally, the stacked output vector \( Y_T^T = (y'_1, \ldots, y'_T)^T \) satisfies

\[ (Y_T^T; U_T^T, \theta) \sim N \left( \bar{Y}_T^T(\theta), \Gamma_T(\theta) \right) \]

where

\[ \bar{Y}_T^T(\theta) = Y_T^T(l) = \begin{pmatrix} \bar{y}_1(\theta) \\ \vdots \\ \bar{y}_T(\theta) \end{pmatrix} \quad \Gamma_T(\theta) = \begin{pmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(-T+1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(-T+2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(T-1) & \gamma(T-2) & \cdots & \gamma(0) \end{pmatrix} \]

In terms of the state-space matrices in (1.1), these quantities are given by:

\[ \bar{Y}_T^T(\theta) = \begin{pmatrix} D \\ CB \\ C \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{pmatrix} \quad \Gamma_T(\theta) = \begin{pmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(-T+1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(-T+2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(T-1) & \gamma(T-2) & \cdots & \gamma(0) \end{pmatrix} \]

where the notation \( \gamma(i) = \gamma(i; \theta) \) suppresses the dependence on the parameters. Note that \( \gamma(i) \) is given in (1.13) for \( i > 0 \) and by its transpose for \( i < 0 \). Finally, \( \gamma(0) = CPc + \Sigma \) as in (1.12).

Hence, the likelihood function is given by

\[ l_T(Y_T^T; U_T^T, \theta) = \frac{1}{\sqrt{\det(2\pi \Gamma_T(\theta))}} \exp \left( -\frac{1}{2} (Y_T^T - \bar{Y}_T^T(\theta))^T \Gamma_T^{-1}(\theta) [Y_T^T - \bar{Y}_T^T(\theta)] \right) \]

where \( U_T^T \) is defined in the same way as \( Y_T^T \). Usually, \(-2T\) times the logarithm of the likelihood is considered, which is given up to a constant by (2.67) in the following

**Definition 2.3.1 (Coordinate Free Maximum Likelihood Estimate).** Given observations of the output process \( y_t, 1 \leq t \leq T \), observations of the input process \( u_t, 1 \leq t \leq T \) and a model class \( \Theta \) as defined in (2.56), the maximum likelihood estimate (MLE) \( \hat{\theta}_T = (\hat{l}_T, k_T, \sigma(\Sigma_T)) \in \Theta \) is obtained by minimizing the function

\[ L_T(Y_T^T; U_T^T, \theta) = \frac{1}{T} \log \det \Gamma_T(\theta) + \frac{1}{T} (Y_T^T - \bar{Y}_T^T(\theta))^T \Gamma_T^{-1}(\theta) [Y_T^T - \bar{Y}_T^T(\theta)] \]

over the set \( \Theta \).

**Remark 2.3.1.** In the optimization of (2.67), certain boundary points of \( \Theta \) cannot be avoided: In particular, the MLE \( \hat{k}_T \) may have zeros on the unit circle. However, boundary points corresponding to poles of \( k \) of unit modulus and singular \( \Sigma \) can be excluded because the values of the finite sample likelihood become larger and larger in their neighborhood; see remark (2.3.12) in section (2.3.4). This is the reason why the optimization in definition (2.3.1) is performed over the set \( \Theta \).
Remark 2.3.2. We once more emphasize that Gaussianity is in fact not needed for consistency and asymptotic normality to hold; see sections (2.3.4) and (2.3.5) below.

For practical computations, the representation of the likelihood in (2.66) is quite unfavourable. However, using Bayes’ rule together with the property that the conditional distributions of a Gaussian distributed vector are again Gaussian, we can rewrite (2.66) as the following product of conditional densities \( f(\cdot |\cdot) \)

\[
I_T(Y_1^T; U_1^T, \theta) = \prod_{t=1}^{T} f(y_t|Y_{1:t-1}; U_{1:t}^T, \theta)
\]

\[
= \prod_{t=1}^{T} \frac{1}{\sqrt{\det(2\pi \Sigma_{t|t-1}(\theta))}} e^{-\frac{1}{2}(y_t - y_{t|t-1}(\theta))^T \Sigma_{t|t-1}(\theta)^{-1} (y_t - y_{t|t-1}(\theta))}
\]

where \( f(y_t|Y_{1:t-1}; U_{1:t}^T, \theta) = f(y_t|u_1, \theta) \). Here \( y_{t|t-1}(\theta) \) denotes the mean of the Gaussian distribution of \( (y_t|Y_{1:t-1}; U_{1:t}^T, \theta) \) and \( \Sigma_{t|t-1}(\theta) \) denotes the corresponding variance:

\[
(y_t|Y_{1:t-1}; U_{1:t}^T, \theta) \sim N(y_{t|t-1}(\theta), \Sigma_{t|t-1}(\theta)) \quad t > 1
\]

\[
(y_t|u_1, \theta) \sim N(y_{t|0}(\theta), \Sigma_{t|0}(\theta))
\]

\[
y_{t|0}(\theta) = y_1(\theta) = Du_1
\]

\[
\Sigma_{t|0}(\theta) = \gamma(0) = CPC' + \Sigma
\]

Taking again \(-2T^{-1}\) times the logarithm of the likelihood above yields – up to a constant – the following prediction error decomposition form of the likelihood:

\[
L_T(Y_1^T; U_1^T, \theta) = \frac{1}{2} \sum_{t=1}^{T} \log \det \Sigma_{t|t-1}(\theta) + \frac{1}{2} \sum_{t=1}^{T} \text{tr} \left\{ e_t(\theta) e_t(\theta)^T \Sigma_{t|t-1}(\theta)^{-1} \right\}
\]

(2.70)

where \( e_t(\theta) = y_t - y_{t|t-1}(\theta) \). We have shown how the direct version of the likelihood is transformed into this filter version that depends on \( e_t(\theta) \).

Of course, both forms of the likelihood (2.67) and (2.70) are equivalent. Note, however, that for given data \( Y_1^T, U_1^T \) and given \( \Theta = (l, k, \sigma(\Sigma)) \), (2.67) can be evaluated directly (as indicated in (2.65) for state-space representations) whereas the quantities \( e_t(\theta) \) and \( \Sigma_{t|t-1}(\theta) \) in (2.70) have to be computed using some Kalman filtering procedure.

**Theorem 2.3.1.** Given observations of the output process \( y_t, 1 \leq t \leq T \), observations of the input process \( u_t, 1 \leq t \leq T \) and a model class \( \Theta \) as defined in (2.56), the evaluation of the likelihood function (2.67) and its prediction error form (2.70) yields the same values if the quantities \( e_t(\theta) \) and \( \Sigma_{t|t-1}(\theta) \) are obtained from the Kalman filter in theorem (2.1.1) – \( e_t(\theta) \) is calculated from (2.22) and (2.23) as given in the first remark below the theorem – where

- \( \hat{y}_{t|s} = u_t \) for \( t > s \geq 0 \), \( u_t \neq 0 \), i.e. perfect predictability of the input holds true; see remark (2.1.10) in section (2.1).
- \( \hat{x}_{1|0} = 0 \) and
- \( P_1 = P \) is chosen to be the solution of the Lyapunov equation \( P = APA' + KSK' \).

**Proof.** As can be seen from (2.69), the first (unconditional) probability density function \( f(y_t|Y_{1:t-1}; U_1^T, \theta) = f(y_t|u_1, \theta) \) is Gaussian with mean \( y_{t|0}(\theta) = Du_1 \) and variance \( \Sigma_{t|0}(\theta) = \gamma(0) = CPC' + \Sigma \). Choosing \( P_1 \) and \( \hat{x}_{1|0} \) as in the theorem above, equations (2.22) and (2.23) yield \( e_t = y_t - \hat{y}_{t|0} = -Du_1 + y_t \), i.e. \( y_{t|0} = Du_1 \). Moreover, equation (2.20) yields \( \Sigma_1 = CP_1 C' + \Sigma \). This shows that initializing the Kalman filter as described above yields
2.3. MAXIMUM LIKELIHOOD ESTIMATION AND RELATED M-ESTIMATORS

\[ y_{t|0}(\theta) = \tilde{y}_{1|0} \]
\[ \Sigma_{1|0}(\theta) = \Sigma_1 \]

Note that the quantities on the right hand side are given by the Kalman filter. Equalities for \( t > 1 \), i.e. \( y_{t|t-1}(\theta) = \tilde{y}_{t|t-1} \) and \( \Sigma_{t|t-1}(\theta) = \Sigma_t \), are shown as follows: In section (2.1) it was shown that the Kalman filter generates the prediction errors \( e_t = y_t - \tilde{y}_{t|t-1} \), where \( \tilde{y}_{t|t-1} \) denotes the best linear least squares estimate of \( y_t \) from the finite past \( Y_{1:t}^T, U_1^T \). On Gaussian assumptions, this quantity must coincide with the conditional expectation of \( y_t \) given \( Y_{1:t-1}^T, U_1^T \) and \( \theta \), i.e. with \( y_{t|t-1}(\theta) \). Finally, the quantities \( \Sigma_t \) obtained by the filter satisfy \( \Sigma_t = \text{E}(y_t - \tilde{y}_{t|t-1})(y_t - \tilde{y}_{t|t-1})' \) because of perfect predictability of the exogenous input; see (2.20). Hence, the sequence \( \Sigma_t \) obtained by the filter is the covariance of the conditional distribution of \( y_t \) given \( Y_{1:t}^T, U_1^T \) and \( \theta \), i.e. equal to \( \Sigma_{t|t-1}(\theta) \).

\[ (y_t|y_{1:t-1}, \theta) \sim N \left( \tilde{y}_{t|t-1}(\theta), \Sigma_t \right) \]

Remark 2.3.3. In fact, the Kalman filter recursions are sometimes derived under the restriction of Gaussianity. Such proofs commence from the well known result that

\[ \left( \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_{t^2} \end{array} \right) \sim N \left( \left( \begin{array}{c} \tilde{y}_{1|1}(\theta) \\ \tilde{y}_{2|2}(\theta) \\ \vdots \\ \tilde{y}_{t^2|t^2}(\theta) \end{array} \right), \left( \begin{array}{cc} \gamma(0) & \gamma(-1) \\ \gamma(1) & \gamma(0) \end{array} \right) \right) \]

implies that

\[ (y_2|y_1, U_{1:t}^2, \theta) \sim N \left( \tilde{y}_2(\theta) + \gamma(1)\gamma(0)^{-1}[y_1 - \tilde{y}_1(\theta)], \gamma(0) - \gamma(1)\gamma(0)^{-1}\gamma(-1) \right) \]

The mean and the covariance of this conditional distribution are seen to coincide with the quantities \( \tilde{y}_{2|1} \) and \( \Sigma_2 \) given by the Kalman filter in theorem (2.1.1) if we initialize the filter as discussed above. The recursions are the same, the only difference being the interpretation of the quantities \( \tilde{y}_{t|t-1} \) and \( \Sigma_t \) as mean and variance of the conditional distributions in the Gaussian case. Of course, this implies that \( \tilde{y}_{t|t-1} \) is the best least squares predictor for \( y_t \) based on \( Y_{1:t}^T, U_1^T \).

Remark 2.3.4. Of course, any other initialization of the Kalman filter only yields approximations of the likelihood function (2.67). Choosing \( \hat{x}_{1|0} = 0 \) and \( P_1 = 0 \) implies

\[ \Sigma_{t|t-1}(\theta) = \Sigma(\sigma) \quad t = 1, 2, \ldots \]
\[ \hat{x}_{t+1|t}(\theta) = \hat{x}_{t+1|t-1}(l, k) + (A - KC) \hat{x}_{t|t-1}(l, k) + (B - KD) u_t + Ky_t \]
\[ e_t(\theta) = e_t(l, k) - C \hat{x}_{t|t-1}(l, k) - Du_t + y_t \]

This means that \( e_t(\theta) \) is given by the formal inverse of the system corresponding to \( \theta = (l, k, \sigma(\Sigma)) \) which will be denoted by

\[ (\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K}) = (A - KC, B - KD, -C, -D, K) \]

Clearly, \( e_t \) now only depends on \( (l, k) \). Note that (2.72),(2.73) follow directly from equations (2.22) and (2.23) below theorem (2.1.1) and may be rewritten as

\[ \hat{x}_{t+1|t}(l, k) = \bar{A} \hat{x}_{t|t-1}(l, k) + \bar{B} u_t + \bar{K} y_t \]
\[ e_t(l, k) = \bar{C} \hat{x}_{t|t-1}(l, k) + \bar{D} u_t + y_t \]

The prediction error form of the likelihood would thus be

\[ L_T(Y_{1:t}^T; U_1^T, \theta) = \log \det \Sigma(\sigma) + \frac{1}{2} \sum_{t=1}^{T} \text{tr} \left\{ e_t(l, k) e_t(l, k)' \Sigma^{-1}(\sigma) \right\} \]

\[ \text{tr} \]
Note, once more, that $\Sigma$ depends only on the covariance parameters $\sigma$ and $\epsilon_l$ depends only on system parameters for $(l, k)$ now. In the case where the Kalman filter is initialized differently (see, e.g., theorem (2.3.1)), both quantities $\Sigma_{l1t-1}(\theta)$ and $\epsilon_l(\theta)$ will be functions of covariance parameters $\sigma$ and system parameters for $(l, k)$ in general.

Remark 2.3.5. In the case that $\theta$ corresponds to a stable and strictly minimum phase transfer function, it follows from remark (2.1.14) and theorem (2.1.2) in section (2.1.2) that $\tilde{P} = 0$ is the unique solution to the Algebraic Riccati Equation (2.30) and that convergence of $P_{t+1}$ in (2.19) to $P = 0$ takes place exponentially fast for any symmetric and positive semidefinite $P_l$. Under these conditions (the true transfer function $(i_0, k_0)$ is stable and strictly minimum phase) it can be shown that the asymptotic properties of the estimates obtained by minimization of (2.70) do not depend on the particular initialization of $\hat{x}_{10}$ and $P_l$: Consistency (see section (2.3.4) and asymptotic normality with the same covariance matrix in the asymptotic normal distribution (see section (2.3.5)) hold true for any initialization. In particular, this implies that estimates obtained by minimization of (2.70) and (2.76) are asymptotically equivalent (in the sense discussed above) in this case. Note the difference in this respect to the case where $k_0$ has a zero of unit modulus; see the discussion of example (c) in section (2.1.3) and remark (2.3.17) in section (2.3.5) below.

Remark 2.3.6. If there are no cross restrictions between the unknown and unrestricted free parameters $\sigma(\Sigma)$ and the system parameters for $(l, k)$, the minimization of (2.70) yields the same estimates as the minimization of the concentrated likelihood function

$$L_c^c(Y_1^T; U_1^T, (l, k)) = \log \det \sum_{t=1}^T e_t(l, k)e_t(l, k)'$$

and putting $\Sigma_T = \frac{1}{T} \sum_{t=1}^T e_t(\hat{l}_t, \hat{k}_t)e_t(\hat{l}_t, \hat{k}_t)'$ where $(\hat{l}_T, \hat{k}_T)$ minimizes (2.77). This can be seen as follows: From (2.71) we get $\Sigma = \Sigma(\sigma(\Sigma))$ and from (2.75) we have $e_t(\theta) = e_t(l, k)$, and the two quantities can be parametrized independently according to our assumptions. Hence\(^\text{11}\),

$$\frac{\partial}{\partial \Sigma} \left( \log \det(\Sigma) + \frac{1}{T} \sum_{t=1}^T tr(e_t(l, k)e_t(l, k)'^{-1}) \right) = \Sigma^{-1} - \frac{1}{T} \sum_{t=1}^T (\Sigma^{-1} e_t(l, k)e_t(l, k)'\Sigma^{-1})'$$

and by setting (2.78) equal to zero, we obtain an estimate $\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^T e_t(l, k)e_t(l, k)'$, given the transfer function $(l, k)$. Substituting $\hat{\Sigma}_T$ back in (2.76) yields the concentrated likelihood

$$L_c^c(Y_1^T; U_1^T, (l, k)) = \log \det \left( \frac{1}{T} \sum_{t=1}^T e_t(l, k)e_t(l, k) + \frac{1}{T} \sum_{t=1}^T tr \left( e_t(l, k)e_t(l, k)' \left( \frac{1}{T} \sum_{t=1}^T e_t(l, k)e_t(l, k)' \right)^{-1} \right) \right)$$

$$= \log \det \left( \frac{1}{T} \sum_{t=1}^T e_t(l, k)e_t(l, k)' \right) + s$$

and by neglecting the constant terms, we get (2.77).

\(^{11}\text{Note that } \frac{\partial}{\partial \Sigma} \log \det(\Sigma) = \Sigma^{-1} \text{ and } \frac{\partial}{\partial \Sigma} tr(\Sigma^{-1}) = -\Sigma^{-1} A \Sigma^{-1}.\)
2.3.3 Separable least squares for approximate ML estimation

The starting point of this section is the criterion function given in (2.77) above. However, we will now assume that the transfer function \((l, k)\) is described by some vector of free parameters, \(\tau\) say. Using the state-space framework, we could therefore write \((A, B, C, D, K) = (A(\tau), B(\tau), C(\tau), D(\tau), K(\tau))\).

Let us now consider a modified criterion function (as compared to (2.77)) of the following form:

\[
\tilde{L}_T(Y^T_1; U^T, \tau) = tr \sum_{t=1}^{T} e_t(\tau)e_t(\tau)'\Sigma^{-1}
\]  

(2.79)

where \(\Sigma\) is fixed, but is known to be a consistent estimate for the true innovations covariance \(\Sigma_0\) (obtained, e.g. by subspace estimation beforehand).

We continue with a few remarks concerning this criterion function:

**Remark 2.3.8.** It is well known that estimates obtained by minimizing the concentrated likelihood function (2.77) are asymptotically equivalent (in the sense discussed in remark (2.3.5) above) to estimates obtained by minimizing (2.79) if \(\Sigma = \Sigma_0\); see, e.g. (Ljung, 1999) or (Söderström and Stoica, 1989). In fact, this situation amounts to having a priori knowledge of the true innovations covariance, and it follows immediately that (2.79) yields the exact maximum likelihood estimate as would be obtained from (2.70) in this case.

**Remark 2.3.9.** If, on the other hand, one replaced the consistent estimate \(\hat{\Sigma}\) in (2.79) by the identity matrix (as it is often seen in applications), then (2.79) would definitely yield a worse estimate as compared to the one that would be obtained from (2.76) or (2.77) in the sense that the covariance matrix in the asymptotic distribution would be strictly larger. This is true except for the case where the true innovations covariance happens to be a scalar multiple of the identity matrix. Note that this in particular means that in the single-output case, the estimates are indeed asymptotically equivalent which is clear because for \(s = 1\) (2.77) and (2.79) become \(\log \text{det} \sum_{t=1}^{T} e_t(l, k)^2 = \log \sum_{t=1}^{T} e_t(l, k)^2\) and \(tr \sum_{t=1}^{T} e_t(l)^2 = \sum_{t=1}^{T} e_t(l)^2\), respectively, clearly yielding the same estimates.

**Remark 2.3.10.** Finally, we conjecture (without having a proof yet) that using a consistent estimate \(\hat{\Sigma}\) in (2.79) suffices to guarantee that the estimates obtained by minimizing (2.79) are indeed asymptotically equivalent (in the sense discussed in remark (2.3.5) above) to the estimates obtained from (2.76) or (2.77).

In practice, (2.79) is minimized by some iterative algorithm and we could use this fact to further modify (2.79) slightly: Assume that \(\hat{\tau}\) is obtained in the previous iteration of the estimation algorithm. Then we could calculate \(\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} e_t(\hat{\tau})e_t(\hat{\tau})'\) and consider \(\hat{\Sigma}\) to be fixed for the next iteration which again consists of minimizing (2.79). For such a procedure, \(\hat{\Sigma}\) would be fixed in each iteration, but clearly change from one iteration to the other\(^{12}\). However, we will not pursue this possibility here and rather consider \(\Sigma\) to be fixed throughout the optimization procedure.

When referring to (2.79), we implicitly make the assumptions described in remarks (2.3.4) and (2.3.6). Hence, the quantities \(e_t(\tau)\) in (2.79) can be rewritten as (see (2.71) and (2.75)):

\[
\begin{align*}
e_1 &= \tilde{D}u_1 + y_1 \\
e_2 &= \tilde{C}x_{2\tau} + \tilde{D}u_2 + y_2 = \tilde{C}\tilde{B}u_1 + \tilde{C}\tilde{K}y_1 + \tilde{D}u_2 + y_2 \\
e_3 &= \tilde{C}x_{3\tau} + \tilde{D}u_3 + y_3 = \tilde{C}\tilde{A}\tilde{B}u_1 + \tilde{C}\tilde{A}\tilde{K}y_1 + \tilde{C}\tilde{B}u_2 + \tilde{C}\tilde{K}y_2 + \tilde{D}u_3 + y_3 \\
&\vdots
\end{align*}
\]

where the dependence of \(e_t\) on \(\tau\) is not made explicit in the sequel. Hence,

\[
\begin{pmatrix}
e_1 \\
e_2 \\
e_3 \\
\vdots \\
e_r
\end{pmatrix} =
\begin{pmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_r
\end{pmatrix} -
\begin{pmatrix}
\tilde{D} & \tilde{C}\tilde{B} & \tilde{D} \\
\tilde{C}\tilde{A}\tilde{B} & \tilde{C}\tilde{B} & \tilde{D} \\
\tilde{C}\tilde{A}^{r-2}\tilde{B} & \tilde{C}\tilde{B} & \tilde{D}
\end{pmatrix} \begin{pmatrix}
-u_1 \\
u_1 \\
-u_2 \\
u_2 \\
-u_r \\
u_r
\end{pmatrix} -
\begin{pmatrix}
0 & \tilde{C}\tilde{B} & \tilde{D} \\
\tilde{C}\tilde{A}\tilde{B} & \tilde{C}\tilde{B} & \tilde{D} \\
\tilde{C}\tilde{A}^{r-2}\tilde{B} & \tilde{C}\tilde{B} & \tilde{D}
\end{pmatrix} \begin{pmatrix}
0 \\
\tilde{C}\tilde{K} & \tilde{D} \\
\tilde{C}\tilde{A}\tilde{K} & \tilde{D} \\
\tilde{C}\tilde{A}^{r-2}\tilde{K} & \tilde{D}
\end{pmatrix} \begin{pmatrix}
0 \\
\tilde{C}\tilde{K} & \tilde{D} \\
\tilde{C}\tilde{A}\tilde{K} & \tilde{D} \\
\tilde{C}\tilde{A}^{r-2}\tilde{K} & \tilde{D}
\end{pmatrix} \begin{pmatrix}
0 \\
u_1 \\
\tilde{C}\tilde{K} & \tilde{D} \\
\tilde{C}\tilde{A}\tilde{K} & \tilde{D} \\
\tilde{C}\tilde{A}^{r-2}\tilde{K} & \tilde{D}
\end{pmatrix} \begin{pmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_r
\end{pmatrix}
\]

\(^{12}\)In this sense, (2.79) would actually describe a family of criterion functions.
The right hand side of this equation can be rewritten as

\[
\begin{pmatrix}
v_1 \\
v_2 \\
v_3 \\
\vdots \\
v_r
\end{pmatrix}
= -\begin{pmatrix}
-u_1 \odot I_r \vec{D} \\
u_2 \odot I_r \vec{D} + (u_1 \odot C) \vec{B} \\
-u_3 \odot I_r \vec{D} + (u_2 \odot C) \vec{B} + (u_1 \odot C A) \vec{B} \\
\vdots \\
-u_r \odot I_r \vec{D} + \cdots + (u_1 \odot C A^{r-1}) \vec{B}
\end{pmatrix}
- \begin{pmatrix}
0 \\
-u_1 \odot I_r \vec{R} \\
u_2 \odot I_r \vec{R} + (u_1 \odot C) \vec{R} \\
-u_3 \odot I_r \vec{R} + (u_2 \odot C A) \vec{R} \\
\vdots \\
-u_r \odot I_r \vec{R} - \cdots - (u_1 \odot C A^{r-1}) \vec{R}
\end{pmatrix}
\]

such that we finally obtain the following form:

\[
\begin{pmatrix}
e_1 \\
e_2 \\
e_3 \\
\vdots \\
e_r
\end{pmatrix}
= \begin{pmatrix}
v_1 \\
v_2 \\
v_3 \\
\vdots \\
v_r
\end{pmatrix}
- \begin{pmatrix}
-u_1 \odot I_r \\
u_2 \odot I_r - (u_1 \odot C) \\
u_3 \odot I_r - (u_1 \odot C A) \\
\vdots \\
u_r \odot I_r - (u_1 \odot C A^{r-1})
\end{pmatrix}
\begin{pmatrix}
\vec{D} \\
\vec{B} \\
\vec{R}
\end{pmatrix}
\tag{2.80}
\]

Note that \(X \in \mathbb{R}^{I \times m+n(m+s)}\). Let us now make two assumptions: First, as mentioned above, assume that we are given a consistent estimate \(\hat{\Sigma}\) of the innovations covariance. We can then transform (2.80) to obtain

\[
(I_T \otimes \hat{\Sigma}^{-1/2}) E_T = (I_T \otimes \hat{\Sigma}^{-1/2}) Y_T - (I_T \otimes \hat{\Sigma}^{-1/2}) X (Y_T^{-1}, U_T, \hat{C}, \hat{A}) \begin{pmatrix}
\vec{D} \\
\vec{B} \\
\vec{R}
\end{pmatrix}
\tag{2.81}
\]

where \(\hat{\Sigma}^{-1/2}\) denotes the inverse of any square root of \(\hat{\Sigma}\). Secondly, let us now assume that we parametrize the inverse system \((\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{R})\) using a parameter vector of the form \(\tau = (\tau^o, \tau^u)\) where the parameters in \(\tau^o \in \mathbb{R}^d\) only appear in \((\hat{C}, \hat{A}) = (\hat{C}(\tau^o), \hat{A}(\tau^o))\) and the parameters in \(\tau^u \in \mathbb{R}^{m+n(m+s)}\) are simply taken to be the matrix entries of \((\hat{D}, \hat{B}, \hat{R})\). Then we have the following

**Lemma 2.3.2.** Let \(\hat{\tau}^o \in \mathbb{R}^d\), \(\hat{\Sigma}\) and the observations \(Y_1^T, U_1^T\) be given where \(T > (m+n)+mn/s\). Assume that \((\hat{C}(\tau^o), \hat{A}(\tau^o))\) are analytic functions of \(\tau^o\) and that the matrix \(\hat{X}(Y_1^T, U_1^T, \hat{C}(\tau^o), \hat{A}(\tau^o), \hat{\Sigma})\) in (2.81) has full column rank. Then \(\hat{X}\) has full column rank for all \(\tau^o \in T^o\) where \(T^o\) is an open and dense subset of \(\mathbb{R}^d\) and the unique \(\hat{\tau}^u = ((\vec{D})', (\vec{B})', (\vec{R})')'\) that solves the optimization problem

\[
\min_{\tau^u \in \mathbb{R}^{m+n(m+s)}} \hat{L}_{T^u}(Y_1^T, U_1^T, (\hat{\tau}^o, \tau^u))
\tag{2.82}
\]

where \(\hat{L}_{T^u}\) is given in (2.79) is given by

\[
\hat{\tau}^u = \begin{pmatrix}
\vec{D} \\
\vec{B} \\
\vec{R}
\end{pmatrix} = \left(\hat{X}' \hat{X}\right)^{-1} \hat{X}' Y_1^T
\tag{2.83}
\]

**Proof.** Clearly, the matrix \(\hat{X}(Y_1^T, U_1^T, \hat{C}(\tau^o), \hat{A}(\tau^o), \hat{\Sigma})\) is an analytic function of \(\tau^o\). Hence, the function \(\Delta_{\hat{X}} : \mathbb{R}^d \to \mathbb{R}\) attaching det\(X'X\) to \(\tau^o\) is also analytic (and therefore trivially continuous). The inverse image \(\Delta_{\hat{X}}^{-1}(\mathbb{R} \setminus \{0\})\) is an open subset of \(\mathbb{R}^d\). Moreover, as \(\Delta_{\hat{X}}(\hat{\tau}^o) \neq 0\), \(\Delta_{\hat{X}} = 0\) can only hold true on a thin subset of \(\mathbb{R}^d\), showing that \(T^o\) is open and dense in \(\mathbb{R}^d\). The fact that \(\hat{\tau}^u\) in (2.83) is the unique solution to the optimization problem above is clear as (2.82) can be rewritten as

\[
\min_{\tau^u \in \mathbb{R}^{m+n(m+s)}} \|Y_1^T - \hat{X} \cdot \tau^u\|_2^2
\tag{2.84}
\]

the unique solution \(\hat{\tau}^u\) of which is given by (2.83). \(\square\)
We can immediately use the lemma above to obtain a new criterion function from (2.79):

\[
L_T^{(2)}(Y'; U'_T, \tau^o) = \text{tr} \sum_{t=1}^{T} e_t(\bar{A}(\tau^o), \bar{C}(\tau^o)) e_t(\bar{A}(\tau^o), \bar{C}(\tau^o)) \Sigma^{-1}
\]  

(2.85)

Here, \(e_t(\bar{A}(\tau^o), \bar{C}(\tau^o))\) are the residuals in (2.80) with \((\text{vec}(\bar{D})'), (\text{vec}(\bar{B})'), (\text{vec}(\bar{K}'))'\) replaced by (2.83). The following result is obtained in (Golub and Pereyra, 1973); note that a critical point is a point where the gradient is zero:

**Lemma 2.3.3.** Let \(\hat{L}_T^{(2)}(Y'; U'_T, (\tau^o, \tau^w)) = \hat{L}_T^{(2)}(\tau^o, \tau^w)\) and \(L_T^{(2)}(Y'; U'_T, \tau^o) = L_T^{(2)}(\tau^o)\) be given, let the assumptions of lemma (2.3.2) be valid and let \(T^0 \subset \mathbb{R}^d\) be as described in lemma (2.3.2). Then the following statements hold true:

(i) If \(\hat{\tau}^o \in T^0\) is a critical point (or a global minimizer in \(T^0\)) of \(L_T^{(2)}(\tau^o)\) and \(\hat{\tau}^w\) is given by (2.83), then \((\hat{\tau}^o, \hat{\tau}^w)\) is a critical point of \(\hat{L}_T^{(2)}(\tau^o, \tau^w)\) (or a global minimizer for \(\tau^o \in T^0\)) and \(\hat{L}_T^{(2)}(\hat{\tau}^o, \hat{\tau}^w) = L_T^{(2)}(\hat{\tau}^o)\).

(ii) If \((\hat{\tau}^o, \hat{\tau}^w)\) is a global minimizer of \(\hat{L}_T^{(2)}(\tau^o, \tau^w)\) for \(\tau^o \in T^0\), then \(\hat{\tau}^o\) is a global minimizer of \(L_T^{(2)}(\tau^o)\) in \(T^0\) and \(L_T^{(2)}(\hat{\tau}^o) = L_T^{(2)}(\hat{\tau}^o, \hat{\tau}^w)\). Furthermore, if there is an unique \(\hat{\tau}^w\) among the minimizing pairs of \(L_T^{(2)}(\tau^o, \tau^w)\), then \(\hat{\tau}^w\) must satisfy (2.83).

**Proof.** Having shown lemma (2.3.2) above, (i) and (ii) are special cases of theorem 2.1 in (Golub and Pereyra, 1973).

Note that instead of considering the original nonlinear optimization problem of finding minimizing arguments \((\tau, \sigma) = ((\tau^o, \tau^w), \sigma)\) for (2.76) we now obtain

- a nonlinear optimization problem in the parameters \(\tau^o\) (for \(L_T^{(2)}(\tau^o)\)), followed by
- a generalized linear least squares problem in \(\tau^w = (\text{vec}(\bar{D})'), (\text{vec}(\bar{B})'), (\text{vec}(\bar{K}'))'\) (the solution of which is given analytically) and
- a final minimization problem in \(\Sigma\) (the solution of which is given analytically).

In the literature, approaches of this type are termed **separable least squares methods** and are known for a long time; see, for instance, (Golub and Pereyra, 1973) for a general treatment.

The applicability of separable least squares methods clearly depends on the structure of the criterion function and the parametrization that is used. In a system identification context, we could clearly also apply this idea for a criterion function of the form \(P_T(Y'; U'_T, \tau) = \text{tr} \sum_{t=1}^{T} e_t(\tau) e_t(\tau)'\) where the second step of determining \(\bar{D}, \bar{B}\) and \(\bar{K}\) for given \(\bar{A}\) and \(\bar{C}\) becomes a linear least squares problem (instead of the generalized linear least squares problem above); see, e.g., (Bruel et al., 1997). As mentioned already, the corresponding estimates are known to lack the property of asymptotic efficiency if the true innovations covariance is not a scalar multiple of the identity matrix. Note, however, that the generalized least squares step as given in (2.83) is in general not an optimal concentration step for the concentrated likelihood in (2.77)! Consider, for instance, the following example: Put \(n = 1, s = 2\) and \(m = 0\) and

\[
\bar{a} = 0 \quad \bar{C} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \bar{\sigma}^2 = 1 \quad Y^3_i = (1,1,1,0,1,0)'
\]

Then

\[
\bar{X} = \bar{X}(Y^2_1, \bar{C}, \bar{a}, \bar{\sigma}^2) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad Y^3_i - \bar{X}_{\tau^w} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -(\bar{K}_1 + \bar{K}_2) \\ 1 \\ 1 - \bar{K}_1 \end{pmatrix}
\]
Clearly, \( \bar{X} \) has full column rank and

\[
\hat{\tau}^u = \bar{K} = (\bar{X}'\bar{X})^{-1} \bar{X}'Y_i^3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}
\]

We show that this \( \bar{K} \) does not satisfy the first order conditions for (2.77), or, equivalently for

\[
L_T = \det \sum_{i=1}^{3} e_i e_i' = \det \left( \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -(\bar{K}_1 + \bar{K}_2) \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right)
\]

\[
= \det \left( 1 + 3 \bar{K}_1 \bar{K}_2 + (1 - \bar{K}_1) \left( 1 - (\bar{K}_1 + \bar{K}_2) \right) \right)
\]

\[
= 6 + 6\bar{K}_1^2 + 6\bar{K}_1 \bar{K}_2 + 3\bar{K}_2^2 - 6\bar{K}_1 - (2 - 2\bar{K}_1 - \bar{K}_2)^2
\]

The first order conditions are

\[
\frac{\partial L_T}{\partial \bar{K}_1} = 12\bar{K}_1 + 6\bar{K}_2 - 6 + 4(2 - 2\bar{K}_1 - \bar{K}_2) \Rightarrow \frac{\partial L_T}{\partial \bar{K}_1}(1, -1) = 4 \neq 0
\]
\[
\frac{\partial L_T}{\partial \bar{K}_2} = 6\bar{K}_1 + 6\bar{K}_2 + 2(2 - 2\bar{K}_1 - \bar{K}_2) \Rightarrow \frac{\partial L_T}{\partial \bar{K}_2}(1, -1) = 2 \neq 0
\]

Hence, the generalized least squares step as given in (2.83) cannot be applied to (2.77) in general. Clearly, the single-output case \( (s = 1) \) is again an exception because (2.83) is easily seen to be optimal for both criterion functions (2.77) and (2.79).

### 2.3.4 Consistency

We recall that the consistency result can be formulated in a coordinate free manner. This is done in the following

**Theorem 2.3.4 (Coordinate-free consistency).** Let the following assumptions hold true:

1. \( \theta_0 \in \Theta^* \)
2. \( (\varepsilon_t) \) satisfies (2.58) and (2.59).
3. \( (u_t) \) is asymptotically stationary; see (2.57).
4. \( (\varepsilon_t) \) and \( (u_t) \) are wholly independent (see the brief discussion below (2.57)).
5. The persistent exciting condition (2.63) holds true.
6. The set \( U \) has finite degree, i.e. there exists a natural number \( n \in \mathbb{N} \) such that \( U \subset \mathbb{M}(n) \).
7. For \( s > 1 \), the matrix \( \Gamma_u = [\gamma_u(i-j)]_{i,j=1,...,sp} \) has full rank.

Then the maximum likelihood estimators \( \hat{\theta}_T \) over \( \hat{\Theta} \) are strongly consistent, i.e. \( \lim_{T \to \infty} \hat{\theta}_T = \theta_0 \) a.s.

**Proof.** We will only present the idea of the proof. It comes in several steps:
1. Consider the function

\[
L(\theta) = \log \det \Sigma + \frac{1}{2\pi} \int_{-\pi}^{\pi} \text{tr} \left\{ \left( k \Sigma k^* \right)^{-1} \left( k_0 \Sigma_0 k_0^* \right) \right\} d\omega + \int_{-\pi}^{\pi} \frac{\text{tr} \left\{ \left( k \Sigma k^* \right)^{-1} (l - l_0) dF_u(\omega) (l - l_0)^* \right\}}{Q(\theta)} 
\]

(2.86)

where the arguments \( e^{i\omega} \) for \( k, l, k_0 \) and \( l_0 \) have been omitted. Note that (2.86) depends on both \( \theta = (l, k, \sigma(\Sigma)) \) and \( \theta_0 = (l_0, k_0, \sigma(\Sigma_0)) \). This function will be shown to be the asymptotic form of the likelihood. Clearly, the contribution of the second term in \( Q(\theta) \) depends on the form of the input: If \( u_t \) is a sum of sinusoids of fixed (and finitely many) frequencies, for instance, then the only contribution to \( L(\theta) \) will be from the frequencies being present in \( u_t \).

Moreover, note that

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} \text{tr} \left\{ \left( k \Sigma k^* \right)^{-1} \left( k_0 \Sigma_0 k_0^* \right) \right\} d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} \text{tr} \left\{ k^{* -1} \Sigma^{-1} (k - k_0) \right\} d\omega \\
= \frac{1}{2\pi} \int_{-\pi}^{\pi} \text{tr} \left\{ \Sigma^{-1} (k - k_0) \right\} E e_i \epsilon_i^t (k - k_0)^* d\omega \\
= \frac{1}{2\pi} \text{tr} \left\{ \Sigma^{-1} \int_{-\pi}^{\pi} e^{i\omega} (k - k_0) E e_i \epsilon_i (k - k_0)^* d\omega \right\}
\]

where \( \Sigma_0 = E e_i \epsilon_i^t \). Applying the same argument to the second term of \( Q(\theta) \), we see that the asymptotic likelihood can also be written as

\[
L(\theta) = \log \det \Sigma + E e_i^t \epsilon_i, \quad e_i = k^{-1}(z) (k_0(z) \epsilon_i + (l(z) - l_0(z)) u_t)
\]

(2.87)

where the expression for \( e_i \) in (2.87) is a bit sloppy because we have to assume that \( (u_t) \) is extended on the negative integers such that it is already in ”stationary form” in formula (2.87).

2. We have the following results for the asymptotic likelihood:

\[
\theta_0 \in \tilde{\Theta}, \theta \in \tilde{\Theta} \quad \Rightarrow \quad L(\theta) > \log \det \Sigma_0 + s \quad \text{if} \quad \theta \neq \theta_0 \\
\theta_0 \in \Theta^* \quad \Rightarrow \quad L(\theta) = \log \det \Sigma_0 + s \quad \text{if} \quad \theta = \theta_0
\]

(2.88)

This shows that the asymptotic likelihood is uniquely minimized at \( \theta = \theta_0 \) if the true \( (l_0, k_0, \sigma(\Sigma_0)) \) is in \( \Theta^* \).

3. The first term of the finite sample likelihood given in (2.67) is shown to converge for \( T \to \infty \):

\[
\theta \in \tilde{\Theta} \Rightarrow \frac{1}{T} \log \det \Gamma_T(\theta) \to \log \det \Sigma \\
\text{and} \quad \frac{1}{T} \log \det \Gamma_T(\theta) \geq \log \det \Sigma \quad \forall T \in \mathbb{N}
\]

Note that the term \( \frac{1}{T} \log \det \Gamma_T(\theta) \) depends only on \( k \) and \( \Sigma \) and is independent of the data; see (2.65), (1.12) and (1.13). Moreover, the limit turns out to be independent of the system parameters for \( k \) and is only depending on \( \sigma(\Sigma) \).
4. The second term of the finite sample likelihood (2.67) contains the observations \( y_t \) and the exogenous inputs \( u_t \) and is therefore clearly a random variable. It can also be shown to converge for \( T \to \infty \):

\[
\theta_0 \in \hat{\Theta}, \theta \in \Theta^* \Rightarrow \frac{1}{T}[y_1^T - \hat{\gamma}_T^T(\theta)][\Gamma_T^{-1}(\theta)[y_1^T - \hat{\gamma}_T^T(\theta)]] \to Q(\theta) \quad \text{a.s.} \quad (2.89)
\]

Note that this convergence is only pointwise in the parameter space. It is rather involved to find a compact subset of the parameter space \( \hat{\Theta} \) where the following holds:

- The convergence in (2.89) is uniform within the compact parameter subset.
- The likelihood estimate \( \hat{\theta}_T \) enters this compact subset from a certain \( T \geq T_0 \) onwards.

Once these main ingredients are available, consistency can be shown rather quickly. For a detailed proof see theorem 4.2.1 in (Hannan and Deistler, 1988).

\[ \square \]

**Remark 2.3.11.** Although the Gaussian likelihood is considered, the result also holds true for other distributions of \( \varepsilon_t \) (with finite second moments).

**Remark 2.3.12.** For fixed, but large enough \( T \), it can be shown that if \( \theta \in \hat{\Theta} \) corresponds to \( (l, k, \sigma(\Sigma)) \) where \( \Sigma \) is singular or \( k(z) \) has a pole on the unit circle, then \( L_T(\theta_n) \to \infty \) if \( \theta_n \to \theta, \theta_n \in \hat{\Theta} \). In other words, such boundary points can be avoided in the course of the optimization process.

**Remark 2.3.13.** Note that boundary points corresponding to zeros of \( k(z) \) on the unit circle cannot be avoided. In fact, simulation studies showed that estimates in \( \hat{\Theta} \setminus \Theta^* \), i.e. estimates with zeros of unit modulus, will be obtained with positive probability in finite samples even if \( \theta_0 \in \Theta^* \), but \( k_0(z) \) has zeros close to the unit circle; see (Poetscher, 1987), for instance.

### 2.3.5 Asymptotic normality

In this section we present the most important results on asymptotic normality. Note that \( \theta \) still denotes the coordinate-free parameter of the form \( (l, k, \sigma(\Sigma)) \), whereas the coordinatized form of \( (l, k, \sigma(\Sigma)) \) is denoted by the real-valued vector \( (\tau', \sigma')' \). Here, \( \tau \) denotes the system parameters for \( (l, k) \); see chapter 4 for a number of different parametrizations of \( (l, k) \).

**Theorem 2.3.5 (Central Limit Theorem, CLT).** Let the conditions of theorem (2.3.4) hold true and assume that the following strengthened conditions are satisfied:

- \( \Theta = \mathbb{M}(n) \times \Sigma \), i.e. the order of the transfer function is fixed and no cross restrictions between the system parameters \( \tau \) and the covariance parameters \( \sigma \) are imposed.
- \( \theta_0 \in \Theta^* \cap (\mathbb{M}(n) \times \Sigma) \), i.e. the true parameter value does not correspond to a system of lower McMillan degree (this may well be the case for systems in \( \Theta^* \)).
- The white noise process is ergodic, i.e. the \( \sigma \)-algebra of invariant events is trivial; see, e.g. the first chapter in (Hannan and Deistler, 1988) for details.

Additionally, assume that

- The exogenous input fulfills

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} u_t = \mu_u \quad (2.90)
\]

- The white noise process fulfills

\[
E(\varepsilon_1|\mathcal{F}_{t-1}) = 0 \]

where \( \mathcal{F}_t \) denotes the sigma algebra generated by \( \varepsilon_s, s \leq t \). Note that \( \varepsilon_t \) denotes the innovations process, i.e. \( \varepsilon_t = y_t - \hat{y}_t|_{t-1} \) is the error of the linear prediction. In fact, the condition above is necessary and sufficient for the best linear mean squares predictor of the stationary process \( (y_t) \) to be the best mean squares predictor and thus amounts to absence of non-linearities.
2.4. FIGURES

The white noise process \((\varepsilon_t)\) additionally satisfies

\[
\lim_{k \to \infty} \mathbb{E}(\varepsilon_t \varepsilon'_t | \mathcal{F}_{t-k}) = \Sigma \quad \text{a.s.} \tag{2.91}
\]

\[
\lim_{k \to \infty} \mathbb{E}(\varepsilon_t^{(a)} \varepsilon_t^{(b)} \varepsilon_{t-r}^{(c)} | \mathcal{F}_{t-k}) = \mathbb{E}(\varepsilon_t^{(a)} \varepsilon_t^{(b)} \varepsilon_{t-r}^{(c)}) = \sigma_{abc}(r) \quad \text{a.s.} \tag{2.92}
\]

\[
\mathbb{E}(\varepsilon_t^{(a)})^4 < \infty \tag{2.93}
\]

where \(r > 0, a, b, c = 1, \ldots, s\) and \(\varepsilon_t^{(a)}\) denotes the \(a\) component of the random vector \(\varepsilon_t\); trivially, \(\varepsilon_t^{(b)}\) etc. are given analogously. The conditions imply that \(\varepsilon_t\) has finite moments up to fourth order. Note that (2.91) and (2.92) can be viewed as requirements of (nonlinear) nondeterminism, that is, that apart from their mean values there is nothing from the infinitely far past influencing \(\varepsilon_t \varepsilon'_t\) or \(\varepsilon_t^{(a)} \varepsilon_t^{(b)} \varepsilon_{t-r}^{(c)}\).

Then

\[
\sqrt{T}(\hat{\tau}_T - \tau_0) \to^d N(0, I(\tau_0)) \tag{2.94}
\]

where \(\to^d\) denotes convergence in distribution. If (2.91), (2.92) and (2.93) are replaced by the assumption

\[
\mathbb{E}(\varepsilon_t \varepsilon'_t | \mathcal{F}_{t-1}) = \Sigma \tag{2.95}
\]

then the covariance matrix \(I(\tau_0)\) in (2.94) has to be replaced by \(I(\tau_0)\) which is the Fisher information matrix as for the Gaussian case.

Proof. See theorem 4.3.1 in (Hannan and Deistler, 1988).

A few remarks conclude this short summary:

Remark 2.3.14. The theorem also holds true for \(\Theta = V \times \Sigma\) where \(V\) denotes some subset of \(\mathbb{M}(n)\) corresponding to any of the parametrizations in chapter 4 as long as \(\theta_0\) is an inner point of \(V \times \Sigma\).

Remark 2.3.15. If the theorem is to hold for the whole parameter vector \((\tau, \sigma)^t\) (and not only for the system parameters), then (2.91) and (2.92) have to be replaced by the stronger assumption (2.95) in any case. Now if \(\mathbb{E}(\varepsilon_t^{(a)} \varepsilon_t^{(b)} \varepsilon_t^{(c)} | \mathcal{F}_{t-1}) = \sigma_{abc}(0)\) is added, we get

\[
\sqrt{T} \left( \begin{array}{c} \hat{\tau}_T \\ \hat{\sigma}_T \end{array} \right) - \left( \begin{array}{c} \tau_0 \\ \sigma_0 \end{array} \right) \to^d N(0, I(\tau_0, \sigma_0))
\]

with a covariance matrix as in the Gaussian case.

If the latter condition is weakened to \(\lim_{k \to \infty} \mathbb{E}(\varepsilon_t^{(a)} \varepsilon_t^{(b)} \varepsilon_{t-k}^{(c)} | \mathcal{F}_{t-k}) = \sigma_{abc}(0)\), then the CLT still holds true for \((\tau', \sigma')^t\), but the covariance matrix will be more complicated.

Remark 2.3.16. The theorem can be generalized to the case where (twice continuously differentiable) constraints on the true parameter value \(\theta_0\) are imposed. In this case one must distinguish between the case where each constraint is a function of either \(\tau\) alone or of \(\sigma\) alone ("constraints do not mix parameters") and the case where this does not hold; see (Hannan and Deistler, 1988), chapter 4.

Remark 2.3.17. Note that if \(\theta_0\) corresponded to a system with zeros of unit modulus, the limiting distribution would not be Gaussian any more. Consider the ARMA(1,1) case \(y_t + ay_{t-1} = \varepsilon_t + b\varepsilon_{t-1}\) with \(\tau = (a, b) \in T = (-1, 1) \times [-1, 1]\), for instance. Then \(\tau_0 = (a, 1)\) implies that the second component of the estimation error \(\hat{\tau}_T - \tau_0\) cannot be positive which then trivially also holds true for \(\sqrt{T}(\hat{\tau}_T - \tau_0)\).

2.4 Figures
Figure 2.1: Plot (1) shows a typical trajectory of the input and output series. The plots labelled (a,1) and (a,2) show the performance of the Kalman filter for stable and minimum phase models.
Figure 2.2: The plots labelled (b,1) and (b,2) show the performance of the Kalman filter for stable, nonminimum phase models without zeros of unit modulus, whereas the plots labelled (c,1) and (c,2) show the performance of the Kalman filter for stable and minimum phase models with zeros of unit modulus.
Chapter 3

Manifolds in the space $S(n)$

In this chapter, a number of different submanifolds of $S(n)$ is investigated; recall from section (1.4) that $S(n) = \mathbb{R}^{n^2 + 2n + m(n + s)}$ is the Euclidean space obtained by stacking the entries of the system matrices $(A, B, C, D, K)$ on top of each other. For the general definition of real analytic manifolds and submanifolds, see the appendix (A.6). Note that each of the manifolds treated in this chapter is easily seen to be a submanifold of $S(n)$. Hence, the tangent spaces of these submanifolds are subspaces of the tangent space to $S(n)$, and the latter clearly coincides with $S(n)$ itself because $S(n)$ is a Euclidean vector space.

The results of this chapter give us some idea of the geometry of certain sets of state-space realizations, but will also be of importance for the full state-space parametrization and the data driven local parametrizations introduced in chapter 4 below.

For completeness, we will present two proofs of well known and frequently needed results on real non singular and orthogonal matrices in section (3.1). In section (3.2) the class of observationally equivalent systems $E(A, B, C, D)$ will then be treated. Section (3.3) treats a subset of $E(A, B, C, D)$, namely the set of all $(A, B, C, D, K)$ having minimum Frobenius norm. Observationally equivalent balanced stable allpass systems $(A, B, C, D)$ will be dealt with in section (3.4), and section (3.5) enlarges this class by not restricting the balanced stable allpass systems $(A, B, C, D)$ to be observationally equivalent. Finally, the manifold of $L_2^G$-equivalent systems will be introduced and investigated in section (3.6), and a number of illustrative plots in section (3.7) concludes this chapter.

3.1 Non singular and orthogonal matrices

Let us start with two lemmas which will be needed frequently in this chapter. Note that $GL(n)$ denotes the set of nonsingular real $n \times n$ matrices and $O(n) \subset GL(n)$ denotes the set of real orthogonal $n \times n$ matrices. Connectivity always refers to pathwise connectivity in this thesis:

Lemma 3.1.1. The set $O(n)$ consists of two disconnected compact components containing the identity $I_n$ and $\text{diag}(I_{n-1}, -1)$, respectively.

Proof. For $n = 1$ this is trivial because $O(1) = \{1, -1\}$. For $n = 2$ one can write $U = (u_{i,j}), i, j \in \{1, 2\}$. As $U^T U = I$, one can set $u_{11} = \cos(\phi)$ and $u_{12} = \sin(\phi)$ and thus gets $u_{21} = \pm u_{12}$ and $u_{22} = \mp u_{11}$. Hence,

$$O(2) = \left\{ \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix} \right\}, \phi \in [0, 2\pi]$$

Note that matrices of the first form are rotations ($det = 1$) which can be connected through a (continuous) path within $O(2)$ to $I_2$ whereas matrices of the second type are products of a rotation and a reflection.
(\text{det} = -1) which can be connected through a (continuous) path within \(O(2)\) to \(\text{diag}(1, -1)\). Now for an arbitrary orthogonal matrix \(U = (u_{ij}), j i \in \{1, \ldots, n\} \in O(n)\) one can always find a matrix \(R_{12} = R_{12}(\phi_1)\), which performs a rotation of the angle \(\phi_1\) in the plane spanned by the first and second coordinate axis, such that \(u_{21} = 0\):

\[
\begin{pmatrix}
\cos(\phi_1) & \sin(\phi_1) \\
-\sin(\phi_1) & \cos(\phi_1)
\end{pmatrix}
\begin{pmatrix}
1 \\
\ldots \\
1
\end{pmatrix}
\]

\[
U = \begin{pmatrix}
  \text{diag}(1, \ldots, 1) & * & \ldots & * \\
  0 & * & \ldots & * \\
  * & \ldots & * \\
  * & \ldots & *
\end{pmatrix}
\]

where * denotes arbitrary entries. The angle \(\phi_1\) is then given by \(\tan^{-1}(\frac{\text{sum}}{u_{11}}) \in [-\frac{\pi}{2}, \frac{\pi}{2}]\). Note that \(\{R_{12}(\phi)|\phi \in [0, \phi_1]\}\) is a (continuous) path within \(O(n)\) connecting \(U\) and \(\tilde{U}_{12}\). Next, one premultiplies by \(R_{13}(\phi_2)\) (which is defined accordingly; trigonometric entries appear in the (1, 1), (3, 1), (1, 3) and (3, 3) positions) and chooses \(\phi_2\) such that the (3, 1) element cancels out. This will leave the second row unchanged. Writing \(R_{1n}(\phi_{n-1})\ldots R_{13}(\phi_2)R_{12}(\phi_1)U = R_{1n}U = U_1 \in O(n)\) finally yields a matrix with a first column of the form \((\tilde{u}_{11}, 0, \ldots, 0)'\). Note that \(\tilde{u}_{11} = \pm 1\) because \(\tilde{U}_1 \in O(n)\). If \(\tilde{u}_{11} = -1\), one applies a final rotation \(R_{12}(\pi)\) in order to get \(\tilde{u}_{11} = 1\). Clearly, the first row must then also be zero everywhere except for the (1, 1) position. Next, \(R_{2n}(\rho_{n-2})\ldots R_{24}(\rho_2)R_{23}(\rho_1)U_1 = R_{2\tilde{U}_1}\), and after \(n-1\) iterations one gets \(R_{n-1}\ldots R_2R_1U = \text{diag}(I_{n-1}, \tilde{u}_{nn})\) with \(\tilde{u}_{nn} = \pm 1\) (because the product is again an element of \(O(n)\)), dependent on whether \(\text{det}(U) = 1\) or \(\text{det}(U) = -1\). It is evident that there is no (continuous) path within \(O(n)\) between \(I_n\) and \(\text{diag}(I_{n-1}, -1)\). To show the compactness of the two disconnected components, observe that the entries of any \(U \in O(n)\) are bounded because each column (row) has unit length. Moreover, each of the two components is closed (in \(\mathbb{R}^n\)) because the complement of each component is given by the open sets \(\text{det}^{-1}(\mathbb{R} \setminus \{1\})\) and \(\text{det}^{-1}(\mathbb{R} \setminus \{-1\})\), respectively.

\textbf{Lemma 3.1.2.} The set \(GL(n)\) consists of two open and disconnected components containing the identity \(I_n\) and \(\text{diag}(I_{n-1}, -1)\), respectively.

\textbf{Proof.} First, consider the continuous mapping

\[
\text{det} : \mathbb{R}^{n \times n} \to \mathbb{R}, \\
T \mapsto \text{det}(T)
\]

Clearly, \(GL(n) = \text{det}^{-1}((0, \infty)) \cup \text{det}^{-1}((0, 0))\) and these two sets are disjoint and open (as inverse images of open sets). Hence, \(GL(n)\) consists of at least two open and disconnected components. To show that there are exactly these two components, consider the singular value decomposition of an arbitrary \(T \in GL(n)\): \(T = U_1 \Sigma V_1'\). Here, \(\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)\) with \(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0\) and \(U_1, V_1 \in O(n)\). The singular values \(\sigma_i, i = 1, \ldots, n\) are unique. Let us assume that \(U_1 \in O(n)\) is fixed. The corresponding \(V_1 \in O(n)\) will then be uniquely determined. Clearly, then there is a (continuous) path in \(GL(n)\) from \(\Sigma\) to the identity \(I_n\). It therefore remains to show that for any \(U \in O(n)\), there exists a (continuous) path within \(O(n)\) from \(U = U_1V_1'\) to either the identity \(I_n\) or to \(\text{diag}(I_{n-1}, -1)\). Hence, use of lemma (3.1) completes the proof.

\section{3.2 The manifold of observationally equivalent systems}

Let us first recall the definition of the \((l, k)\)-equivalence classes \(E(A, B, C, D)\); see definition (1.4.4). The lemma below clarifies the structure of \(E(A, B, C, D)\); for a definition of real analytic manifolds see definition (A.6.1) in the appendix (A.6):
Lemma 3.2.1 (Structure of the set $\mathcal{E}(A, \bar{B}, C, D)$). Let $(A, B, C, D, K)$ be a minimal system and let $\pi(A, B, C, D, K) = (l, k)$ be the corresponding transfer function. The $(l, k)$-equivalence class $\mathcal{E}(A, \bar{B}, C, D)$ constitutes a $n^2$ dimensional real analytic manifold consisting of two disconnected components and is given by

$$\mathcal{E}(A, \bar{B}, C, D) = \left\{ (TAT^{-1}, T\bar{B}, C\bar{T}^{-1}, D), \ T \in GL(n) \right\}$$

(3.1)

where $GL(n)$ denotes the set of non singular real $n \times n$ matrices.

Remark 3.2.1. For $n = s = 1$ and $m = 0$, the $k$-equivalence classes $\mathcal{E}(a, k, c)$ in $S(1)$ are (one dimensional) hyperbolae with two disconnected branches being determined by a fixed $a$ and $kc = \text{const}$; see plot (a) in figure (3.1) in section (3.7).

Proof. For a given minimal $(A, B, C, D, K)$, the set (3.1) clearly represents the same transfer function. This is because both systems $(A, B, C, D)$ and $(TAT^{-1}, T\bar{B}, C\bar{T}^{-1}, D)$ are easily seen to correspond to the same Hankel matrix $\mathcal{H}$. Note that $\mathcal{H} = \mathcal{O}C$ where $\mathcal{O}$ and $\mathcal{C}$ denote the infinite observability and controllability matrix, respectively; see remark (1.4.3) in chapter 1. Moreover, $\mathcal{H}$ is, together with the constant feedthrough matrix $D$, in a one-to-one correspondence with the transfer function $(l, k)$. Conversely, let two systems $(A, B, C, D, K)$ and $(\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K})$ correspond to the same transfer function $(l, k)$. Then, using an obvious notation, $\mathcal{H} = \mathcal{O}C = \bar{\mathcal{O}}\bar{C}$, and it is straightforward to construct a unique nonsingular $T = (\bar{\mathcal{O}}\mathcal{O})^{-1} \bar{\mathcal{O}}\mathcal{O}$ such that $(A, B, C, D, K) = (TAT^{-1}, T\bar{B}, C\bar{T}^{-1}, D, TK)$. Hence, we have shown that for any fixed minimal $(A, B, C, D, K)$ with $\pi(A, B, C, D, K) = (l, k)$ the mapping

$$\phi : GL(n) \rightarrow \mathcal{E}(A, \bar{B}, C, D)$$

$$T \mapsto \begin{pmatrix} \text{vec}(TAT^{-1}) \\ \text{vec}(TB) \\ \text{vec}(CT^{-1}) \\ \text{vec}(D) \end{pmatrix}$$

(3.2)

is bijective. Continuity of $\phi$ is trivial, and continuity of $\phi^{-1}$ follows from the continuity of the construction of a unique $T = (\bar{\mathcal{O}}\mathcal{O})^{-1} \bar{\mathcal{O}}\mathcal{O}$ above; in other words, $\phi$ is a homeomorphism. Hence, $(\mathcal{E}(A, B, C, D), \phi^{-1})$ is a system of local coordinates consisting of a single coordinate neighborhood and a single corresponding coordinate mapping. Note that $\phi^{-1}(\mathcal{E}(A, B, C, D)) = GL(n)$ is an open subset of $\mathbb{R}^{n^2}$ where $GL(n)$ is embedded into $\mathbb{R}^{n^2}$. This shows that $\mathcal{E}(A, \bar{B}, C, D)$ is a $n^2$ dimensional real analytic manifold.

For the second part, it suffices to show that the set $GL(n)$ consists of two disconnected components; see lemma (3.1.2) above. These two components are open in $\mathbb{R}^{n^2}$, their images under $\phi$ are open and disjoint subsets of the equivalence class ($\phi$ is a homeomorphism) and thus they are disconnected. However, each subset is itself connected as the image of a connected set under a continuous mapping is connected.

Remark 3.2.2. Using the relation

$$\text{vec}(XYZ) = (Z' \otimes X) \cdot \text{vec}(Y) \text{ where } X \otimes Y = \begin{pmatrix} X_{11}Y & \cdots & X_{1q}Y \\ \vdots & \ddots & \vdots \\ X_{p1}Y & \cdots & X_{pq}Y \end{pmatrix}$$

(3.3)

with $X \in \mathbb{R}^{p \times q}$ and $Y \in \mathbb{R}^{r \times s}$, the vectorized form of $\mathcal{E}(A, \bar{B}, C, D)$ in (3.1) becomes

$$\left\{ \begin{pmatrix} T^{-1} \otimes T & 0 & 0 & 0 \\ 0 & I_{m+s} \otimes T & 0 & 0 \\ 0 & 0 & T^{-1} \otimes I_s & 0 \\ 0 & 0 & 0 & I_m \otimes I_s \end{pmatrix}, \begin{pmatrix} \text{vec}(A) \\ \text{vec}(\bar{B}) \\ \text{vec}(C) \\ \text{vec}(D) \end{pmatrix}, \ T \in GL(n) \right\}$$

(3.4)
Note that $T_{vec}$ has full rank for all $T \in GL(n)$ as $rk(A \otimes B) = rk(A) \cdot rk(B)$.

Next, we will consider the tangent space to the \((l, k)\)-equivalence class in $S_m(n)$ at a given \((A, B, C, D, K)\). For a general definition of the tangent space to a real analytic manifold, see the appendix (A.6). We recall that the real analytic manifolds considered in this chapter are submanifolds of $S(n)$, and that the tangent spaces to these manifolds can thus be considered as linear subspaces of $S(n)$ itself. However, at the expense of being a bit imprecise, we will view all tangent spaces as affine spaces (instead of linear vector spaces) in the sequel. Hence, by $Q_{(A, B, C, D)}$ we denote the (affine) tangent space to $E(A, B, C, D)$ at $(A, B, C, D)$, and we proceed analogously for all tangent spaces considered in this chapter:

**Lemma 3.2.2 (Tangent space $Q_{(A, B, C, D)}$ to $E(A, B, C, D)$).** Let the system $(A, B, C, D, K)$ be minimal. Then $(A, B, C, D) + (A_s, B_s, C_s, D_s) \in Q_{(A, B, C, D)}$ if and only if $(A_s, B_s, C_s, D_s) = (\tilde{T}A - AT, \tilde{T}B, -CT, 0)$ for some $\tilde{T} \in \mathbb{R}^{n \times n}$.

**Remark 3.2.3.** From lemma (3.2.1) it is clear that the dimension of the affine subspace $Q_{(A, B, C, D)}$ is $n^2$. For $n = s = 1$ and $m = 0$, $Q_{(a, k, c)}$ is the one dimensional tangent line to the hyperbola at some minimal $(a, k, c)$; see plot (a) in figure (3.1) in section (3.7).

**Proof.** We have to show that $(A_s, B_s, C_s, D_s)$ is in the tangent space to the equivalence class given in (3.1) at $T = I$, i.e. at $(A, B, C, D)$. Let $T = T(\theta), \theta \in [0, 1]$ be a differentiable path in $GL(n)$ and let $\tilde{T} = \tilde{T}(\theta)$ be the derivative with respect to $\theta$ where $T(0) = I$. Then differentiation yields

$$\{(\tilde{T}AT^{-1} - TAT^{-1}T^{-1}, \tilde{T}B, -CT^{-1}T, 0), \tilde{T} \in \mathbb{R}^{n \times n}\}$$

and at $\theta = 0$, i.e. at the given minimal realization $(A, B, C, D)$, (3.5) reduces to

$$\{(\tilde{T}A - AT, \tilde{T}B, -CT, 0), \tilde{T} \in \mathbb{R}^{n \times n}\}$$

where $\tilde{T} = \tilde{T}(0)$ in this case. Hence, $(A_s, B_s, C_s, D_s)$ must have a representation of the form (3.6). □

**Remark 3.2.4.** Using (3.3) together with $(A \otimes B)(C \otimes D) = (AC \otimes BD)$, (3.5) can be vectorized to yield

$$\begin{align*}
\left\{ \begin{array}{l}
T^{-1}A' \otimes I_n - T^{-1} \otimes TAT^{-1} \\
- T^{-1} \otimes CT^{-1}
\end{array} \right\} \cdot vec(T), \quad T \in \mathbb{R}^{n \times n}
\end{align*}$$

\[
\begin{pmatrix}
T^{-1} \\
0
\end{pmatrix}
\begin{pmatrix}
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
T' & 0 & 0 & 0 \\
0 & I_{m \times n} & 0 & 0
\end{pmatrix}

\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}

\begin{pmatrix}
A' \otimes I_n - I_n \otimes A \\
- I_n \otimes A
\end{pmatrix}

\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}

\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}

\begin{pmatrix}
I_n \otimes T^{-1} \cdot vec(T), \quad T \in \mathbb{R}^{n \times n}
\end{pmatrix}

At $T = I$, (3.7) then becomes

$$\left\{ Q \cdot vec(T), \quad T \in \mathbb{R}^{n \times n}\right\}$$

which is the vectorization of (3.6).

By the ortho-complement $Q_{(A, B, C, D)}^\perp$ to the tangent space $Q_{(A, B, C, D)}$, being a bit sloppy, we mean the ortho-complement to the linear space corresponding to $Q_{(A, B, C, D)}$ shifted back by the initial system $(A, B, C, D)$. Dealing with affine spaces, the term ortho-complement is always used in this sense in the sequel. Of course, orthogonality is to be understood with respect to the inner product deriving from the Euclidean norm on $S(n)$:

**Lemma 3.2.3 (Ortho-complement $Q_{(A, B, C, D)}^\perp$).** Let the system $(A, B, C, D, K)$ be minimal. Then $(A, B, C, D) + (A_s, B_s, C_s, D_s) \in Q_{(A, B, C, D)}^\perp$ if and only if $A_sA' + B_sB' = A'A_s + C'C_s$. 

Remark 3.2.5. It is clear that the dimension of the affine subspace \( Q^+_{\alpha,b,c,d} \) is \( 2ns + m(n + s) \). For \( n = s = 1 \) and \( m = 0 \), \( Q^+_{\alpha,b,c,d} \) is the two dimensional affine plane being orthogonal to the tangent line described in remark (3.2.3); see plot (a) in figure (3.1) in section (3.7).

Proof. Clearly, from lemma (3.2.2), we get

\[
(A, \tilde{B}, C, D) + (A_{\bar{s}}, \tilde{B}_{\bar{s}}, C_{\bar{s}}, D_{\bar{s}}) \in Q^+_{\alpha,b,c,d} \Leftrightarrow \\
(A_{\bar{s}}, \tilde{B}_{\bar{s}}, C_{\bar{s}}, D_{\bar{s}}) \perp \{ (\tilde{T}A - A\tilde{T}, \tilde{T}\tilde{B}, -C\tilde{T}, 0), \quad \tilde{T} \in \mathbb{R}^{n \times n} \}
\]

Hence,

\[
\begin{align*}
tr\{A_{\bar{s}}(\tilde{T}A - A\tilde{T})'\} + tr\{\tilde{B}_{\bar{s}}(\tilde{T}\tilde{B})'\} - tr\{C_{\bar{s}}(C\tilde{T})'\} &= 0 \quad \forall \tilde{T} \in \mathbb{R}^{n \times n} \\
tr\{A_{\bar{s}}A'T'\} - tr\{A_{\bar{s}}A'T'\} + tr\{\tilde{B}_{\bar{s}}\tilde{B}'T'\} - tr\{C_{\bar{s}}C'T'\} &= 0 \quad \forall \tilde{T} \in \mathbb{R}^{n \times n} \\
tr\{A_{\bar{s}}A'T'\} - tr\{A'A_{\bar{s}}T'\} + tr\{\tilde{B}_{\bar{s}}\tilde{B}'T'\} - tr\{C'C_{\bar{s}}T'\} &= 0 \quad \forall \tilde{T} \in \mathbb{R}^{n \times n}
\end{align*}
\]

By considering matrices \( \tilde{T} \) with only one nonzero entry and by varying the position of this entry, it is immediately clear that the equation holds true if and only if \( A_{\bar{s}}A' + \tilde{B}_{\bar{s}}\tilde{B}' = A'A_{\bar{s}} + C'C_{\bar{s}} \). \( \square \)

### 3.3 The manifold of observationally equivalent MFN systems

Let us first note that MFN stands for minimum Frobenius norm. Defining the Frobenius norm of a realization \((A, \tilde{B}, C, D)\) by

\[
||(A, \tilde{B}, C, D)||_F^2 := tr\{AA'\} + tr\{\tilde{B}\tilde{B}'\} + tr\{CC'\} + tr\{DD'\}
\]  \hspace{1cm} (3.9)

we are able to give the following

**Definition 3.3.1 (MFN Realizations).** A minimal realization \((A, B, C, D, K)\) is said to be of minimum Frobenius norm or, briefly, MFN, if \(||(A, B, C, D)||_F \leq ||(TAT^{-1}, T\tilde{B}, C\tilde{T}^{-1}, D)||_F\) for all \((TAT^{-1}, T\tilde{B}, C\tilde{T}^{-1}, D) \in \mathcal{E}(A, B, C, D)\).

Next, a characterization of MFN realizations is given:

**Lemma 3.3.1 (Characterization of MFN Realizations).** Let the system \((A, B, C, D, K)\) be minimal. Then the following statements are equivalent:

(i) \((A, B, C, D, K)\) is MFN.

(ii) \(AA' + \tilde{B}\tilde{B}' = A'A + C'C\).

(iii) \((0, 0, 0, 0) \in Q^+_{\alpha,b,c,d}\).  

Proof. We have to consider the problem

\[
\min_{T \in GL(n)} ||(TAT^{-1}, T\tilde{B}, C\tilde{T}^{-1}, D)||_F
\]

or, equivalently

\[
\min_{T \in GL(n)} \underbrace{tr\{(TAT^{-1})(TAT^{-1})'\} + tr\{(T\tilde{B})(T\tilde{B})'\} + tr\{(C\tilde{T}^{-1})(C\tilde{T}^{-1})'\} + tr\{DD'\}}_{G(T)}
\]  \hspace{1cm} (3.10)
"(i) ⇒ (ii)": Note that \( G : \text{GL}(n) \to \mathbb{R} \) is a differentiable function. A necessary first order condition is given by
\[
tr\{TAT^{-1} - TAT^{-1}TT^{-1}\} + tr\{TAT^{-1}\}(-T^{-1}i'T^{-1}A'T + T^{-1}A'T) + \]
\[
tr\{TBB'T\} + tr\{TBB'T\} + tr\{-CT^{-1}i'T^{-1}C'\} + tr\{-CT^{-1}i'T^{-1}C'\} = 0 \quad \forall T \in \mathbb{R}^{n \times n}
\]
For the system \((A, B, C, D)\) to be MFN, this condition has to be satisfied at \(T = I\):
\[
tr\{TAA'\} - tr\{AT'A\} + tr\{AA'T\} + tr\{Bbb'T\} - tr\{CT'C\} = 0 \quad \forall T \in \mathbb{R}^{n \times n}
\]
\[
tr\{AA'T\} - tr\{AAT\} + tr\{AA'T\} + tr\{BB'T\} + tr\{BB'T\} - tr\{CT'C\} = 0 \quad \forall T \in \mathbb{R}^{n \times n}
\]
Now, using the same argument as at the end of the proof of lemma (3.2.3), this holds true if and only if \(AA' + BB' = A'A + C'C\).

"(i) ⇔ (ii)": Given a minimal system \((A, B, C, D)\), the condition \(AA' + BB' = A'A + C'C\) is also sufficient for the realization to be MFN; see (Helmkie and Moore, 1994). In fact, the authors use several complex variable theory (see theorem 5.1 on page 223 of (Helmke and Moore, 1994)) to deduce
- the existence of a minimal realization satisfying \(AA' + BB' = A'A + C'C\) and
- the result that a minimal \((A, B, C, D)\) minimizes \(G(T)\) if and only if \(AA' + BB' = A'A + C'C\).

"(ii) ⇔ (iii)": This follows readily from lemma (3.2.3) because \((0, 0, 0, 0) = (A, B, C, D) + (-A, -B, -C, -D) \in Q_{(A, B, C, D)}^n\) if and only if \(AA' + BB' = A'A + C'C\).

The next definition deals with the set of observably equivalent MFN systems:

**Definition 3.3.2 (The MFN \((l, k)\)-equivalence class \(\mathcal{E}^{MFN}(A, B, C, D)\)).** Let \((A, B, C, D, K)\) be an MFN realization and let \(\pi(A, B, C, D, K) = (l, k)\) be the corresponding transfer function. The set \(\mathcal{E}^{MFN}(A, B, C, D) = \{ (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) \in \mathcal{E}(A, B, C, D) : (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) \text{ is MFN} \} \) is called the class of observably equivalent MFN systems or, briefly, the MFN \((l, k)\)-equivalence class.

The lemma below clarifies the structure of \(\mathcal{E}^{MFN}(A, B, C, D)\):

**Lemma 3.3.2 (Structure of the set \(\mathcal{E}^{MFN}(A, B, C, D)\)).** Let \((A, B, C, D, K)\) be an MFN realization and let \(\pi(A, B, C, D, K) = (l, k)\) be the corresponding transfer function. The MFN \((l, k)\)-equivalence class \(\mathcal{E}^{MFN}(A, B, C, D)\) constitutes a \(\frac{n(n-1)}{2}\) dimensional real analytic manifold consisting of two disconnected components and is given by
\[
\mathcal{E}^{MFN}(A, B, C, D) = \left\{ (QAQ', QB, CQ', D), \quad Q \in O(n) \right\}
\]
where \(O(n)\) denotes the set of real orthogonal \(n \times n\) matrices.

**Remark 3.3.1.** For \(n = s = 1\) and \(m = 0\), the MFN \(k\)-equivalence class in \(S(1)\) consists of two separated points on the hyperbola \(\mathcal{E}(a, k, c): \mathcal{E}^{MFN}(a, k, c) = \left\{ (a, \frac{k}{\sqrt{k^2 + c^2}}, \frac{c}{\sqrt{k^2 + c^2}}), (a, -\frac{k}{\sqrt{k^2 + c^2}}, -\frac{c}{\sqrt{k^2 + c^2}}) \right\}\); see plot (b) in figure (3.1) in section (3.7).

**Proof.** First, note that orthogonal state transformations leave the Frobenius norm of the realizations unchanged. Next, we again refer to theorem 5.1 on page 223 of (Helmke and Moore, 1994): If two minimal realizations \((A_1, B_1, C_1, D_1)\) and \((A_2, B_2, C_2, D_2)\) of the same transfer function satisfy \(A'A + BB' = A'A + C'C\), then there exists a unique orthogonal \(Q \in O(n)\) such that \((A_2, B_2, C_2, D_2) = (QA_1Q', QB_1, C_1Q', D_1)\). This results implies that \(\mathcal{E}^{MFN}(A, B, C, D)\) is characterized by (3.11).

To show that \(\mathcal{E}^{MFN}(A, B, C, D)\) is a real analytic manifold of dimension \(\frac{n(n-1)}{2}\), we proceed as follows: Let us consider the restriction of the mapping \(\phi\) as defined in (3.2) to \(O(n)\). From what was said above, it is clear that the image space \(\phi|_{O(n)}(O(n))\) is \(\mathcal{E}^{MFN}(A, B, C, D)\). Hence,
\[ \phi|_{O(n)} : O(n) \to E_{MFN}(A,\tilde{B},C,D) \]
\[ Q \mapsto \begin{pmatrix}
\text{vec}(QAQ') \\
\text{vec}(Q\tilde{B}) \\
\text{vec}(CQ') \\
\text{vec}(D)
\end{pmatrix} \quad (3.12) \]

Since \( \phi \) is a homeomorphism, the same holds true for \( \phi|_{O(n)} \). The fact that \( O(n) \) is itself a real analytic submanifold of \( \mathbb{R}^{n^2} \) of dimension \( n(n-1)/2 \) is well known; see, e.g., (Bröcker and Jänich, 1990). Hence, by considering a system of local coordinates for \( O(n) \), it is straightforward to construct a system of local coordinates for \( E_{MFN}(A,\tilde{B},C,D) \), showing that \( E_{MFN}(A,\tilde{B},C,D) \) is also a real analytic manifold of dimension \( n(n-1)/2 \).

\( E_{MFN}(A,\tilde{B},C,D) \) consists of two disconnected components because this is true for \( O(n) \) as has been shown in lemma (3.1.1).

\[ \square \]

**Remark 3.3.2.** Note that the set of constant Frobenius norm realizations for an arbitrary minimal realization may be larger than the set that is obtained by applying orthogonal state transformations. An example would be following:

\[ (A,K,C) = \left( \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \right) \]

Both \( (A,K,C) \) and \( (TAT^{-1},TK,CT^{-1}) \) have the same Frobenius norm. By the lemma above, this is not possible for MFN realizations.

Next we consider the tangent space to the MFN \((l,k)\)-equivalence class \( E_{MFN}(A,\tilde{B},C,D) \) at a given \((A,\tilde{B},C,D)\):

**Lemma 3.3.3 (Tangent space \( Q_{MFN}^{(A,\tilde{B},C,D)} \) to \( E_{MFN}(A,\tilde{B},C,D) \)).** Let \((A,\tilde{B},C,D,K)\) be an MFN realization. Then \((A,\tilde{B},C,D) + (A_s,\tilde{B}_s,C_s,D_s) \in Q_{MFN}^{(A,\tilde{B},C,D)} \) if and only if \((A_s,\tilde{B}_s,C_s,D_s) = (QA +AQ', \tilde{Q}B,CQ', 0)\) for some skew symmetric \(Q \in \mathbb{R}^{n \times n}\), i.e. \(Q' = -Q\).

**Remark 3.3.3.** From lemma (3.3.2) it is clear that the dimension of the affine subspace \( Q_{MFN}^{(A,\tilde{B},C,D)} \) is \( n(n-1)/2 \). Moreover, it will be a subset of \( Q_{(A,\tilde{B},C,D)} \), the tangent space to the whole \((l,k)\)-equivalence class \( E(A,\tilde{B},C,D) \). For \( n = s = 1 \), \( Q_{MFN}^{(A,k,c)} \) reduces to the (zero dimensional) point \((a,k,c)\) itself; see plot (b) in figure (3.1) in section (3.7).

**Proof.** According to lemma (3.3.2), \( E_{MFN}(A,\tilde{B},C,D) \) is given by (3.11). Differentiation yields

\[ (QAQ' + QAQ', \tilde{Q}B,CQ', 0) \]

(3.13)

Note that \( \tilde{Q} \) is clearly not allowed to vary in \( \mathbb{R}^{n \times n} \) as it was the case in lemma (3.2.2) because the tangent space to \( Q \in O(n) \) at \( Q = I \) is not the whole \( \mathbb{R}^{n \times n} \). In fact, from \( \tilde{Q}' = I \) we get \( \tilde{Q}'Q + Q'\tilde{Q} = 0 \) and at \( Q = I \) this reduces to \( \tilde{Q}' = -\tilde{Q} \): all skew symmetric matrices are in the tangent space to \( O(n) \) at \( Q = I \). Using this, (3.13) becomes

\[ \left\{ (QA + AQ', \tilde{Q}B, CQ', 0), \quad \tilde{Q} \in \mathbb{R}^{n \times n}_{\text{skew}} \right\} \]

(3.14)

where \( \mathbb{R}^{n \times n}_{\text{skew}} \) denotes the set of skew symmetric real \( n \times n \) matrices. Hence, \((A_s,\tilde{B}_s,C_s,D_s)\) must have a representation of the form (3.14).
The set of all MFN systems

Clearly, we may not restrict ourselves to MFN realizations within one particular \((l, k)\)-equivalence class \(\mathcal{E}(A, B, C, D)\):

**Definition 3.3.3 (The set \(\mathcal{MFN}(n)\) of all MFN realizations).** The set \(\mathcal{MFN}(n) = \{ (A, B, C, D, K) \in S_m(n) : (A, B, C, D, K) \text{ is MFN} \} \) is called the set of all MFN realizations.

The following two remarks treat the simple case where \(n = s = 1\) and \(m = 0\):

**Remark 3.3.4.** For \(n = s = 1\) and \(m = 0\), the set \(\mathcal{MFN}(1)\) in \(S(1)\) is the union of the two dimensional planes given by \(k = c\) and \(k = -c\), excluding the \(a\)-axis \((k = c = 0)\); see plot (b) in figure (3.1) in section (3.7). Note that this is a two dimensional real analytic manifold consisting of four disconnected components.

**Remark 3.3.5.** For \(n = s = 1\) and \(m = 0\), let \(Q_{(n,k,c)}^{\mathcal{MFN}}\) denote the tangent space to \(\mathcal{MFN}(1)\) at \((a, k, c)\). \(Q_{(a,k,c)}^{\mathcal{MFN}}\) is one of the two planes described in remark (3.3.4), including the \(a\)-axis; see plot (b) in figure (3.1) in section (3.7).

For the case when \(n, s\) and \(m\) are arbitrary, we continue with the following

**Remark 3.3.6.** We conjecture that the set \(\mathcal{MFN}(n)\) constitutes a \(2ns + m(n + s) + \frac{n(n-1)}{2}\) dimensional real analytic manifold which is given by

\[
\mathcal{MFN}(n) = \left\{ (A, \bar{B}, C, D) \in S_m(n) : AA' + \bar{B}\bar{B}' = A'A + C'C \right\} \quad \text{(3.15)}
\]

One way to show such a statement would be to consider the function \(f : S_m(n) \to \mathbb{R}^{(n+1)(n+1)/2}\) attaching the diagonal and upper diagonal part of the symmetric matrix \((AA' + \bar{B}\bar{B}' - A'A - C'C)\) to the matrices \((A, B, C, D)\). Then, clearly, \(\mathcal{MFN}(n) = f^{-1}(0)\) where \(0 \in \mathbb{R}^{(n+1)(n+1)/2}\) and by a result which will be used in the proof of statement (iv) of theorem (4.9.1), it suffices to show that the Jacobian of this mapping has full rank \(n(n+1)/2\) at any system \((A, B, C, D) \in f^{-1}(0)\). Differentiation yields \(\hat{A}'A' + A\hat{A}' + \bar{B}\bar{B}' + \bar{B}'\bar{B} = \hat{A}'A' - A'A - C'C - C'C\) and we must show that any symmetric matrix \(S \in \mathbb{R}^{n \times n}\) can be represented in this form for some \(\hat{A} \in \mathbb{R}^{n \times n}, \bar{B} \in \mathbb{R}^{n \times (m+s)}\) and \(C \in \mathbb{R}^{s \times n}\). However, we do not have a proof for this statement at the moment.

**Remark 3.3.7.** In case our conjecture in remark (3.3.6) is right, it is straightforward to compute the tangent space \(Q_{(A, B, C, D)}^{\mathcal{MFN}}\) to \(\mathcal{MFN}(n)\): \((A, B, C, D) + (A_s, B_s, C_s, D_s) \in Q_{(A, B, C, D)}^{\mathcal{MFN}}\) if and only if \((A_s A' + \bar{B}_s \bar{B}') + (A_s A' + \bar{B}_s \bar{B}')' = (A_s A' + C'C_s) + (A_s A' + C'C_s)'\).

**Remark 3.3.8.** In case our conjecture in remark (3.3.6) is right, it can be shown rather easily that the tangent space \(Q_{(A, B, C, D)}^{\mathcal{MFN}}\) can be decomposed in a nice way: \(Q_{(A, B, C, D)}^{\mathcal{MFN}} = Q_{(A, B, C, D)}^{\mathcal{MFN}^+} \oplus Q_{(A, B, C, D)}^{\mathcal{MFN}^-}\).

### 3.4 The manifold of observationally equivalent BSA systems

Let us first note that BSA stands for balanced, stable allpass. We will now explain both terms, starting with the term balanced which will be used frequently in this chapter for what should more precisely be called Lyapunov balanced:

**Definition 3.4.1 (Lyapunov Balancing).** A minimal state-space representation \((A, B, C, D, K) \in M_s(n)\) is called Lyapunov balanced, if the (unique) solutions to the Lyapunov equations

\[
P - APA' - \bar{B}\bar{B}' = 0 \quad \text{(3.16)}
\]

\[
Q - A'QA - C'C = 0 \quad \text{(3.17)}
\]

satisfy \(P = Q = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)\), where in addition \(\sigma_1 \geq \cdots \geq \sigma_n > 0\). The \(\sigma_i\) are called second order modes or Hankel singular values.
Let us continue with a few remarks:

**Remark 3.4.1.** It is easy to see that the controllability Gramian $P$ and the observability Gramian $Q$ as given in definition (1.4.3) in section (1.4) are exactly the solutions to (3.16) and (3.17) above.

**Remark 3.4.2.** Let $(A, B, C, D)$ be a minimal and stable system where $P$ and $Q$ are the corresponding controllability and observability Gramian, respectively. Then, the Gramians $\bar{P}$ and $\bar{Q}$ corresponding to the observationally equivalent system $(\bar{A}, \bar{B}, \bar{C}, \bar{D}) = (TAT^{-1}, TB, CT^{-1}, D)$, $T \in GL(n)$ are given by $\bar{P} = TPT'$ and $\bar{Q} = T^{-1}PT^{-1}$, respectively.

**Remark 3.4.3.** Balanced state-space realizations always exist: Using remark (3.4.2), it is easy to construct an observationally equivalent balanced state-space realization for any given minimal and stable $(A, B, C, D)$. Note, however, that this construction is not unique, i.e. there is more than one balanced realization in each $(l, k)$-equivalence class $\mathcal{E}(A, B, C, D)$. In fact, section (4.6) is concerned with the question of how to select a unique state-space realization within the set of balanced realizations.

**Remark 3.4.4.** From remark (3.4.2) it follows that the eigenvalues of $PQ$ and $\bar{P}\bar{Q}$ are the same. They are invariant within $\mathcal{E}(A, B, C, D)$. Trivially, the same holds true for the positive square roots of these eigenvalues. Note that the latter quantities are the Hankel singular values, and thus the $\sigma_i$ in definition (3.4.1) may more precisely be called the Hankel singular values of the transfer function $(l, k)$.

Having defined balanced state-space realizations for stable transfer functions $(l, k) \in M_\sigma(n)$, we will now consider a special subset of $M_\sigma(n)$:

**Definition 3.4.2 (Stable Allpass Transfer Functions).** A rational and causal $s \times s$ discrete time transfer function $k(z)$ is said to be stable allpass (or inner) if it satisfies $k^*(z)k(z) = I_s$ for all $z = e^{i\omega}$ where $k^*(z) = k(\frac{1}{z})'$.

Note that stable allpass transfer functions $k(z)$ are square by definition and need not necessarily satisfy $k(0) = I$. Therefore, a "constant feedthrough matrix" $D \in \mathbb{R}^{s \times s}$ has to be included for the corresponding state-space realizations:

**Definition 3.4.3 (BSA Realizations).** A minimal realization $(A, B, C, D)$ is said to be balanced stable allpass or, briefly, BSA, if it is Lyapunov balanced and $\pi(A, B, C, D)$ is a stable allpass transfer function.

**Remark 3.4.5.** We are only dealing with the real rational case – the coefficients in the numerator and denominator polynomials of each matrix entry are real. Clearly, $k^*(z) = k^H(z)$ if $|z| = 1$. Here, $k^H(z)$ denotes the Hermite transpose of the matrix $k(z)$.

**Remark 3.4.6.** Of course, also unstable transfer functions $k(z) = C(z^{-1}I - A)^{-1}B + D$ may satisfy $k^*(z)k(z) = I_s$, allowing for an extension of the definition to allpass transfer functions. If, however, $k(z)$ has a pole at some $z_0$ of unit modulus, then it cannot be allpass. To show this, let us assume for a moment that $z_0$ was a pole of $k(z)$ of unit modulus. As any invertible $k(z)$ has a pole at $z_0$ if and only if $k^{-1}(z)$ has a zero at $z_0$ (see, e.g., lemma 3.38 in [Zhou et al., 1996]), this implies the existence of a zero of $k^*(z)$ at $z_0$. Note that $k^*(z) = B'(zI - A')^{-1}C^* + D'$ is proper and has full normal rank. Under these assumptions (see lemma 3.28 in [Zhou et al., 1996]), $z_0$ is a zero if and only if there exists a vector $0 \neq u_0 \in \mathbb{C}^s$ such that $u_0^Hk^*(z_0) = 0$. This is the case even if the transfer function $k^*(z)$ also has a pole at $z_0$ (note that this is only possible for $s = 1$) because multiplication with $u_0^H$ will then yield a well-defined product $u_0^Hk^*(z_0) = 0$. Using the fact that $z_0 = e^{i\omega}$ and hence $k^*(z_0) = k^H(z_0)$, we get $0 = u_0^Hk^H(z_0)k(z_0)u_0 = u_0^Hk^*(z_0)k(z_0)u_0 = u_0^H0 = u_0^H0$ which is a contradiction because $u_0 \neq 0$.

We continue with the following:

**Definition 3.4.4 (Realization Matrix).** Let $(A, B, C, D) \in S(n)$ be a state-space realization of a square $s \times s$ transfer function $\pi(A, B, C, D) = k$. The matrix

$$R = \begin{pmatrix} D & C \\ B & A \end{pmatrix} \in \mathbb{R}^{(n+s) \times (n+s)}$$

is called the realization matrix corresponding to $(A, B, C, D)$.

It is obvious that the Frobenius norm of the realization $(A, B, C, D)$ (as given in (3.9)) is equal to the Frobenius norm of the corresponding realization matrix which is given by $\|R\|^2_F = tr(RR^T)$.

The following lemma can be found on page 552 of [Zhou et al., 1996], yet with a less detailed proof:
Lemma 3.4.1. Let \( (A, B, C, D) \in S(n) \) (not necessarily minimal) be given and suppose that \( Q = Q' \) satisfies \( Q = A' Q A + C' C \). Then the following statements hold:

(i) \( D' C + B' Q A = 0 \) implies \( k^*(z)k(z) = D' D + B' Q B \).

(ii) If \( A \) is stable, \( (A, B) \) is controllable and \( k^*(z)k(z) = D' D + B' Q B \), then \( D' C + B' Q A = 0 \).

Proof. Note that

\[
k^*(z)k(z) = (B'(zI - A)^{-1}C' + D') \left( C(z^{-1} I - A)^{-1} B + D \right)
= D' D + D' C(z^{-1} I - A)^{-1} B + B'(zI - A)^{-1} C' D + B'(zI - A)^{-1} C C(z^{-1} I - A)^{-1} B
= D' D + D' C(z^{-1} I - A)^{-1} B + B'(zI - A)^{-1} C' D + B'(zI - A)^{-1} (Q - A' Q A)(z^{-1} I - A)^{-1} B
\]

where the last equality is obtained by substituting \( C' C = Q - A' Q A \). Now, the power series

\[
(z^{-1} I - A)^{-1} = z(I - zA)^{-1} = \sum_{i=0}^{\infty} A^i z^i \quad |z| < \frac{1}{|\lambda_{\text{max}}(A)|}
\]

converges in the given neighborhood of \( z = 0 \). Note that we immediately get a Laurent series for

\[
(zI - A')^{-1} = z^{-1}(I - z^{-1} A') = z^{-1} \sum_{j=0}^{\infty} A'^j z^{-j} \quad |z| > |\lambda_{\text{max}}(A)|
\]

which is guaranteed to converge in the region specified above. Clearly, if \( |\lambda_{\text{max}}(A)| < 1 \) (i.e. if \( A \) is stable), both series converge in a common region. We can continue as follows:

\[
B'(zI - A')^{-1}(Q - A' Q A)(z^{-1} I - A)^{-1} B = B' \left( \sum_{j=0}^{\infty} A'^j z^{-j} \right) (Q - A' Q A) \left( \sum_{i=0}^{\infty} A^i z^i \right) B
\]

A comparison of coefficients yields:

\[
\begin{align*}
z^0 : B'(Q - A' Q A + A' Q A - A'^2 Q A^2 + A'^2 Q A^2 - \ldots)B &= B'QB \\
z^1 : B'(Q A - A' Q A^2 + A' Q A^2 - A'^2 Q A^3 + A'^2 Q A^3 - \ldots)B &= (B'QA)B \\
z^2 : B'(Q A^2 - A' Q A^3 + A' Q A^3 - A'^2 Q A^4 + A'^2 Q A^4 - \ldots)B &= (BQA)AB
\end{align*}
\]

In general, we have for the integer \( k > 0 \)

\[
\begin{align*}
z^k : \quad & (B'QA)A^{k-1} B \\
z^{-k} : \quad & B'A^{k-1'}(A'QB)
\end{align*}
\]

yielding

\[
B'(zI - A')^{-1}(Q - A' Q A)(z^{-1} I - A)^{-1} B = B'QB + \sum_{j=1}^{\infty} A'^j z^{-j} Bz^j + \sum_{i=1}^{\infty} B'A^{i-1'} z^{-i}(A'QB) = B'QB + B'QA(z^{-1} I - A)^{-1} B + B'(zI - A')^{-1} A'QB
\]
such that (3.18) can be written as:

$$k^*(z)k(z) = D'D + B'QB + (D'C + B'QA)(z^{-1}I - A)^{-1}B + B'(zI - A')^{-1}(C'D + A'QB)$$ (3.19)

It is important to note that the representation of $k^*(z)k(z)$ in (3.19) holds true in any case (independent of stability of $A$), but that the representation of $k^*(z)k(z)$ of the form

$$k^*(z)k(z) = (D'D + B'QB) + (D'C + B'QA)\sum_{j=1}^{\infty} A^{-1}Bz^j + \sum_{i=1}^{\infty} B'A^{-1}z^{-i}(C'D + A'QB)$$ (3.20)

only converges in some region containing the unit circle if $A$ is stable. Now (i) is evident, and for (ii) we can use (3.20). The coefficient matrices corresponding to $z^j$, $j \neq 0$ have to be zero, and from the first $n$ restrictions of this type (considering the coefficient matrices corresponding to $z, \ldots, z^n$) it is clear that $(D'C + B'QA) = 0$ by the controllability assumption.

This lemma proves to be a powerful tool:

**Corollary 3.4.2.** All Hankel singular values of a stable allpass transfer function $k(z)$ are equal to one.

**Proof.** Let $(A, B, C, D)$ be any minimal realization of $k(z)$ and let us assume that the observability Gramian is the identity matrix (if this is not the case, we have to apply an appropriate non singular state space transformation; see remark (3.4.2) above). Then, $I = A'A + C'C$, stability of $A$ and controllability of $(A, B)$ are satisfied as well as $k^*(z)k(z) = I$. Writing $k^*(z)k(z)$ in terms of a Laurent series expansion (see (3.20)), we see that $I = D'D + B'QB = D'D + B'B$. Applying lemma (3.4.1), (ii), yields $D'C + B'A = 0$. Hence,

$$R'R = \begin{pmatrix} D' & B' \\ C' & A' \end{pmatrix} \begin{pmatrix} D & C \\ B & A \end{pmatrix} = \begin{pmatrix} D'D + B'B & D'C + B'A \\ C'D + A'B & C'C + A'A \end{pmatrix} = I$$

and $RR' = I$ must also hold true. From the bottom right block of $RR'$ we see that $BB' + AA' = I$. Hence, $P = I$ is the unique solution to the controllability Lyapunov equation (3.16). The realization is balanced with $\Sigma = I$, which completes the proof.

The proof of corollary (3.4.2) also shows that the realization matrix corresponding to a BSA system is orthogonal. In fact, we have the following

**Lemma 3.4.3 (Characterization of BSA realizations).** Let $(A, B, C, D)$ be a BSA realization. Then the corresponding realization matrix $R$ is orthogonal, i.e. $R \in O(n + s)$. Conversely, let an orthogonal realization matrix $R \in O(n + s)$ be given. Then the corresponding system $(A, B, C, D)$ is a state-space realization of a stable allpass transfer function. If $(A, B, C, D)$ is minimal, $A$ is stable and the realization is balanced. Otherwise, $A$ has all its eigenvalues in the closed unit disk and some of its eigenvalues lie on the unit circle.

**Proof.** The first part has been shown in the proof of corollary (3.4.2). In order to show the second part, note that $R'R = RR' = I$ immediately implies that $A'A + C'C = I$, $AA' + BB' = I$, $D'D + B'B = I$ and $D'C + B'A = 0$. Hence, by (i) of lemma (3.4.1), we get $k^*(z)k(z) = I$, i.e. $k(z)$ is allpass.

By a standard result on Lyapunov equations the existence of a positive definite solution to (3.16) implies that $A$ is stable if and only if $(A, B)$ is controllable. Analogously, the existence of a positive definite solution to (3.17) implies that $A$ is stable if and only if $(C, A)$ is observable.

Hence, if $(A, B, C, D)$ is minimal, then $A$ is stable, and $A'A + C'C = I = AA' + BB'$ shows balancedness. If $(A, B, C, D)$ is non minimal, then $A$ is not stable. However, another standard result (see statement 2 (a) in lemma 21.6 in (Zhou et al., 1996)) guarantees that all eigenvalues of $A$ are of modulus less than or equal to one in case of existence of a positive definite solution to (3.16). Hence, $A$ must have eigenvalues of unit modulus, but these eigenvalues cannot be poles of $k(z)$ because $k(z)$ is allpass; see remark (3.4.6) above.
Remark 3.4.7. Note that the Frobenius norm of any BSA realization is \( \sqrt{n+s} \) as the Frobenius norm of any \( R \in O(n+s) \) is \( \sqrt{n+s} \).

Another interesting observation is the following:

**Lemma 3.4.4 (MFN and BSA realizations).** Let the system \((A,B,C,D)\) be a minimal realization of a square \( s \times s\) stable allpass transfer function. Then \((A,B,C,D)\) is MFN if and only if it is BSA.

**Proof.** "\( \Leftarrow \)" If \((A,B,C,D)\) is a BSA realization, then clearly \((A A' + B B') = (A' A + C' C)\), and by lemma (3.3.1) this is the necessary and sufficient condition for a minimal realization to be MFN.

"\( \Rightarrow \)" Conversely, let us assume for a moment that there exists an MFN realization of a stable allpass transfer function which is not balanced. Then it must have a representation of the form

\[
R(T) = \begin{pmatrix} I & 0 \\ 0 & T \end{pmatrix} R \begin{pmatrix} I & 0 \\ 0 & T^{-1} \end{pmatrix}
\]

where the realization matrix \( R \) corresponds to some BSA and hence MFN system and \( T \in GL(n) \setminus O(n) \). Clearly, \( \det(R(T)) = \pm \prod_{i=1}^{n+s} \sigma_i(R(T)) \) where \( \sigma_i(R(T)) \) denotes the \( i \)-th largest singular value of \( R(T) \). Moreover, \( \det(R(T)) = \det(R) = \pm 1 \) because \( R \) is orthogonal. Hence, \( \prod_{i=1}^{n+s} \sigma_i^2(R(T)) = 1 \), and by the well known arithmetic-geometric-mean inequality, we have \( \sum_{i=1}^{n+s} \sigma_i^2(R(T)) \geq n+s \). Note that this sum of squared singular values equals \( \|R(T)\|^2 \). The minimum \( n+s \) (the arithmetic mean equals the geometric mean) is attained if and only if all singular values \( \sigma_i(R(T)), \ i=1, \ldots , n+s \) are equal and hence one, implying that \( R(T) \) is orthogonal. As \( R(T) \) corresponds to a minimal system, use of lemma (3.4.3) shows that this system also has to be balanced, yielding a contradiction. \( \square \)

We are now ready to give the following

**Definition 3.4.5 (The BSA k-equivalence class \( \mathcal{E}^{BSA}(A,B,C,D) \)).** Let \((A,B,C,D)\) be a BSA realization and let \( \pi(A,B,C,D) = k \) be the corresponding square \( s \times s \) transfer function. The set \( \mathcal{E}^{BSA}(A,B,C,D) = \{ (A,\tilde{B},\tilde{C},\tilde{D}) \in \mathcal{E}(A,B,C,D) : (A,\tilde{B},\tilde{C},\tilde{D}) \text{ is BSA} \} \) is called the class of observationally equivalent BSA systems or, briefly, the BSA k-equivalence class.

The lemma below clarifies the structure of \( \mathcal{E}^{BSA}(A,B,C,D) \):

**Lemma 3.4.5 (Structure of the set \( \mathcal{E}^{BSA}(A,B,C,D) \)).** Let \((A,B,C,D)\in S(n)\) be a BSA realization and let \( \pi(A,B,C,D) = k \) be the corresponding square \( s \times s \) transfer function. The BSA k-equivalence class \( \mathcal{E}^{BSA}(A,B,C,D) \) constitutes a \( \frac{n(n-1)}{2} \) dimensional real analytic manifold consisting of two disconnected components and is given by

\[
\mathcal{E}^{BSA}(A,B,C,D) = \{ (QAQ^t, QB, CCQ^t, D), \quad Q \in O(n) \} \quad (3.21)
\]

or, equivalently, by

\[
\mathcal{E}^{BSA}(A,B,C,D) = \left\{ \left( \begin{array}{cc} I & 0 \\ 0 & Q \end{array} \right) R \left( \begin{array}{cc} I & 0 \\ 0 & Q^t \end{array} \right), \quad Q \in O(n) \right\} \quad (3.22)
\]

where \( R \) denotes the realization matrix corresponding to \((A,B,C,D)\).

**Remark 3.4.8.** For \( n = s = 1 \), the BSA k-equivalence class \( \mathcal{E}^{BSA}(a,b,c,d) \) in \( S(1) \) consists of two separated points (hence, its dimension is zero): \( \mathcal{E}^{BSA}(a,b,c,d) = \{ (a,b,c,d), (a,-b,-c,d) \} \); see plot (c) in figure (3.1) in section (3.7).

**Proof.** Note that \((A,B,C,D)\) is minimal and that the solutions \( P \) and \( Q \) of (3.16) and (3.17) satisfy \( P = Q = I \) by assumption. It follows from lemma (3.2.1) that the whole class of observationally equivalent state-space realizations is obtained by applying non singular state transformations \( T \in GL(n) \). The Gramians \( P \) and \( Q \) are then easily seen to be transformed to \( TPT^t \) and \( T^{-1}QT^{-1} \). Hence, for the new realization \((TAT^{-1}, TB, CT^{-1}, D)\) to remain balanced, \( TT^t = T^{-1}T^{-1} = I \) must hold true. This
is the case if and only if $T \in O(n)$, showing that $E^{BSA}(A, B, C, D)$ is given by (3.21) or, equivalently, (3.22). The fact that this set is a real analytic manifold of dimension $\frac{n(n-1)}{2}$ consisting of two disconnected components has already been established in the proof of lemma (3.3.2).

As the manifolds $E^{BSA}(A, B, C, D)$ and $E^{MFN}(A, \hat{B}, C, D)$ have the same structure, the following lemma is evident:

**Lemma 3.4.6 (Tangent space $Q_{(A,B,C,D)}^{BSA}$ to $E^{BSA}(A, B, C, D)$).** Let $(A, B, C, D)$ be a BSA realization. Then $(A, B, C, D) + (A_s, B_s, C_s, D_s) \in Q_{(A,B,C,D)}^{BSA}$ if and only if $(A_s, B_s, C_s, D_s) = (QA + AQ', QB, CQ', 0)$ for some skew symmetric $Q \in \mathbb{R}^{n \times n}$, i.e. $Q' = -Q$.

**Remark 3.4.9.** From lemma (3.4.5) it is clear that the dimension of the affine subspace $Q_{(A,B,C,D)}^{BSA}$ is $\frac{n(n-1)}{2}$. Moreover, it will be a subset of $Q_{(A,B,C,D)}$, the tangent space to the whole $k$-equivalence class $E(A, B, C, D)$. For $n = s = 1$, $Q_{(a,b,c,d)}^{BSA}$ reduces to the (zero dimensional) point $(a, b, c, d)$ itself; see plot (c) in figure (3.1) in section (3.7).

**Proof.** The proof is completely analogous to the proof of lemma (3.3.3).

The following lemma will be important in analyzing numerical properties of the orthoDLC algorithm:

**Lemma 3.4.7 (Transformation of systems in $E^{BSA}(A, B, C, D)$).** Let $(A, B, C, D)$ be a BSA realization and let $(A, B, C, D) + (A_s, B_s, C_s, D_s)$ be an arbitrary system. Then for any system $(QAQ', QB, CQ', D) \in E^{BSA}(A, B, C, D)$ we have:

(i) $\pi((A, B, C, D) + (A_s, B_s, C_s, D_s)) = \pi((QAQ', QB, CQ', D) + (QA_sQ', QB_s, C_sQ', D_s))$

(ii) $\llVert (A_s, B_s, C_s, D_s) \rrVert_F = \llVert (QA_sQ', QB_s, C_sQ', D_s) \rrVert_F$

(iii) If $(A, B, C, D) + (A_s, B_s, C_s, D_s) \in Q_{(A,B,C,D)}^+$, then $(QAQ', QB, CQ', D) + (QA_sQ', QB_s, C_sQ', D_s) \in Q_{(QAQ', QB, CQ', D)}^+$.

**Proof.** We proceed step by step:

(i) is obvious as

\[
\begin{pmatrix}
I & 0 \\
0 & Q
\end{pmatrix}
\begin{pmatrix}
D + D_s & C + C_s \\
B + B_s & A + A_s
\end{pmatrix}
\begin{pmatrix}
I & 0 \\
0 & Q'
\end{pmatrix} =
\begin{pmatrix}
D + D_s & (C + C_s)Q' \\
Q(B + B_s) & Q(A + A_s)Q'
\end{pmatrix}
\]

(ii) is clear because $\text{tr}(Q_sR_sQ'_s)(Q_sR_sQ'_s) = \text{tr}R'_sR_s$ for any $Q_s \in O(n + s)$ and any $R_s \in \mathbb{R}^{(n+s) \times (n+s)}$.

(iii) According to lemma (3.2.3), the condition

\[(QA_sQ')(QAQ')' + (QB_s)(QB)' = (QAQ')'(QA_sQ') + (CQ')'(C_sQ')\]

has to be satisfied. This is obvious, too, because $Q'Q = I$ and $A_sA' + B_sB' = A'A_s + C'C_s$ hold by assumption.

**Remark 3.4.10.** Considering the whole $k$-equivalence class $E(A, B, C, D)$ of the form (3.1), it is clear that (i) still holds true, whereas (ii) and (iii) will in general be violated; of course, $Q$ and $Q'$ have to be replaced by $T$ and $T^{-1}$ in this case.
3.5 The manifold of all BSA systems

Again, we may not restrict ourselves to a particular BSA $k$-equivalence class $E^{BSA}(A, B, C, D)$, but consider the set of all BSA realizations instead:

**Definition 3.5.1 (The set $BSA(n)$ of all BSA realizations).** The set $BSA(n) = \{(A, B, C, D) \in S(n) : (A, B, C, D) \text{ is BSA}\}$ is called the set of all BSA realizations.

The lemma below clarifies the structure of $BSA(n)$:

**Lemma 3.5.1 (Structure of the set $BSA(n)$).** The set $BSA(n)$ constitutes a $\frac{(n+s)(n+s-1)}{2}$-dimensional real analytic manifold. It is given by

$$BSA(n) = \{(A, B, C, D) \in S_m(n) : RR^t = I_{n+s}\} \hspace{1cm} (3.23)$$

where $R$ is the realization matrix corresponding to $(A, B, C, D)$. $BSA(n)$ is an open and dense subset of $O(n+s)$ and hence a thin subset of $K(0, \sqrt{n+s})$, the ball in $S(n)$ of radius $\sqrt{n+s}$ with its center in the origin.

**Remark 3.5.1.** For $n = s = 1$, the set $BSA(1)$ in $S(1)$ is obtained by either one of the choices

$$\begin{pmatrix} d & c \\ b & a \end{pmatrix} = \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix}, \begin{pmatrix} d & c \\ b & a \end{pmatrix} = \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ \sin(\phi) & -\cos(\phi) \end{pmatrix}, \phi \in (0, \pi) \cup (\pi, 2\pi) \hspace{1cm} (3.24)$$

Note that $\phi = 0$ and $\phi = \pi$ have to be excluded because then $b = c = 0$ would hold, i.e. the system would be non minimal. If we neglect the scalar $d$, then the set of all balanced stable alpass systems in $\mathbb{R}^2$ is seen to consist of two unit circles in the $b = c$ and the $b = -c$ plane, respectively, where in both cases the intersections with the $a$-axis have to be excluded. Thus, $BSA(1)$ consists of four disconnected components; see plot (c) in figure (3.1) in section (3.7).

**Proof.** The fact that the set $BSA(n)$ is characterized by (3.23) follows from lemma (3.4.3). As mentioned above, $O(n+s)$ is known to be a real analytic submanifold of $S(n) = \mathbb{R}^{s+2ns+s^2}$ of dimension $(n+s)(n+s-1)/2$; see again (Bröcker and Jänich, 1990), for instance. The set $BSA(n)$ is an open subset of $O(n+s)$: This follows from the fact that for each minimal $(A, B, C, D)$ there exists an open neighborhood in $S(n)$ (and hence an open neighborhood in $O(n+s)$ if we endow $O(n+s)$ with the relative topology) containing only minimal realizations; see statement (i) in theorem (4.1.2) in section (4.1). Clearly, any open subset of a real analytic manifold is itself a real analytic manifold. The last statement is evident as the dimension of the sphere $K(0, \sqrt{n+s})$ is $n^2+2ns+s^2$ and thus strictly greater than the dimension of $BSA(n)$. It remains to show that $BSA(n)$ is a dense subset of $O(n+s)$. Let $R$ be any realization matrix in $O(n+s)$ corresponding to a non minimal system. Proceeding in the same way as in the proof of lemma (3.1.1), it is easy to construct a coordinate chart $(U_1, \phi_1)$ such that $U_1$ contains $R$ and is open and dense in the connected component of $O(n+s)$ to which $R$ belongs. Thus $T_1 = \phi_1(U_1)$ is an open and (by continuity) connected subset of $\mathbb{R}^{s+2ns+s^2-1}$. Consider the mapping $\Delta_1 : T_1 \rightarrow \mathbb{R}$ attaching $det(W_n^a(\tau_1)W_n^c(\tau_1))$ to $\tau_1 \in T_1$ where $W_n^a(\tau_1) = O_n(\tau_1)^tC_n(\tau_1)$ and $W_n^c(\tau_1) = \mathbb{R}^{s+n}$ with $O_n(\tau_1) = O_n(A_{\tau_1}, B_{\tau_1}, C_{\tau_1}, D_{\tau_1})$

being the corresponding (finite) observability matrix; see definition (1.4.2). The Gramian $W_n^c(\tau_1)$ is given analogously. Note that $W_n^a(\tau_1)$ and $W_n^c(\tau_1)$ have full rank if and only if $(A, B, C, D)$ is minimal. Clearly, $\Delta_1$ is an analytic function, and $\Delta_1(\tau_1^R) = 0$ by assumption, where $\tau_1^R = \phi_1(R)$ is the parameter vector corresponding to $R$. This can only hold true either on a thin subset of $T_1$ (which would complete the proof) or everywhere in $T_1$ (note that $T_1$ is connected). The latter implies that an open and dense subset of the connected component to which $R$ belongs would correspond to non minimal alpass systems, and from the knowledge that $BSA(n)$ is an open subset of $O(n+s)$ we can conclude that the whole component would correspond to non minimal alpass systems. However, this is a contradiction as it is easy to find a
realization matrix corresponding to a BSA (and, therefore, minimal) system in each of the two connected components of $O(n + s)$. Consider, e.g., the two realization matrices

$$R_+^T\left( \begin{array}{cc}
0 & 1 \\
0 & 0 \\
1 & 0 \\
0 & 1 \\
\pm 1 & 0 \\
\end{array} \right)$$

where $n$ and $s$ are arbitrary. Evidently, $|\lambda_{\text{max}}(A)| = 0$ and $A$ is in Jordan canonical form. It is straightforward to see that $(A, \pm B, C, D)$ are minimal and that $\det R_+^T = \pm 1 \cdot (-1)^{n+s+1}$, implying that the matrices $R_+^T$ come to lie in two different components. This completes the proof.

The tangent space to the manifold $BSA(n)$ is described in the next lemma:

**Lemma 3.5.2 (Tangent space $Q_{(A,B,C,D)}^{BSA}$ to $BSA(n)$).** Let $(A, B, C, D)$ be a BSA realization and let $R$ be the corresponding realization matrix. Then $(A, B, C, D) + (A_s, B_s, C_s, D_s) \in Q_{(A,B,C,D)}^{BSA}$ if and only if

$$\left( \begin{array}{cc}
D_s & C_s \\
B_s & A_s \\
\end{array} \right) = R \dot{Q}, \quad \dot{Q} \in \mathbb{R}_n^{(n+s) \times (n+s)}$$

where $\mathbb{R}_n^{n \times n}$ denotes the set of skew symmetric real $n \times n$ matrices.

**Remark 3.5.2.** From lemma (3.5.1) it is clear that the dimension of the affine subspace $Q_{(A,B,C,D)}^{BSA}$ is $\frac{(n+s)(n+s+1)}{2}$. For $n = s = 1$, it follows from remark (3.5.1) that $Q_{(a,b,c,d)}^{BSA}$ is given by the tangent line to the unit circle in the $b = c$ or the $b = -c$ plane (neglecting the scalar $d$ again); see plot (c) in figure (3.1) in section (3.7).

**Proof.** For any given $(A, B, C, D) \in BSA(n)$ with a corresponding realization matrix $R$, it is evident from lemma (3.5.1) that $BSA(n)$ can be locally described by

$$\{RQ : Q \in U(I_{n+s})\}$$

where $U(I_{n+s})$ is some open neighborhood of the identity matrix in $O(n + s)$. Hence, differentiation yields $R \dot{Q}$, where $\dot{Q}$ has to be an element of the tangent space to the manifold $O(n + s)$ at $I_{n+s}$. We know already from the proof of lemma (3.4.6) that the set of all skew symmetric matrices $\dot{Q} \in \mathbb{R}_n^{(n+s) \times (n+s)}$ spans this tangent space, and this shows the result.

### 3.6 The manifold of $L_{T_T}^{EC}$-equivalent systems

In this section, we consider the function $L_{T_T}^{EC}(Y_1^T; U_1^T, \tau^o)$ in (2.85) in section (2.3.3). Moreover, we take

$$\tau^o = ((v e \tilde{A})', (\bar{v} e \tilde{C})')' \in \mathbb{R}^{n^2 + ns}$$

to be a vector consisting of all matrix entries of the pair $(C, \tilde{A})$. Thus, we may write $L_{T_T}^{EC}(Y_1^T; U_1^T, \tilde{A}, \tilde{C})$. Recall that $Y_1^T$ and $U_1^T$ denote the stacked vectors of outputs and inputs, respectively.

Let us now assume that we are given a pair $(\tilde{C}, \tilde{A})$, $\tilde{C} \in \mathbb{R}^{n \times n}$, $\tilde{A} \in \mathbb{R}^{n \times n}$ and that we consider arbitrary matrices $B \in \mathbb{R}^{n \times s}$ and $D \in \mathbb{R}^{s \times s}$ such that $(\tilde{A}, B, \tilde{C}, D)$ is a square system. If $R$ denotes the corresponding realization matrix, the criterion function $L_{T_T}^{EC}(Y_1^T; U_1^T, \tilde{A}, \tilde{C})$ can also be viewed as a function of the form
\[
L_T^c(Y_1^T; U_1^T, (\tilde{A}, B, \tilde{C}, D)) \equiv L_T^c(Y_1^T; U_1^T, R) \tag{3.26}
\]
where \(B\) and \(D\) are unnecessary in the sense that \(L_T^c\) does not vary for varying \(B\) and \(D\). We then have the following

**Lemma 3.6.1.** Let the system \((\tilde{A}_0, B_0, \tilde{C}_0, D_0)\) be given. Moreover, let \(\Sigma\) and \((Y_1^T, U_1^T)\) be given where \(T > (m + n) + mn/s\) and let the matrix \(X(Y_1^{T-1}; U_1^T, \tilde{C}_0, \tilde{A}_0, \Sigma)\) in (2.81) have full column rank. Then there exists an open and dense subset \(S\) of \(\mathbb{R}^{n^2 + 2ns + s^2}\) such that

\[
\Delta_V^U : S \rightarrow S(n) \quad (\tilde{A}, B, \tilde{C}, D) \mapsto (\tilde{A} - \tilde{K}\tilde{C}, B - \tilde{K}D, -\tilde{C}, -\tilde{D}, \tilde{K}) \tag{3.27}
\]
defines an analytic function, where \(\tilde{D}, \tilde{B}\) and \(\tilde{K}\) are given by (2.83). The subset \(S_m\) of \(S\) corresponding to all minimal \((\tilde{A}, B, \tilde{C}, D)\) in \(S\) is also open and dense in \(\mathbb{R}^{n^2 + 2ns + s^2}\).

**Proof.** Note that \(\Delta_V^U\) can be decomposed as follows: First, \((\tilde{A}, B, \tilde{C}, D)\) is mapped onto the system \((\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K})\) which is then simply inverted to obtain \((\tilde{A} - \tilde{K}\tilde{C}, B - \tilde{K}D, -\tilde{C}, -\tilde{D}, \tilde{K})\) in the second step. It is evident that this second transformation is an analytic function for any \((\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K})\). Thus, it remains to show that \((\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K})\) is an analytic function of \((\tilde{A}, B, \tilde{C}, D)\) in \(S\) where \(S\) is open and dense in \(\mathbb{R}^{n^2 + 2ns + s^2}\).

By lemma (2.3.2), the matrix \(\tilde{X}(Y_1^{T-1}; U_1^T, \tilde{C}, \tilde{A}, \Sigma)\) has full column rank for all \(\tau^o = ((\text{vec}\tilde{A})', (\text{vec}\tilde{C})')' \in T^o \subseteq \mathbb{R}^{n^2 + ns}\) where \(T^o\) is an open and dense subset of \(\mathbb{R}^{n^2 + ns}\). As \(\tilde{X}\) does not depend on \(B\) and \(D\), this statement holds true for \((\tilde{A}, \tilde{B}, \tilde{C}, D)\in S\), where \(S\) is open and dense in \(\mathbb{R}^{n^2 + 2ns + s^2}\). It then follows immediately (from the structure of (2.83)) that (2.83) defines an analytic function on \(S\).

The last statement follows from the fact that the set of minimal systems is known to form an open and dense subset of \(\mathbb{R}^{n^2 + 2ns + s^2}\) (see statement (i) of theorem (4.1.2) in section (4.1)) and that the intersection with \(S\) maintains these properties as \(\mathbb{R}^{n^2 + 2ns + s^2}\) is a complete metric space; clearly, openness of both sets is important at this point.

The following assumption will be needed frequently:

**Assumption 3.6.1.** Let the pair \((\tilde{C}, \tilde{A})\) be observable and let \((\tilde{A}, B, \tilde{C}, D)\) be a realization of a square \(s\times s\) transfer function \(k\) where \(B \in \mathbb{R}^{n \times s}\) and \(D \in \mathbb{R}^{s \times s}\) are arbitrary. Moreover, let \(\Sigma\) and \((Y_1^T, U_1^T)\) be given where \(T > (m + n) + mn/s\). Assume that the matrix \(\tilde{X}(Y_1^{T-1}; U_1^T, \tilde{C}, \tilde{A}, \Sigma)\) in (2.81) has full column rank and that \(\Delta_V^U(\tilde{A}, B, \tilde{C}, D)\) as given in (3.27) is minimal.

We are now able to continue with the following

**Definition 3.6.1 (The \(L_T^c\) k-equivalence class \(E^c(\tilde{A}, B, \tilde{C}, D)\)).** Let **Assumption 3.6.1** be satisfied. The set \(E^c(\tilde{A}, B, \tilde{C}, D) = \{(\tilde{A}, B, \tilde{C}, D) \in S(n) : \Delta_V^U((\tilde{A}, B, \tilde{C}, D)) \in E(\Delta_V^U(\tilde{A}, B, \tilde{C}, D))\}\) is called the class of \(L_T^c\) observationally equivalent systems or, briefly, the \(L_T^c\) k-equivalence class.

The following lemma clarifies the structure of \(E^c(\tilde{A}, B, \tilde{C}, D)\):

**Lemma 3.6.2 (Structure of the set \(E^c(\tilde{A}, B, \tilde{C}, D)\)).** Let **Assumption 3.6.1** be satisfied. Then the \(L_T^c\) k-equivalence class \(E^c(\tilde{A}, B, \tilde{C}, D)\) constitutes a \(n^2 + ns + s^2\) dimensional real analytic manifold consisting of two disconnected components and is given by

\[
E^c(\tilde{A}, B, \tilde{C}, D) = \{(T\tilde{A}T^{-1}, \Theta_B, \tilde{C}T^{-1}, \Theta_D), \quad T \in GL(n), \Theta_B \in \mathbb{R}^{n \times s}, \Theta_D \in \mathbb{R}^{s \times s}\} \tag{3.28}
\]

**Proof.** Given a square system satisfying **Assumption 3.6.1**, we first show that another square system is in the same \(L_T^c\) k-equivalence class if and only if there exists a non-singular state transformation \(T \in GL(n)\) relating the system matrices to each other according to (3.28).

The following assumption will be needed frequently:

**Assumption 3.6.1.** Let the pair \((\tilde{C}, \tilde{A})\) be observable and let \((\tilde{A}, B, \tilde{C}, D)\) be a realization of a square \(s\times s\) transfer function \(k\) where \(B \in \mathbb{R}^{n \times s}\) and \(D \in \mathbb{R}^{s \times s}\) are arbitrary. Moreover, let \(\Sigma\) and \((Y_1^T, U_1^T)\) be given where \(T > (m + n) + mn/s\). Assume that the matrix \(\tilde{X}(Y_1^{T-1}; U_1^T, \tilde{C}, \tilde{A}, \Sigma)\) in (2.81) has full column rank and that \(\Delta_V^U(\tilde{A}, B, \tilde{C}, D)\) as given in (3.27) is minimal.
"Φ": Using the shorthand notation \( \tilde{X} = \hat{X}(Y_{1}^{T-1}, U_{C}^{T}, \tilde{C}, \tilde{A}, \tilde{S}) \) and \( \tilde{X}_{T} = \hat{X}(Y_{1}^{T-1}, U_{C}^{T}, \tilde{C}T^{-1}, T\tilde{A}T^{-1}, \tilde{S}) \) and noting that

\[
(u', y') \otimes (\tilde{C}\tilde{A}^{k})T^{-1} = (1 \cdot (u', y')) \otimes (\tilde{C}\tilde{A}^{k} \cdot T^{-1})
\]

\[
= (1 \otimes \tilde{C}\tilde{A}^{k}) \cdot ((u', y') \otimes T^{-1})
\]

\[
= (\tilde{C}\tilde{A}^{k}) \cdot ((u', y') \otimes T^{-1})
\]

we have

\[
\tilde{X}_{T} = -(I_{T} \otimes \tilde{S}^{-\frac{1}{2}})
\begin{pmatrix}
  u_{1}' \otimes I_{n} & \tilde{C}[(u_{1}', y_{1}') \otimes T^{-1}] \\
  u_{2}' \otimes I_{n} & \tilde{C}A[(u_{1}', y_{1}') \otimes T^{-1}] + \tilde{C}[(u_{2}', y_{2}') \otimes T^{-1}] \\
  \vdots & \vdots \\
  u_{m}' \otimes I_{n} & \tilde{C}A^{-2}(u_{m}', y_{m}') \otimes T^{-1} + \cdots + \tilde{C}[(u_{m}', y_{m}') \otimes T^{-1}] + \tilde{C}[(u_{1}', y_{1}') \otimes I_{n}][I_{m+s} \otimes T^{-1}]
\end{pmatrix}
\begin{pmatrix}
  I_{m} \otimes I_{n} \otimes T^{-1} \otimes \tilde{S}^{-\frac{1}{2}} \\
  0_{n(m+s) \times n} \otimes I_{n} \otimes T^{-1} \otimes \tilde{S}^{-\frac{1}{2}}
\end{pmatrix}
\]

where \( \tilde{X}_{1} \in \mathbb{R}^{n \times m} \) and \( \tilde{X}_{2} \in \mathbb{R}^{n \times n(m+s)} \) denote the first and second block column of \( \tilde{X} \), respectively. As \( rk(\tilde{X}) = rk(\tilde{X}_{T}) \), \( \Delta_{V}(T\tilde{A}^{-1}, \Theta_{B}, \tilde{C}T^{-1}, \Theta_{D}) \) is well defined and from formula (2.83) we get

\[
\bar{\tilde{X}}^{u} = \left[ \begin{pmatrix}
  I & 0 \\
  0 & I \otimes T^{-1}
\end{pmatrix} \right]^{T} \left( \begin{pmatrix}
  \tilde{X}_{1}^{'} \tilde{X}_{1} & \tilde{X}_{1}^{'} \tilde{X}_{2} \\
  \tilde{X}_{2}^{'} \tilde{X}_{1} & \tilde{X}_{2}^{'} \tilde{X}_{2}
\end{pmatrix} \right) \left[ \begin{pmatrix}
  I & 0 \\
  0 & I \otimes T^{-1}
\end{pmatrix} \right]^{-1} \left[ \begin{pmatrix}
  I & 0 \\
  0 & I \otimes T^{-1}
\end{pmatrix} \right]^{T} \left( \begin{pmatrix}
  \bar{\tilde{X}}_{1} \\
  \bar{\tilde{X}}_{2}
\end{pmatrix} \right)
\]

yielding \( \bar{\tilde{X}}^{u} = (vec(\bar{D})', vec(T\bar{B})', vec(T\bar{K})')' \). Inverting \( (T\tilde{A}^{-1}, T\tilde{B}, \tilde{C}T^{-1}, \tilde{D}, T\tilde{K}) \) yields

\[
(T(\tilde{A} - \tilde{K}\tilde{C})T^{-1}, T(\tilde{B} - \tilde{K}\tilde{D}), -\tilde{C}T^{-1}, -\tilde{D}, T\tilde{K})
\]

showing that \( \Delta_{V}(T\tilde{A}^{-1}, \Theta_{B}, \tilde{C}T^{-1}, \Theta_{D}) \in \mathcal{E}(\Delta_{V}((\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}))). \)

"⇒": Let us assume that the systems \( (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) \) and \( (\hat{A}, \hat{B}, \hat{C}, \hat{D}) \) are in the same \( L_{T}^{p} \) k-equivalence class, i.e. the minimal systems \( \Delta_{V}(\tilde{A}, \tilde{B}, \tilde{C}, \hat{D}) = (\hat{A}, \hat{B}, \hat{C}, \hat{D}, K) \) and \( \Delta_{V}(\hat{A}, \hat{B}, \hat{C}, \hat{D}) = (\hat{A}, \hat{B}, \hat{C}, \hat{D}, K) \) are observationally equivalent. Then, by the proof of Lemma (3.2.1), there exists a unique (1) state transformation \( T \in GL(n) \) such that \( (\tilde{A}, \tilde{B}, \tilde{C}, \hat{D}, K) = (T\tilde{A}^{-1}, T\hat{B}, CT^{-1}, D, TK) \), Hence \( \hat{A} = T\tilde{A}^{-1} \) and \( \hat{C} = CT^{-1} \), and using (3.27), we get

\[
\hat{A} - \hat{K}\hat{C} = T(\tilde{A} - \tilde{K}\tilde{C})T^{-1} \Rightarrow \hat{A} + \hat{K}\hat{C} = T\tilde{A}^{-1} + TKCT^{-1} \Rightarrow \hat{A} + TKCT^{-1} = T\tilde{A}^{-1} + TKCT^{-1}
\]

Therefore, \( \hat{A} = T\tilde{A}^{-1} \) and \( \hat{C} = CT^{-1} \), showing that a unique \( T \in GL(n) \) relates the \( L_{T}^{p} \) equivalent systems to each other according to (3.28).
The fact that \( \mathcal{E}^c(\bar{A}, B, \bar{C}, D) \) is a \( n^2 + ns + s^2 \) dimensional real analytic manifold consisting of two disconnected components follows from the observation that the mapping

\[
\phi^c : GL(n) \times \mathbb{R}^{n \times s} \times \mathbb{R}^{s \times s} \rightarrow \mathcal{E}^c(\bar{A}, B, \bar{C}, D)
\]

\[
(T, \Theta_B, \Theta_D) \mapsto \begin{pmatrix}
\text{vec}(TAT^{-1}) \\
\text{vec}(\Theta_B) \\
\text{vec}(\bar{C}T^{-1}) \\
\text{vec}(\Theta_D)
\end{pmatrix}
\]

(3.29)

is a homeomorphism. This is easy to see as the function \( \phi_{\text{red}} : GL(n) \rightarrow \mathbb{R}^{n^2 + ns} \) attaching the vector \( \langle \text{vec}(TAT^{-1})', \text{vec}(CT^{-1})' \rangle \) to \( T \in GL(n) \) is injective for any observable pair \( (C, A) \). The proof is analogous to the proof of lemma (3.2.1).

Remark 3.6.1. If \( (\bar{A}, B, \bar{C}, D) \) is minimal, then \( \mathcal{E}(\bar{A}, B, \bar{C}, D) \) is a subset of \( \mathcal{E}^c(\bar{A}, B, \bar{C}, D) \). For \( n = s = 1 \), neglecting the scalar \( D \), the \( L^c_T \) \( k \)-equivalence classes \( \mathcal{E}^c(a, B, \bar{C}, D) \) in \( S(1) \) are planes given by a fixed \( a \), excluding the \( B \)-axis; see plot (d) in figure (3.1) in section (3.7).

The corresponding tangent space is described in the following

**Lemma 3.6.3 (Tangent space \( Q^c_{\bar{A}, B, \bar{C}, D} \) to \( \mathcal{E}^c(\bar{A}, B, \bar{C}, D) \)).** Let assumption 3.6.1 be satisfied. Then \( (\bar{A}, B, \bar{C}, D), (\bar{A}_s, B_s, \bar{C}_s, D_s) \in Q^c_{\bar{A}, B, \bar{C}, D} \) if and only if

\[
(\bar{A}_s, B_s, \bar{C}_s, D_s) = (\bar{T} \bar{A} - \bar{A} \bar{T}, \bar{B} - \bar{C} \bar{T}, \bar{D})
\]

(3.30)

for some \( \bar{T} \in \mathbb{R}^{n \times n}, \bar{B} \in \mathbb{R}^{n \times s} \) and \( \bar{D} \in \mathbb{R}^{s \times s} \).

Remark 3.6.2. From lemma (3.6.2) it is clear that the dimension of the affine subspace \( Q^c_{\bar{A}, B, \bar{C}, D} \) is \( n^2 + ns + s^2 \). For \( n = s = 1 \), neglecting the scalar \( D \), \( Q^c_{\bar{A}, B, \bar{C}, D} \) is the plane given in remark (3.6.1), now including the \( B \)-axis; see plot (d) in figure (3.1) in section (3.7).

Proof. The proof is completely analogous to the proof of lemma (3.2.2). Consider a differentiable path \( (T, \Theta_B, \Theta_D) = (T(\theta), \Theta_B(\theta), \Theta_D(\theta)) \) in \( GL(n) \times \mathbb{R}^{n \times s} \times \mathbb{R}^{s \times s} \) where \( \theta \in [0, 1] \), \( (T, \Theta_B, \Theta_D) = (T(\theta), \Theta_B(\theta), \Theta_D(\theta)) \) is the derivative with respect to \( \theta \), \( (T(0), \Theta_B(0), \Theta_D(0)) = (I, B, D) \) and, finally, \( (T(0), \Theta_B(0), \Theta_D(0)) = (T, B, D) \) (with slight abuse of notation for \( T \)). Differentiation of (3.28) yields

\[
\{(\bar{T} \bar{A} - \bar{A} \bar{T}, \bar{B} - \bar{C} \bar{T}, \bar{D}): \bar{T} \in \mathbb{R}^{n \times n}, \bar{B} \in \mathbb{R}^{n \times s}, \bar{D} \in \mathbb{R}^{s \times s}\}
\]

(3.31)

and at \( \theta = 0 \), i.e. at the given realization \( (\bar{A}, B, \bar{C}, D) \), (3.31) reduces to

\[
\{(\bar{T} \bar{A} - \bar{A} \bar{T}, \bar{B} - \bar{C} \bar{T}, \bar{D}): \bar{T} \in \mathbb{R}^{n \times n}, \bar{B} \in \mathbb{R}^{n \times s}, \bar{D} \in \mathbb{R}^{s \times s}\}
\]

(3.32)

showing the result.

Remark 3.6.3. Again, (3.31) can be vectorized to yield

\[
\begin{pmatrix}
T^{-1} \otimes T & 0 & 0 & 0 \\
0 & I_s \otimes I_n & 0 & 0 \\
0 & 0 & T^{-1} \otimes I_s & 0 \\
0 & 0 & 0 & I_s \otimes I_s
\end{pmatrix}
\cdot
\begin{pmatrix}
\bar{A} \otimes I_n - I_n \otimes \bar{A} & 0 & 0 & 0 \\
0 & I_n \otimes I_n & 0 & 0 \\
- I_n \otimes \bar{C} & 0 & 0 & 0 \\
0 & 0 & 0 & I_s \otimes I_s
\end{pmatrix}
\]

(3.33)

\[
\begin{pmatrix}
I_n \otimes T^{-1} & 0 & 0 & 0 \\
0 & I_s \otimes I_n & 0 & 0 \\
0 & 0 & I_s \otimes I_s
\end{pmatrix}
\cdot
\begin{pmatrix}
\text{vec}(\bar{T}) \\
\text{vec}(\Theta_B) \\
\text{vec}(\bar{C}) \\
\text{vec}(\Theta_D)
\end{pmatrix}, \quad \bar{T} \in \mathbb{R}^{n \times n}, \bar{B} \in \mathbb{R}^{n \times s}, \bar{D} \in \mathbb{R}^{s \times s}
\]
At \((T, \Theta_B, \Theta_D) = (I, B, D)\), (3.33) then becomes

\[
Q_{cc} \cdot \begin{pmatrix}
vec(\hat{T}) \\
vec(\hat{B}) \\
vec(\hat{D})
\end{pmatrix}, \quad \hat{T} \in \mathbb{R}^{n \times n}, \hat{B} \in \mathbb{R}^{n \times s}, \hat{D} \in \mathbb{R}^{s \times s}
\]

(3.34)

which is the vectorization of (3.30).

The ortho-complement to the tangent space \(Q_{cc}^{\perp}(A, B, C, D)\) is denoted by \(Q_{cc}^{\perp}(A, B, C, D)\):

**Lemma 3.6.4 (Ortho-complement) \(Q_{cc}^{\perp}(A, B, C, D)\).** Let **Assumption 3.6.1** be satisfied. Then \((A, B, C, D) + (A_s, B_s, C_s, D_s) \in Q_{cc}^{\perp}(A, B, C, D)\) if and only if \(A_s \bar{A} - \bar{A} A_s - C^t C_s = 0\) and \(B_s = 0, D_s = 0\).

**Remark 3.6.4.** It is clear that the dimension of the affine subspace \(Q_{cc}^{\perp}(A, B, C, D)\) is \(ns\). For \(n = s = 1\), neglecting the scalar \(D\), \(Q_{cc}^{\perp}(A, B, C, D)\) is the straight line being orthogonal to the tangent space described in remark (3.6.2): \(C\) and \(\bar{c}\) are fixed, and \(\bar{a}\) is varying; see plot (d) in figure (3.1) in section (3.7).

**Proof.** We proceed completely analogous to the proof of lemma (3.2.3) to obtain

\[
\text{tr}\{A_s \bar{A} \hat{T}'\} - \text{tr}\{\bar{A}' A_s \hat{T}'\} + \text{tr}\{B_s \hat{B}'\} - \text{tr}\{C^t C_s \hat{T}'\} + \text{tr}\{D_s \hat{D}'\} = 0
\]

\[\forall \hat{T} \in \mathbb{R}^{n \times n}, \hat{B} \in \mathbb{R}^{n \times s}, \hat{D} \in \mathbb{R}^{s \times s}\]

We can again consider matrices \(\hat{T}, \hat{B}\) and \(\hat{D}\) with only one nonzero entry and vary the position of these entries to get the result. \(\square\)

### 3.7 Figures
Figure 3.1: (a) The case \( n = s = 1 \) and \( m = 0 \) where \( S(1) = \mathbb{R}^2 \): The thick hyperbola corresponds to the \( k \)-equivalence class \( \mathcal{E}(a, k, c) \), the red straight line to the tangent space \( Q_{(a, k, c)} \) and the affine plane to the orthogonal complement \( Q_{(a, k, c)}^\perp \). The minimal system \( (a, k, c) = (0, 3, 1) \) is represented by the black point.

(b) The case \( n = s = 1 \) and \( m = 0 \) where \( S(1) = \mathbb{R}^3 \): The two black points correspond to the set \( E_{MFN}^{MFN}(a, k, c) \) and the upper right black point to the tangent space \( Q_{(a, k, c)}^{MFN} \) where the minimal system \( (a, k, c) \) equals \( (0, \sqrt{3}, \sqrt{3}) \). The set \( MFN(1) \) is given by the union of the two planes, excluding the \( a \)-axis. The tangent space \( Q_{(a, k, c)}^{MFN} \) coincides with the plane given by \( k = c \).

(c) The case \( n = s = 1 \) in the \((\bar{a}, B, \bar{c})\)-space (neglecting the scalar \( D \)): The two black points correspond to the set \( E_{BSA}^{BSA}(\bar{a}, B, \bar{c}, D) \) and the upper right black point to the tangent space \( Q_{(\bar{a}, B, \bar{c}, D)}^{BSA} \) where the minimal system \( (\bar{a}, B, \bar{c}, D) \) equals \( (1/\sqrt{2}, 1/\sqrt{2}, 1/\sqrt{2}, -1/\sqrt{2}) \). The set \( BSA(1) \) is given by the union of the two unit circles excluding the intersections with the \( a \)-axis, respectively, yielding four disconnected components. The tangent space \( Q_{(\bar{a}, B, \bar{c}, D)}^{BSA} \) is given by the red straight line.

(d) The case \( n = s = 1 \) and \( m = 0 \) where \( S(1) = \mathbb{R}^3 \): The affine plane corresponds to the \( L_{r}^{\perp} \) \( k \)-equivalence class \( E_{\infty}(a, B, \bar{c}, D) \), excluding the \( B \)-axis. The tangent space \( Q_{(\bar{a}, B, \bar{c}, D)}^{\infty} \) is given by the same affine plane and the orthogonal complement \( Q_{(\bar{a}, B, \bar{c}, D)}^{\infty, \perp} \) corresponds to the red straight line. The system \((\bar{a}, B, \bar{c}, D) = (0, 3, 1, 1) \) is represented by the black point.
Chapter 4

Parametrizations of linear dynamic systems

The aim of this chapter is to compare different parametrizations, mainly with respect to their topological and geometrical properties, in a unified framework. As has been mentioned in sections (1.4) and (1.5) in the introductory chapter 1, the parametrization problem for linear dynamic systems may be considered using, for instance, ARMAX or state-space representations. In spite of the fact that the same set of rational transfer functions can be described in both frameworks – see lemmas (1.4.1) and (1.5.1) –, the classes of observational equivalence are different.

Commencing from \( \mathbb{M}(n) \), we could distinguish between four different approaches to the parametrization problem:

(i) First, we could use the full state space parametrization where all entries in the state-space matrices are free. In this situation we do not have identifiability because there are \( n^2 \) essentially unnecessary coordinates. This parametrization will be treated in section (4.1).

(ii) Second, we could make use of canonical forms which reduce the number of free parameters to be estimated. In the state-space case, we will treat echelon state-space forms in section (4.2) and different types of balanced state-space forms in sections (4.6), (4.7) and (4.8). In the ARMAX case, the echelon and reversed echelon ARMAX parametrizations are discussed in section (4.4).

(iii) Another approach is to use a description of \( \mathbb{M}(n) \) by overlapping parametrizations. Again, there are no redundant parameters in these parametrizations. In fact, we will discuss one (well known) construction of a system of local coordinates for the set \( \mathbb{M}(n) \) and thereby show the manifold structure of \( \mathbb{M}(n) \). In the state-space case, this is done in section (4.3) and in the ARMAX case in section (4.5).

(iv) Finally, we treat the class of so called data driven local parametrizations, the investigation of which is a main contribution of this thesis. For this class of parametrizations, there is no a priori fixed parameter space, but the parametrization is adapted in the course of a numerical search procedure. The main idea is to choose the parametrization in a data driven way in order to obtain favourable properties. This was the motivation for the introduction of the "dynamic minimal parametrization" in (McKelvey and Helmersson, 1999). We shall use the term data driven local coordinates, or, briefly, DDLC instead; DDLC will be introduced and investigated in detail in section (4.9).

Potential and actual drawbacks of DDLC were the main motivation for modifying the DDLC concept. This modification leads to an alternative analogous parametrization for a suitable concentrated likelihood-type criterion function. The concentration step can be done by a generalized least squares step. An obvious advantage is the reduced number of parameters resulting in less computational burden. Of course, the criterion function itself is changed by the concentration step. The resulting new parametrization is described in section (4.10) and will be called \( s1sDDLC \).

A drawback of DDLC is that an unfavourable choice of the state-space representation in the initial equivalence class can yield arbitrarily large condition numbers e.g. of the Fisher information
matrix of the parameter estimates in the region of primary interest, that is in the neighborhood of
the parameter vector corresponding to the initial system. Making use of the s1sDDLc parametrization,
this drawback is then eliminated by a slight modification of an arbitrary iterative estimation
algorithm: The choice of the initial state-space representation is done in a favourable (but still
non unique) way, and the sequence of system representatives in the course of the iterative search
procedure is then constructed in an easily implementable manner such that each system shares
certain favourable orthogonality properties. This approach is called orthoDDLc and will be treated
in section (4.11).

In section (4.12), the state-space parametrizations considered in this chapter are explicitly given for a
particularly simple case. Finally, section (4.13) presents a number of plots for this simple case.

4.1 Full state-space parametrization

4.1.1 Introduction

The attentive reader may have noticed that the term parametrization itself has not been defined yet.
One possibility of doing this would be the following:
A bijective mapping \( \psi : V \rightarrow T, (l,k) \mapsto \tau \) where \( V \subseteq \mathbb{M}(n), \tau \in T \subseteq \mathbb{R}^d, d \in \mathbb{N} \), is called a
parametrization of \( V \). The vector \( \tau = \psi(l,k) \) is called a parameter vector, its components are called
parameters and if the set \( T \) is open in \( \mathbb{R}^d \), the parameters are called free.

On the one hand, a parametrization in the sense of this definition need not yield a well posed estimation
problem because \( (T_{pt}) \) convergence of the transfer function estimates to the true transfer function (see
theorem (2.3.4)) need not imply convergence of the parameter estimates to the true parameter vector\(^1\).
On the other hand, the definition is already quite restrictive: It would not allow us to use the terms full
state-space parametrization or driven driven local parametrization (as will be seen below).

In the sequel we will thus interpret the term parametrization in a very wide sense. For instance, we
will also call non injective mappings from a set \( T \subseteq \mathbb{R}^d, d \in \mathbb{N} \), into a set \( V \) of transfer functions a
parametrization. At the same time it is clear that the main effort will be to show desirable properties of
such parametrizations.

A (seemingly trivial) approach to parametrize transfer functions in \( \mathbb{M}(n) \) is described in the following

**Theorem 4.1.1 (Full state-space parametrization).** Let a transfer function \( (l,k) \in \mathbb{M}(n) \) be given.
Then there exists a state-space realization \( (A,B,C,D,K) \in S_m(n) \). Conversely, let a state-space
representation \( (A,B,C,D,K) \in S_m(n) \) be given. Then \( (A,B,C,D,K) \) is a state-space realization of a transfer
function in \( \mathbb{M}(n) \).

**Proof.** The theorem follows from lemma (1.4.2) and remark (1.4.3) below definition (1.4.1). \( \square \)

**Remark 4.1.1.** Evidently, \( S_m(n) \) is taken to be the parameter space for the full state-space parametrization.
Note, however, that the structure of \( S_m(n) \) and its corresponding set of transfer functions \( \mathbb{M}(n) \) is of
interest: Is \( S_m(n) \) open in \( \mathbb{R}^{m^2+2ns+2m(n+1)} \), which transfer functions are described on the boundary of
\( S_m(n) \) (if there is any), is the parameter space bounded etc.? Questions of this type are important for
the actual identification procedure and will be treated for all parametrizations introduced in this thesis.

An illustrative example

We consider the case where \( n = s = 1 \) and \( m = 0 \); no exogenous inputs are present and thus \( l(z) \) vanishes.
Clearly, \( S(1) = \mathbb{R}^3 \) and \( S_m(1) = \{(a,k,c) \in \mathbb{R}^3 : (a,k,c) \text{ is minimal } \} = \mathbb{R}^3 \setminus \{(x,0,z), x \in \mathbb{R}, z \in \mathbb{R} \} \cup \{(x,y,0), x \in \mathbb{R}, y \in \mathbb{R} \} \}. \) In other words, non minimal systems are represented by the union
of the planes given by \( k = 0 \) and \( c = 0 \), respectively. Clearly, \( S_m(1) \) is open and dense in \( \mathbb{R}^3 \). All systems
in the union of the two planes described above correspond to the trivial transfer function \( k(z) = 1 \), i.e.

\(^1\) Observe, e.g., that \( \psi \) need not be a homeomorphism.
they comprise one equivalence class in \( \tilde{S}_m(1) \setminus S_m(1) \). The \( k \)-equivalence classes within \( S_m(1) \) are given by hyperbolas with two disconnected branches; see remark (3.2.1) and plot (a) in figure (3.1) in section (3.7). Clearly, \( \pi(\tilde{S}_m(1)) = \pi(\mathbb{R}^3) = M(1) \cup M(0) \).

### 4.1.2 Topological and geometrical properties

The following theorem states the main topological and geometrical properties of the full state-space parametrization:

**Theorem 4.1.2.** The full state-space parametrization as given in theorem (4.1.1) above has the following properties:

(i) \( S_m(n) \) is open and dense in \( S(n) = \mathbb{R}^{n^2 + 2ns + m(n+s)} \).

(ii) The mapping \( \pi|_{S_m(n)} : S_m(n) \to M(n) \) is an open mapping, i.e. the image \( \pi(O) \) of any open subset \( O \subset S_m(n) \) is open in \( M(n) \).

(iii) \( \pi(\tilde{S}_m(n)) = \bigcup_{i \leq n} M(i) \).

(iv) For every \((l, k) \in M(n)\), the \((l, k)\)-equivalence class \( E(A, B, C, D) \) in \( S_m(n) \) constitutes a \( n^2 \) dimensional real analytic manifold consisting of two disconnected components and is given by (3.1) in section (3.2).

(v) \( M(n) \) is open (and trivially dense) in \( \tilde{M}(n) \).

(vi) \( \pi(\tilde{S}_m(n)) = \tilde{M}(n) \).

**Proof.** (i) Let us consider the mapping \( \Delta : S(n) \to \mathbb{R} \) attaching \( \det(W^nW^n) = (A, B, C, D, K) \) where \( W^n = O^n \cap \mathbb{R}^{n^2} \) and \( C^n = C^n \cap \mathbb{R}^{n^2} \) and \( O^n \) and \( C^n \) are given in definition (1.4.2). Note that \( W^n \) and \( W^n \) have full rank if and only if \( (A, B, C, D, K) \) is minimal. Moreover, the determinant of \( W^nW^n \) is a polynomial in the entries of the system matrices and thus analytic (and trivially continuous). Openness of \( S_m(n) \) in \( S(n) \) is now straightforward as \( S_m(n) = \Delta^{-1}(\mathbb{R} \setminus \{0\}) \) is the inverse image of an open set and \( \Delta \) is continuous. Denseness of \( S_m(n) \) in \( S(n) \) follows from a well known result for analytic functions: \( \Delta(A, B, C, D, K) = 0 \) can only hold true on a thin subset of \( S(n) \) (\( \Delta \) cannot vanish everywhere in \( S(n) \)).

(ii) Let us assume that \( \pi(O) \) were not open in \( M(n) \). Then one can find a transfer function \((l_0, k_0) \in \pi(O)\) and a sequence \((l_t, k_t) \to (l_0, k_0) \) with \((l_t, k_t) \in M(n) \setminus \pi(O) \) for all \( t \in \mathbb{N} \). Thus, the inverse image \( \pi|_{S_m(n)}^{-1}(l_t, k_t) \) satisfies \( \pi|_{S_m(n)}^{-1}(l_t, k_t) \cap O = \emptyset \). Let us now consider an arbitrary point \((A_0, B_0, C_0, D_0) \in \pi|_{S_m(n)}^{-1}(l_0, k_0) \cap O \). Clearly one can find an \( \varepsilon > 0 \) such that the open ball \( K_{\varepsilon}(A_0, B_0, C_0, D_0) \) with radius \( \varepsilon \) and center \((A_0, B_0, C_0, D_0)\) satisfies \( K_{\varepsilon}(A_0, B_0, C_0, D_0) \subset O \). The condition above then means that

\[
\inf_{(A_t, B_t, C_t, D_t) \in \pi|_{S_m(n)}^{-1}(l_t, k_t)} \|[A_0, B_0, C_0, D_0] - (A_t, B_t, C_t, D_t)|| \geq \varepsilon \quad (4.1)
\]

where \((A_t, B_t, C_t, D_t)\) is an arbitrary point in \( \pi|_{S_m(n)}^{-1}(l_t, k_t) \). This is shown to yield a contradiction: It is well known and will be briefly discussed in section (4.3) that \( M(n) \) has the structure of a real analytic manifold of dimension \( 2ns + m(n + s) \). Thus, one can always find a coordinate neighborhood \( U_\alpha \) containing the transfer function \((l_0, k_0)\) together with a homeomorphism (coordinate mapping) \( \psi_\alpha^{(1)} \) from \( U_\alpha \) onto an open subset of \( \mathbb{R}^{2ns + m(n + s)} \), \( T_\alpha^{(1)} \) say; see theorem (4.3.2). Convergence of \((l_t, k_t)\) to \((l_0, k_0)\) therefore implies convergence of the corresponding parameters \( \tau_\alpha^{(1)} \in T_\alpha^{(1)} \) to \( \psi_\alpha^{(1)}(l_0, k_0) \) and thus convergence of the corresponding system matrices. Next, one considers the unique transformation \( T_0 \in GL(n) \) mapping the system matrices corresponding to \( \psi_\alpha^{(1)}(l_0, k_0) \) to the chosen representative \((A_0, B_0, C_0, D_0) \in \pi|_{S_m(n)}^{-1}(l_0, k_0) \cap O \). This mapping is also continuous, which finally implies convergence of the corresponding points \((A_t, B_t, C_t, D_t) \to (A_0, B_0, C_0, D_0)\), clearly yielding a contradiction to (4.1).

(iii) From (i) it is clear that any system at the boundary \( \tilde{S}_m(n) \setminus S_m(n) \) corresponds to a transfer function of order strictly less than \( n \), implying that \( \pi(\tilde{S}_m(n)) \subset \bigcup_{i \leq n} M(i) \). On the other hand, any transfer function in \( M(i) \) has a state-space representation with state dimension \( i \) by lemma (1.4.2), and
this system can always be written as an (uncontrollable and/or unobservable) system with augmented state dimension \( n > i \). Hence, \( \cup_{i \leq n} M(i) \subset \pi(\bar{S}_m(n)) \), showing the result.

(iv) This has been shown in the proof of lemma (3.2.1).

(v) This is a classical result obtained in (Hannan and Deistler, 1988), for instance.

(vi) From (iii) we know that \( \pi(\bar{S}_m(n)) = \cup_{i \leq n} M(i) \). Thus it remains to show that \( \bar{M}(n) = \cup_{i \leq n} M(i) \). If \( (l_0, k_0) \in \bar{M}(n) \), then, by definition, there exists a sequence \( (l_t, k_t) \in M(n) \) such that \( (l_t, k_t) \to (l_0, k_0) \), or, for the corresponding Hankel matrices, \( H_t \to H_0 \). As \( rk(H_t) = n \) for all \( t \in \mathbb{N} \), all minors of \( H_t \) larger than \( n \times n \) must be zero and thus, by the continuity of the determinant, this must also hold true for \( H_0 \). This shows that \( (l_0, k_0) \) has order smaller than or equal to \( n \), i.e. \( \bar{M}(n) \subset \cup_{i \leq n} M(i) \). Conversely, if \( (l_0, k_0) \in M(i) \) for \( i < n \), then by (iii) and (i) there exists a sequence of state-space realizations in \( S_m(n) \) such that \( (A_i, B_i, C_i, D_i, K_i) \to (A_0, B_0, C_0, D_0, K_0) \in S(n) \), where \( (A_0, B_0, C_0, D_0, K_0) \) is chosen to be an arbitrary (non minimal) realization of \( (l_0, k_0) \). Continuity of \( \pi \) shows that the corresponding sequence of transfer functions satisfies \( (l_t, k_t) \to (l_0, k_0) \) where \( (l_t, k_t) \in M(n) \) for all \( t \in \mathbb{N} \). Hence, \( (l_0, k_0) \in \bar{M}(n) \) and \( \cup_{i \leq n} M(i) \subset \bar{M}(n) \), which completes the proof.

\[ \square \]

Remark 4.1.2. We give a short discussion of these results:

(i) This means that the parameters are really free in the sense that they are not restricted to a thin subset of \( \mathbb{R}^{n^2 + 2ns + m(n + s)} \). More specifically, almost any point in \( S(n) \) corresponds to a transfer function in \( \bar{M}(n) \). Note that openness is an important requirement for many numerical optimization procedures to work properly.

(ii) yields a nice picture: open sets in the open and dense subset \( S_m(n) \) of the Euclidean space \( S(n) \) are mapped onto open sets in the transfer function space \( \bar{M}(n) \). This result will also be important for the proof of statement (ii) in theorem (4.9.1).

(iii) The closure of the parameter space \( S_m(n) \) – note that \( \bar{S}_m(n) = \mathbb{R}^{n^2 + 2ns + m(n + s)} \) by (i) – corresponds to all transfer functions of McMillan degree less than or equal to \( n \).

(iv) Clearly, this implies that the full state-space parametrization is not identifiable. More specifically, the \((l, k)\)-equivalence classes \( \mathcal{E}(A, B, C, D) \) have the structure of smooth manifolds of dimension \( n^2 \).

(v) The set \( \bar{M}(n) \) is always open in its closure. This is important in connection with the coordinate free consistency result in section (2.3.4): If a sequence of likelihood estimates is known to satisfy \( (l_t, k_t) \to (l_0, k_0) \), where \( (l_0, k_0) \in \bar{M}(n) \) and \( (l_t, k_t) \in \bar{M}(n) \), then \( (l_t, k_t) \) will be in \( \bar{M}(n) \) from a certain \( t \geq T_0 \) onwards. This means that the transfer functions \( (l_t, k_t) \) for \( t \geq T_0 \) can be represented in the parameter space \( S_m(n) \). However, consistency in terms of the parameter estimates cannot hold true in this case because there is no well defined (continuous) mapping from the transfer function space \( \bar{M}(n) \) to the parameter space \( S_m(n) \) due to the non identifiability of the full state-space parametrization; see (iv) above.

(vi) states that the same transfer functions are described in the closure of the parameter space \( S_m(n) \) and in the closure of the corresponding set of transfer functions \( \bar{M}(n) \); this will not be true in general for other parametrizations. Moreover, note that (vi) and (iii) imply that \( \bar{M}(n) = \cup_{i \leq n} M(i) \).

Remark 4.1.3. It would also be interesting to investigate the structure of the equivalence classes corresponding to non minimal systems at the boundary of \( S_m(n) \), i.e. in the set \( \bar{S}_m(n) \setminus S_m(n) \). However, these equivalence classes are given by non linear restrictions and are thus difficult to describe in general. For a motivation of this statement, consider, e.g., the equivalence class corresponding to the trivial transfer function \( (l, k) = (D, I) \). Here, \( CB = CAB = \cdots = CA^{n-1}B = 0 \) must hold true, which amounts to (highly) non-linear restrictions on the matrix entries.

### 4.2 Echelon state-space forms

#### 4.2.1 Introduction

Echelon forms are based on the selection of the first linearly independent rows of the Hankel matrix \( \mathcal{H} \) corresponding to a given transfer function \( (l, k) \). This choice yields a unique Kronecker index \( \alpha = \)
(n_1, \ldots, n_s) indicating that the rows h_1, \ldots, h_{(m-1)s+1}, \ldots, h_s, \ldots, h_{(n-1)s+1} are the first basis rows for the row space of the Hankel matrix \( \mathcal{H} \); here \( h_i \) denotes the \( i \)th row of \( \mathcal{H} \). Then expressing the rows \( h_{n_i+1}, i = 1, \ldots, s \) as linear combinations of all preceding basis rows yields the echelon form.

To be more precise, let us consider a transfer function \( (l, k) \in \mathbb{M}(n) \) for some \( n \in \mathbb{N} \). Let

\[
\mathcal{H} = \begin{pmatrix}
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & h(1,1) & \cdots & \cdots \\
  \vdots & h(1,2) & \cdots & \cdots \\
  \vdots & h(2,1) & \cdots & \cdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
\end{pmatrix}
\]

be the corresponding Hankel matrix, where \( h(i,j) \) denotes the \( j \)th row in the \( i \)th block row of \( \mathcal{H} \). As already mentioned, \( \mathcal{H} \) together with \( l(0) = D \) is in a one-to-one correspondence with the transfer function \( (l, k) \). By remark (1.4.5), \( r(k(\mathcal{H})) = n \) must hold true. Starting with the first row of \( \mathcal{H} \) and continuing row by row, we select the first \( n \) linearly independent basis rows of \( \mathcal{H} \). It follows from the structure of the Hankel matrix \( \mathcal{H} \) that once a \( h(i,j) \) is not selected as a basis row (i.e., \( h(i,j) \) is in the linear span of the rows selected before), then \( h(i+1,j), h(i+2,j), \) etc. will also not be selected (i.e., \( h(i+1,j), h(i+2,j), \) etc. are in the linear span of the rows selected before). This selection procedure yields a unique Kronecker index \( \alpha = (n_1, \ldots, n_s) \) indicating that the first row was selected in the first \( n_1 \) blocks, the second row was selected in the first \( n_2 \) blocks, etc. Note that \( |\alpha| = \sum_{i=1}^{s} n_i = n \) holds true. The \( n \) selected rows are then reordered to obtain the matrix \( \mathcal{H}^{\infty}_\alpha \) given by

\[
\mathcal{H}^{\infty}_\alpha = \begin{pmatrix}
  \vdots & h(1,1) & \cdots & \cdots \\
  \vdots & h(n_1,1) & \cdots & \cdots \\
  \vdots & h(1,2) & \cdots & \cdots \\
  \vdots & h(n_2,1) & \cdots & \cdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
\end{pmatrix} \in \mathbb{R}^{n \times \infty} \quad \mathcal{H}^{\infty}_\alpha = \begin{pmatrix}
  \vdots & h(2,1) & \cdots & \cdots \\
  \vdots & h(n_1+1,1) & \cdots & \cdots \\
  \vdots & h(2,2) & \cdots & \cdots \\
  \vdots & h(n_2+1,1) & \cdots & \cdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
  \vdots & \vdots & \vdots & \vdots \\
\end{pmatrix} \in \mathbb{R}^{n \times \infty}
\]

(4.2)

Note that the last row in each block of \( \mathcal{H}^{\infty}_\alpha \) has not been selected as a basis row. If \( n_i = 0 \), then no \( h(j,i) \) occurs in \( \mathcal{H}^{\infty}_\alpha \) or in \( \mathcal{H}^{\infty}_{\alpha+1} \). By \( \mathcal{H}^{\alpha}_{\alpha} \in \mathbb{R}^{n \times (m+s)} \) we denote the submatrix of \( \mathcal{H}^{\infty}_\alpha \) consisting of its first \( (m+s) \) columns; \( \mathcal{H}^{\infty}_{\alpha+1} \) is the matrix consisting of the first block row of \( \mathcal{H} \) (i.e., without reordering of rows). It is now straightforward to construct a unique decomposition of \( \mathcal{H} = \mathcal{O} \mathcal{C} \) by setting \( \mathcal{C} = \mathcal{H}^{\infty}_\alpha \). As \( \mathcal{C} \) contains \( n \) basis rows of \( \mathcal{H} \), the matrix \( \mathcal{O} \) is uniquely determined thereby. With this unique decomposition of \( \mathcal{H} \), the system matrices are easily extracted: \( (B, K) \) is the first block column of \( \mathcal{C} \) (see (4.5)), \( C \) is the first block row of \( \mathcal{O} \) (see (4.4)) and is easily seen to be of the form given in (4.7). Finally, \( A \) is obtained from the shift invariance structure of \( \mathcal{C} \): \( A \mathcal{C} = C^\rightarrow \) where \( C^\rightarrow \) is the matrix obtained by omitting the first block column of \( \mathcal{C} \). As \( \mathcal{C} = \mathcal{H}^{\infty}_\alpha \), we have \( C^\rightarrow = \mathcal{H}^{\infty}_{\alpha+1} \), yielding (4.3). Note that \( A \) is uniquely determined by this equation and that additionally \( A \) is of the form described in (4.7) because the rows \( h(n_i+1,i) \), \( i = 1, \ldots, s \) are expressed as linear combinations of all preceding basis rows:

**Theorem 4.2.1** (Echelon state-space canonical form). Let a transfer function \( (l, k) \in \mathbb{M}(n) \) with Kronecker index \( \alpha = (n_1, \ldots, n_s) \) be given and let us denote the whole set of transfer functions with this Kronecker index by \( V_\alpha \). Then we obtain a unique minimal state-space realization by
\[
A^\infty \alpha = \mathcal{H}_\alpha^\infty \\
C^\infty \alpha = \mathcal{H}_1^\infty \\
(B, K) = \mathcal{H}_\alpha^1 \\
D = l(0)
\]

where \((A, B, C, D, K)\) have the following form:

\[
A = (A_{ij})_{i,j \in \{1, \ldots, s\}} \text{ and } A_{ij} \in \mathbb{R}^{n_i \times n_j}
\]

\[
A_{ij} = \begin{pmatrix}
0 & I_{n_i-1} & \cdots & a_{ij}(n_i - 1) \\
& a_{ij}(0) & \cdots & a_{ij}(n_i - 2) \\
& \vdots & \ddots & \vdots \\
& a_{ij}(0) & \cdots & a_{ij}(n_j - 1) & 0
\end{pmatrix}
\]

\[
B = \begin{pmatrix}
b_1 & \cdots & b_m \\
& \vdots & \vdots \\
& b_1 & \cdots & b_m
\end{pmatrix}, \quad K = \begin{pmatrix}
k_1 & \cdots & k_n \\
& \vdots & \vdots \\
& k_1 & \cdots & k_n
\end{pmatrix}
\]

\[
C = \begin{pmatrix}
1, 0, \ldots, 0 \\
0, 1, \ldots, 0 \\
\vdots \\
0, 0, \ldots, 1
\end{pmatrix}
\]

\[
D = \begin{pmatrix}
d_1 & \cdots & d_m \\
& \vdots & \vdots \\
& d_1 & \cdots & d_m
\end{pmatrix}
\]

where \(n_{ij}\) is given by (4.8) below. Conversely, let a parameter vector \(\tau_\alpha\) consisting of the entries in the system matrices \((A, B, C, D, K)\) that are not a priori restricted to zero or one be given. Let \(T_\alpha\) be the parameter space containing all vectors \(\tau_\alpha\) such that \(\varphi_\alpha(\tau_\alpha)\) is minimal where

\[
\varphi_\alpha : T_\alpha \rightarrow S_m(n) \\
\tau_\alpha \mapsto (A(\tau_\alpha), B(\tau_\alpha), C(\tau_\alpha), D(\tau_\alpha), K(\tau_\alpha))
\]

is given by (4.7). Then \(\varphi_\alpha(\tau_\alpha), \tau_\alpha \in T_\alpha\) is an echelon state-space realization of a transfer function in \(V_\alpha\).

**Proof.** For a detailed proof, see the proof of theorem 2.5.2 in (Hannan and Deistler, 1988).

**Remark 4.2.1.** Note that

\[
n_{ij} = \min(n_i + 1, n_j) \text{ for } j < i \\
n_{ij} = \min(n_i, n_j) \text{ for } j \geq i
\]

and the \(a_{ij}(k)\) are the coordinates of the row \(h(n_i + 1, i)\) with respect to the preceding basis rows according to \(\alpha\). Precisely, \(a_{ij}(k)\) denotes the coefficient at row \(h(k + 1, j)\) in the linear combination of \(h(n_i + 1, i)\) in terms of all preceding basis rows. If \(n_i = 0\), then the \(i\)th row of \(C\) consists of zeros except for the columns numbered \((\sum_{j=1}^{i-1} n_j) + 1\) where \(j\) runs over values for which \(n_j > 0, j < i\) and \(n_0 = 0\). In these columns the elements \(a_{ij}(0)\) appear; see the right hand side of (4.10), for instance.

**Remark 4.2.2.** The mapping \(\varphi_\alpha\) from the parameter vectors \(\tau_\alpha\) to the state-space matrices is almost trivial: It simply inserts the components of \(\tau_\alpha\) in the correct positions of \((A, B, C, D, K)\). As the other matrix entries are fixed to either zero or one, \(\varphi_\alpha(T_\alpha)\) is a subset of an affine subspace when embedded into \(S(n)\). For \(n = s = 1\) and \(m = 0\), this is shown in figure (4.2) in section (4.13). Clearly, \(\varphi_\alpha(T_\alpha) = \varphi_\alpha(T_\alpha)\), and in the sequel we will use the symbol \(\pi(T_\alpha) = \pi(\varphi_\alpha(T_\alpha))\) with slight abuse of notation.

Note that for \(|\alpha| = \sum_{i=1}^{s} n_i = n\), the number of free parameters \(d_\alpha\) will be different for different \(\alpha\); see theorem (4.2.2) where the dimension of the parameter space is given in (i). By

\[
\psi_\alpha : V_\alpha \rightarrow T_\alpha
\]

we denote the echelon state-space parametrization attaching the free parameters \(\tau_\alpha\) to the transfer functions \((l, k) \in V_\alpha\).
4.2. ECHELON STATE-SPACE FORMS

An illustrative example

We give an illustrative example for the echelon state-space parametrization in the MIMO case: For \( m = 2, s = 2, n = 4 \) and \( \alpha = (2,2) \), 28 free parameters appear on the left hand side, whereas for \( \alpha = (4,0) \), for instance, we get 25 free parameters on the right hand side:

\[
A = \begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix},
A = \begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
a_{11}(0) & a_{11}(1) & a_{11}(2) & a_{11}(3)
\end{pmatrix}
\]

\[
B = \begin{pmatrix}
b_1 & b_2 \\
b_2 & b_4 \\
b_3 & b_2 \\
b_4 & b_2 \\
b_3 & b_2 \\
b_4 & b_2 \\
b_3 & b_2 \\
b_4 & b_2 \\
\end{pmatrix},
K = \begin{pmatrix}
k_1 & k_2 \\
k_2 & k_2 \\
k_3 & k_2 \\
k_4 & k_2 \\
k_3 & k_2 \\
k_4 & k_2 \\
k_3 & k_2 \\
k_4 & k_2 \\
\end{pmatrix},
B = \begin{pmatrix}
b_1 & b_2 \\
b_2 & b_2 \\
b_3 & b_2 \\
b_4 & b_2 \\
b_3 & b_2 \\
b_4 & b_2 \\
b_3 & b_2 \\
b_4 & b_2 \\
\end{pmatrix},
K = \begin{pmatrix}
k_1 & k_2 \\
k_2 & k_2 \\
k_3 & k_2 \\
k_4 & k_2 \\
k_3 & k_2 \\
k_4 & k_2 \\
k_3 & k_2 \\
k_4 & k_2 \\
\end{pmatrix}
\]

\[
C = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{pmatrix},
D = \begin{pmatrix}
d_1 & d_2 \\
d_2 & d_2 \\
\end{pmatrix},
D = \begin{pmatrix}
d_1 & d_2 \\
d_2 & d_2 \\
\end{pmatrix}
\]

(4.10)

4.2.2 Topological and geometrical properties

The following theorem states the main topological and geometrical properties of the echelon state-space parametrization:

**Theorem 4.2.2.** The echelon state-space parametrization

\[
\psi_\alpha : V_\alpha \rightarrow T_\alpha \\
(l,k) \mapsto \tau_\alpha
\]

where \( \alpha = (n_1, \ldots, n_s) \), has the following properties:

(i) \( T_\alpha \) is an open and dense subset of \( \mathbb{R}^{d_\alpha} \), where

\[
d_\alpha = |\alpha|(s + m + 1) + ms + \sum_{i,j,<i} \{\min(n_i,n_j) + \min(n_j,n_i)\}
\]

(ii) \( \psi_\alpha : V_\alpha \rightarrow T_\alpha \) is a \( (T_{pl'}) \) homeomorphism.

(iii) \( \{V_\alpha | \alpha = n \} \) is a disjoint partition of \( M(n) \) containing \( \binom{n+s-1}{s-1} \) sets.

(iv) \( \pi(\mathbb{T}_\alpha) = \bigcup_{\beta \leq \alpha} V_\beta \) where \( \beta = (m_1, \ldots, m_s) \leq (n_1, \ldots, n_s) = \alpha \) means that all \( m_i \leq n_i \) and \( \beta < \alpha \) means that at least one strict inequality occurs.

(v) For every \( (l,k) \in V_\beta, \beta \leq \alpha \), the \( (l,k) \)-equivalence class in \( \mathbb{T}_\alpha \) is an affine subspace of dimension

\[
\sum_{i=1}^{s} \sum_{j=1}^{s} (n_{ij} - n'_{ij})
\]

where \( n_{ii} := n_i, n_{ij} \) is given in (4.8) and \( n'_{ij} = \min(n_i+1,m_j) \) for \( j < i \) and \( n'_{ij} = \min(n_i,m_j) \) for \( j \geq i \).

Here, \( \alpha = (n_1, \ldots, n_s) \) and \( \beta = (m_1, \ldots, m_s) \).

(vi) \( V_\alpha \) is open in \( V_\alpha \), \( V_\alpha \) is the Kronecker index corresponding to the special case when the first \( n \) rows of \( \mathbb{H} \) are linearly independent.

(vii) \( \pi(\mathbb{T}_\alpha) \subset \mathbb{V}_\alpha \) and equality holds for \( s = 1 \).
**Proof.** See the proofs of theorems 2.5.3 and 2.6.5 in (Hannan and Deistler, 1988).

**Remark 4.2.3.** We give a short discussion of these results:

(i) This means that the parameters are really free in the sense that they are not restricted to a thin subset of $\mathbb{R}^d$. More specifically, almost any point in $\mathbb{R}^d$ corresponds to a transfer function in $V_\alpha$.

(ii) Assures bijectivity of the parametrization on the pieces $V_\alpha$. Note that the inverse mapping $\psi^{-1} = \pi \circ \varphi_\alpha$ is known to be continuous because both $\pi$ and $\varphi_\alpha$ are continuous. As consistency is preserved under continuous transformations, consistency in terms of transfer functions (coordinate free consistency; see section 2.3.4) implies consistency in terms of the parameter estimates.

(iii) Implies that the set of mappings $\{\psi_\alpha(\cdot) = n\}$ constitutes a canonical form for $\mathbb{M}(n)$. Note that for $s = 1$ we get only one piece $V_\alpha$, i.e., $\mathbb{M}(n) = V_{(n)}$. This is evident because the first $n$ rows of the Hankel matrix must necessarily be its basis rows, which follows from the properties of the Hankel matrix $\mathcal{H}$: The block rows of $\mathcal{H}$ comprise a single row for the case $s = 1$, and once a row $h(i, 1)$ is selected as a basis row, then $h(i + 1, 1)$, $h(i + 2, 1)$, etc. will also not be selected.

(iv) The closure of the parameter space $T_\alpha$ – note that $\mathcal{T}_\alpha = \mathbb{R}^d$ by (i) – corresponds to transfer functions of equal and lower McMillan degrees. The boundary $\overline{T}_\alpha \setminus T_\alpha$ is described in

(v) and is seen to consist of equivalence classes corresponding to lower order transfer functions with Kronecker indices smaller than $\alpha$. These equivalence classes have the form of affine subspaces in $\mathbb{R}^d$, implying that, of course, $\psi_\alpha$ is no longer well defined for these lower degree transfer functions.

(vi) The set $V_\alpha$ is always open in its closure. This is important again in connection with the consistency result in section 2.3.4; see (v) in remark 4.1.2. The second statement involves a particular Kronecker index $\alpha_g$ corresponding to the special case when the first $n$ rows of the Hankel matrix are linearly independent and therefore trivially coincide with its first basis rows. The dimension of the parameter vector is $d_{\alpha_g} = 2ns + m(n + s)$, i.e., equal to the dimension of the manifold $\mathbb{M}(n)$ itself. Here, we get $\overline{V}_{2} = \mathbb{M}(n) = \cup_{i < n} \mathbb{M}(i)$, i.e., transfer functions with this particular Kronecker index are not only open in their closure but also form a dense subset of all transfer functions of McMillan degree less than or equal to $n$. This is the reason for calling this $V_\alpha$ a generic neighborhood of $\mathbb{M}(n)$.

(vii) Considering statements (iv) and (v), $\overline{V}_\alpha$ contains at least all lower degree transfer functions with comparable Kronecker indices but, in the multi output case, more than just these transfer functions. As an example, consider the generic neighborhood $V_{\alpha_g}, \alpha_g = (2, 2)$ on the left hand side of (4.10). Although $\overline{V}_{(2, 2)} = \mathbb{M}(2), \pi(\mathcal{T}_{(2, 2)}) = \pi(\mathbb{R}^8)$ does not contain transfer functions in $V_{(4,0)}$ as given on the right hand side of (4.10). Lower degree transfer functions in $V_{(3,0)}$ and $V_{(0,3)}$ are also not contained in $\pi(\mathcal{T}_{(2,2)})$.

**Remark 4.2.4.** From what was said above it follows that a transfer function in $\overline{V}_\alpha \setminus V_\alpha$ can have more than one corresponding parameter vector in $\mathcal{T}_\alpha$ according to (v) or it can also have no representation in $\mathcal{T}_\alpha$ at all. Such a transfer function corresponds to the "point of infinity" in $\mathcal{T}_\alpha$. Phenoa of this kind are discussed in detail in sections (4.9.3) and (4.10.2) for the case of DDLC and DSL below.

**Remark 4.2.5.** Clearly, it can never occur that the same parameter vector at the boundary of $T_\alpha$ corresponds to more than one transfer function. This is to be understood as follows: There do not exist sequences $\tau_{\alpha,n} \to \tau_\alpha$, where $\tau_{\alpha,n} \in T_\alpha$ and $\tau_\alpha \in \overline{T}_\alpha \setminus T_\alpha$, $i = 1, \ldots, l$, converging to the same parameter vector $\tau_\alpha$ such that the corresponding limits of the transfer functions $(l_i, k_i) = \lim_{n \to \infty}(\pi(\tau_{\alpha,n}))$ are different. This is in contrast to balanced canonical forms as will be shown in an example in section 4.6 for the case of Ober’s Lyapunov balanced canonical form.

**Remark 4.2.6.** The subset of $T_\alpha$ corresponding to stable and minimum phase systems is not easy to describe in general, so that both the stability and the minimum phase assumption have to be monitored in the course of a parameter estimation algorithm.

### 4.3 An overlapping state-space parametrization

#### 4.3.1 Introduction

The derivation of an overlapping state-space parametrization may follow along the same lines as described in the previous section (4.2) for echelon canonical forms. However, the Kronecker index has to be replaced
by a so-called structure index which will also be denoted by \( \alpha \) in the sequel and still corresponds to a choice of basis rows with the property that if a row \( h(i, j) \) is not selected as a basis row (i.e. \( h(i, j) \) is in the linear span of the rows selected before), then \( h(i + 1, j), h(i + 2, j), \) etc. is also not selected (i.e. \( h(i + 1, j), h(i + 2, j), \) etc. are in the linear span of the rows selected before). Commencing from the Hankel matrix \( \mathcal{H} \) corresponding to a given transfer function \( (l, k) \), the structure index \( \alpha \) indicates that the rows of \( \mathcal{H}^\infty \alpha \) (as given in (4.2)) form a basis for the row space of \( \mathcal{H} \). In contrast to the Kronecker index, the chosen rows according to \( \alpha \) are not required to be the first basis rows of \( \mathcal{H} \). Of course, by dropping this requirement, the structure index is in general not unique for a given \( \mathcal{H} \), because the Hankel matrix can correspond to more than one structure index \( \alpha \), implying that we get an overlapping description of the model class \( U_A \). Once \( \alpha \) is chosen, however, a unique state-space realization can be constructed in the same manner as has been discussed in section (4.2) for echelon canonical forms. Note that the rows \( h(n_i + 1, i), i = 1, \ldots, s \) are now expressed as a linear combination of all basis rows in \( \mathcal{H}^\infty \alpha \), whereas for echelon state-space forms only the basis rows that precede \( h(n_i + 1, i) \) were used:

**Theorem 4.3.1 (Overlapping state-space parametrization).** Let a transfer function \( (l, k) \in \mathcal{M}(n) \) be given, let \( \alpha = (n_1, \ldots, n_s) \) be a structural index for the Hankel matrix corresponding to \( (l, k) \) and let us denote the whole set of transfer functions for which \( \alpha \) is a structure index by \( U_\alpha \). Then we obtain a minimal state-space realization, unique with respect to \( \alpha \), by

\[
A \mathcal{H}^\infty \alpha = \mathcal{H}^\infty_{\alpha + s} \quad (4.12)
\]

\[
C \mathcal{H}^\infty \alpha = \mathcal{H}^\infty_1 \quad (4.13)
\]

\[
(B, K) = \mathcal{H}^1 \quad (4.14)
\]

\[
D = l(0) \quad (4.15)
\]

where \( (A, B, C, D, K) \) have the following form:

\[
A = (A_{ij}) \quad i,j \in \{1, \ldots, s\} \quad \text{and} \quad A_{ij} \in \mathbb{R}^{n_i \times n_j}
\]

\[
A_i = \begin{pmatrix}
0 & \ldots & 0 & I_{n_i-1} \\
\vdots & \ddots & \vdots & \vdots \\
0 & \ldots & 0 & 1
\end{pmatrix}
\]

\[
A_{ij} = \begin{pmatrix}
0 & \ldots & 0 \\
\vdots & \ddots & \vdots \\
0 & \ldots & 0
\end{pmatrix}
\]

\[
A_{ij} = \begin{pmatrix}
0 & \ldots & 0 \\
\vdots & \ddots & \vdots \\
0 & \ldots & 0
\end{pmatrix}
\]

\[
B = \begin{pmatrix}
b_1 & \ldots & b_m \\
b_1 & \ldots & b_m \\
\vdots & \ddots & \vdots \\
b_1 & \ldots & b_m
\end{pmatrix}
\]

\[
K = \begin{pmatrix}
k_1 \ldots k_m \\
k_1 \ldots k_m \\
k_1 \ldots k_m \\
k_1 \ldots k_m
\end{pmatrix}
\]

\[
D = \begin{pmatrix}
d_1 & \ldots & d_m \\
\vdots & \ddots & \vdots \\
\vdots & \ddots & \vdots \\
\vdots & \ddots & \vdots 
\end{pmatrix}
\]

Conversely, let a parameter vector \( \tau_{\alpha}^{(1)} \) consisting of the entries in the system matrices \( (A, B, C, D, K) \) that are not a priori restricted to zero or one be given. Let \( T_{\alpha}^{(1)} \) be the parameter space containing all vectors \( \tau_{\alpha}^{(1)} \) such that \( \varphi_{\alpha}^{(1)}(\tau_{\alpha}^{(1)}) \) is minimal where

\[
\varphi_{\alpha}^{(1)} : T_{\alpha}^{(1)} \to S_m(n)
\]

\[
\tau_{\alpha}^{(1)} \to (A(\tau_{\alpha}^{(1)}), B(\tau_{\alpha}^{(1)}), C(\tau_{\alpha}^{(1)}), D(\tau_{\alpha}^{(1)}), K(\tau_{\alpha}^{(1)}))
\]

is given by (4.16). Then \( \varphi_{\alpha}^{(1)}(\tau_{\alpha}^{(1)}), \tau_{\alpha}^{(1)} \in T_{\alpha}^{(1)} \) is a state-space realization of a transfer function in \( U_\alpha \).

**Proof.** For a detailed proof, see the proof of theorem 2.6.2 in (Hannan and Deistler, 1988). \qed

**Remark 4.3.1.** Note that the \( a_{ij}(k) \) are the coordinates of the row \( h(n_i + 1, i) \) with respect to all basis rows according to \( \alpha \). If \( n_i = 0 \), then the ith row of \( C \) is therefore given by
\( (a_0(0), \ldots, a_i(n_i - 1), \ldots, a_{i(i-1)}(0), \ldots, a_{i(i-1)}(n_i - 1), a_{i(i+1)}(0), \ldots, a_{i(i+1)}(n_{i+1} - 1), \ldots, a_{i(s)}(n_s - 1)) \)

See the right hand side of (4.18), for instance.

**Remark 4.3.2.** In analogy to remark (4.2.2), we will use the symbol \( \pi(T^{(1)}_\alpha) = \pi(\varphi^{(1)}_\alpha(T^{(1)}_\alpha)) \) with slight abuse of notation. Again, \( \varphi^{(1)}_\alpha(T^{(1)}_\alpha) \) is a subset of an affine subspace when embedded into \( S(n) \). For \( n = s = 1 \) and \( m = 0 \), it coincides with \( \varphi_\alpha(T_\alpha) \) as shown in figure (4.2) in section (4.13).

In contrast to echelon canonical forms, for \( |\alpha| = \sum_{i=1}^s n_i = n \), the number of free parameters is the same for different \( \alpha \); see theorem (4.3.2) where the dimension of the parameter space is given in (i). By

\[
\psi^{(1)}_\alpha : U_\alpha \to T^{(1)}_\alpha
\]

(4.17) we denote the overlapping state-space parametrization attaching the free parameters \( \tau^{(1)}_\alpha \) to the transfer functions \( (l, k) \in U_\alpha \).

**An illustrative example**

We give an illustrative example for the overlapping state-space parametrization in the MIMO case: For \( m = 2, s = 2, n = 4 \) and \( \alpha = (2, 2) \), 28 free parameters appear on the left hand side, and the overlapping parametrization coincides with the echelon form. However, for \( \alpha = (4, 0) \), we get again 28 free parameters on the right hand side:

\[
\begin{align*}
A &= \begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & a_{11}(0) & a_{11}(1) & 0 \\
0 & a_{21}(0) & a_{21}(1) & 0 \\
0 & a_{31}(0) & a_{31}(1) & 0 \\
0 & a_{41}(0) & a_{41}(1) & 0 \\
\end{pmatrix}, &
A &= \begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & a_{11}(0) & a_{11}(1) & a_{11}(2) \\
0 & a_{21}(0) & a_{21}(1) & a_{21}(2) \\
\end{pmatrix} \\
B &= \begin{pmatrix}
b_{11} & b_{12} \\
b_{21} & b_{22} \\
b_{31} & b_{32} \\
b_{41} & b_{42} \\
\end{pmatrix}, &
B &= \begin{pmatrix}
b_{11} & b_{12} \\
b_{21} & b_{22} \\
b_{31} & b_{32} \\
b_{41} & b_{42} \\
\end{pmatrix} \\
K &= \begin{pmatrix}
k_{11} & k_{12} \\
k_{21} & k_{22} \\
k_{31} & k_{32} \\
k_{41} & k_{42} \\
\end{pmatrix}, &
K &= \begin{pmatrix}
k_{11} & k_{12} \\
k_{21} & k_{22} \\
k_{31} & k_{32} \\
k_{41} & k_{42} \\
\end{pmatrix} \\
C &= \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
\end{pmatrix}, &
C &= \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & a_{21}(0) & a_{21}(1) & a_{21}(2) \\
\end{pmatrix} \\
D &= \begin{pmatrix}
d_{11} & d_{12} \\
d_{21} & d_{22} \\
\end{pmatrix}, &
D &= \begin{pmatrix}
d_{11} & d_{12} \\
d_{21} & d_{22} \\
\end{pmatrix}
\end{align*}
\]

(4.18)

**4.3.2 Topological and geometrical properties**

The following theorem is a collection of results in (Hannan and Deistler, 1988). It states the main topological and geometrical properties of the overlapping state-space parametrization, including the result that \( \mathbb{M}_\alpha \) is a real analytic manifold of dimension \( 2ns + m(n + s) \):

**Theorem 4.3.2.** The overlapping state-space parametrization

\[
\psi^{(1)}_\alpha : U_\alpha \to T^{(1)}_\alpha,
\]

where \( \alpha = (n_1, \ldots, n_s) \), has the following properties:

(i) \( T^{(1)}_\alpha \) is an open and dense subset of \( \mathbb{R}^{2ns+m(n+s)} \), where \( n = |\alpha| \).
(ii) \( \psi_\alpha^{(1)} : U_\alpha \to T_\alpha^{(1)} \) is a \( (T_{pl}) \) homeomorphism.

(iii) \((U_\alpha, \psi_\alpha^{(1)}) : |\alpha| = n \) is a system of local coordinates for \( \mathbb{M}(n) \) containing \( \binom{n + s - 1}{s - 1} \) overlapping coordinate neighborhoods \( U_\alpha \), i.e. \( \mathbb{M}(n) \) is a real analytic manifold of dimension \( 2ns + m(n + s) \). Moreover, \( \mathbb{M}(n) \) consists of \( n + 1 \) disconnected components in the single input, single output case and is pathwise connected otherwise.

(iv) \( \pi(T_\alpha^{(1)}) = \bigcup_{\beta \leq \alpha} U_\beta \) where \( \beta = (m_1, \ldots, m_s) \leq (n_1, \ldots, n_s) = \alpha \) means that all \( m_i \leq n_i \) and \( \beta < \alpha \) means that at least one strict inequality occurs.

(v) For every \((l, k) \in U_\beta, \beta \leq \alpha \), the \((l, k)\)-equivalence class in \( T_\alpha^{(1)} \) is an affine subspace of dimension \( s(n - |\beta|) \).

(vi) \( U_\alpha \) is open in \( \bar{U}_\alpha \). \( \bar{U}_\alpha = \bar{\mathbb{M}}(n) \) for any \( |\alpha| = n \).

(vii) \( \pi(T_\alpha^{(1)}) \subset \bar{U}_\alpha \) and equality holds for \( s = 1 \).

Proof. See the proof of theorems 2.6.3 and 2.6.5 in (Hannan and Deistler, 1988). As the set \( \{U_\alpha, |\alpha| = n\} \) clearly covers \( \mathbb{M}(n) \), statements (i), (ii) and (vi) and the fact that coordinate transformations are analytic functions actually imply the first part of statement (iii). The latter two results in (iii) were obtained in (Brockett, 1976) and (Glover, 1975), respectively.

\[ \square \]

Remark 4.3.3. Remarks completely analogous to remarks (4.2.3) - (4.2.6) for echelon state-space parametrizations can also be made for the overlapping state-space parametrization.

Remark 4.3.4. It is clear that \( U_\alpha \supset V_\alpha \) and that equality holds if the \( n \) rows in \( \mathcal{H}_\alpha^{(1)} \) are just the first rows of the Hankel matrix \( \mathcal{H} \): Evidently, there cannot be any preceding rows spanning the row space of \( \mathcal{H} \) in this case. In particular, for \( s = 1 \) (i.e. in the single output case), we have \( \alpha = (n) \) and \( U_\alpha = V_\alpha = \mathbb{M}(n) \).

Remark 4.3.5. It can be shown that every \( U_\alpha \) in statement (iii) of theorem (4.3.2) above is needed to cover \( \mathbb{M}(n) \). However, using other systems of local coordinates, the number of coordinate neighborhoods may be reduced, but in general \( \mathbb{M}(n) \) cannot be described by a single coordinate neighborhood.

### 4.4 Echelon ARMAX and reversed echelon ARMAX forms

#### 4.4.1 Introduction

For a brief treatment of ARMAX realizations, we refer the reader to section (1.5). In (Deistler and Wang, 1989), the authors present a general framework for parametrizations of ARMAX systems including the echelon ARMAX and reversed echelon ARMAX forms of this section as well as the overlapping ARMAX parametrizations in \( z \) and \( z^{-1} \) in section 4.5 below. This paper also treats topological and geometrical properties of the parametrizations and contains many original references.

Let us start with a one-to-one relation between the transfer functions \((l, k)\) and the transfer functions \((\tilde{l}, \tilde{k})\) which are given as follows:

\[
(l, k) = \sum_{j=1}^{\infty} (L_j, K_j)z^j + (D, I)
\]

\[
(\tilde{l}, \tilde{k}) = (l(z^{-1}), k(z^{-1})) - (D, I) = \sum_{j=1}^{\infty} (L_j, K_j)z^{-j}
\]  \hspace{1cm} (4.19)

Note that \((\tilde{l}, \tilde{k})\) is also rational and, additionally, strictly proper. Clearly,

\[
(l(z), k(z)) \text{ is causal } \iff (\tilde{l}(z), \tilde{k}(z)) \text{ is strictly proper}
\]  \hspace{1cm} (4.20)

Given a matrix fraction description (MFD) of \((\tilde{l}, \tilde{k})\)
\[(\tilde{l}(z), \tilde{k}(z)) = a^{-1}(z)(\tilde{d}(z), \tilde{b}(z)) \quad (4.21)\]

one directly obtains

\[
(l(z), k(z)) = a^{-1}(z^{-1})(\tilde{d}(z^{-1}) + a(z^{-1})D, \tilde{b}(z^{-1}) + a(z^{-1}))
\]

Note that these matrices are polynomial matrices in \(z^{-1}\), i.e. they correspond to an ARMAX realization in the forward shift \(z^{-1}\). In order to obtain an MFD of the rational \((l, k)\) in terms of the backward shift \(z\), we can set

\[
(l(z), k(z)) = (\text{diag}(z^{n_1}, \ldots, z^{n_n}) \cdot a(z^{-1}))^{-1} \text{diag}(z^{n_1}, \ldots, z^{n_n})(\tilde{d}(z^{-1}) + a(z^{-1})D, \tilde{b}(z^{-1}) + a(z^{-1}))
\]

Here, \(n_i\) is the (maximum) degree of the \(i\)th row of \(a(z)\) and

\[
a(z) = \text{diag}(z^{n_1}, \ldots, z^{n_n}) \cdot a(z^{-1})
\]

\[
\tilde{d}(z) = \text{diag}(z^{n_1}, \ldots, z^{n_n})(\tilde{d}(z^{-1}) + a(z^{-1})D)
\]

\[
\tilde{b}(z) = \text{diag}(z^{n_1}, \ldots, z^{n_n})(\tilde{b}(z^{-1}) + a(z^{-1}))
\]

(4.23)

The fact that \((\tilde{l}, \tilde{k})\) is strictly proper allows us (from (4.21) and (4.19)) to write

\[
\tilde{a}(z) \left( \sum_{j=1}^{\infty} (L_j, K_j)z^{-j} \right) = (\tilde{d}(z), \tilde{b}(z))
\]

implying that the degree of the \(i\)th row of \((\tilde{d}(z), \tilde{b}(z))\) is strictly less than \(n_i\). Hence, \(a(z), b(z)\) and \(d(z)\) in (4.23) are indeed polynomial matrices! Conversely, if the row degrees \(n_i\) of \(a(z)\) are strictly greater than the corresponding row degrees of \((\tilde{d}(z), \tilde{b}(z))\), and, additionally, \(a(z)\) is row reduced, then

\[
(l(z), k(z)) = a^{-1}(z)(\tilde{d}(z), \tilde{b}(z))\]

is strictly proper, i.e. \((l(z), k(z))\) is causal. More precisely (see lemma 2.4.1 in (Hannan and Deistler, 1988) for a proof):

\[
\tilde{a}(z) \text{ is row reduced } \Rightarrow \tilde{l}(z, \tilde{k}) = a^{-1}(\tilde{d}, \tilde{b}) \text{ is strictly proper } \Leftrightarrow \text{ i. row of } (\tilde{d}(z), \tilde{b}(z)) \text{ has degree } < n_i
\]

Row reducedness is defined as follows:

**Definition 4.4.1.** A polynomial \(s \times s\) matrix \(a(z)\) is called row (column) reduced if the row end matrix \([a]_r\) (the column end matrix \([a]_c\) has full rank \(s\). Here, the \(i\)th row of \([a]_r\) (the \(i\)th column of \([a]_c\) is taken to be the \(i\)th row (the \(i\)th column) of the coefficient matrix corresponding to the power \(n_i\) where \(n_i\) is the (maximum) degree of the \(i\)th row (the \(i\)th column) of \(a(z)\).

**Remark 4.4.1.** It is easy to see that the degree of \(\text{det}(a(z))\) cannot exceed the sum of the row (column) degrees: \(\nu(\text{det}(a)) \leq \sum_{i=1}^{s} n_i = n\) and that the coefficient of \(\text{det}(a(z))\) corresponding to power \(z^n\) is given by \(\text{det}([a]_r) = \text{det}([a]_c)\). Hence, a square polynomial matrix is row (column) reduced, if and only if the degree of its determinant is equal to the sum of the row (column) degrees.

We have now answered the question as to when formula (4.23) yields an MFD of \((l, k)\), i.e. \(a(z), d(z)\) and \(b(z)\) are polynomial matrices. Next, we start from an irreducible MFD of \((\tilde{l}, \tilde{k})\) in (4.21) and treat the question as to when (4.23) will also be an irreducible MFD (see lemma 2.4.2 in (Hannan and Deistler, 1988) for a proof of the special case when \(D = 0\) which is easily seen to generalize to the case of arbitrary \(D\):
\[
\tilde{a}^{-1}(\tilde{d}, \tilde{b}) \text{ is an irreducible MFD of } (\tilde{i}, \tilde{k}) \Rightarrow \\
(4.23) \text{ is an irreducible MFD of } (l, k) \iff A(0) \text{ is nonsingular} \quad (4.25)
\]

We want to repeat that the degree \( n \) of \( \det(\tilde{a}(z)) \) is an invariant for the class of all irreducible MFDs \((\tilde{i}, \tilde{k}) = \tilde{a}^{-1}(d, b)\), in fact it is minimal among all MFDs of \((\tilde{i}, \tilde{k})\) and coincides with the order of \((l, k)\); see definition (1.3.2).

We are now ready to introduce the echelon ARMAX parametrization. The construction is analogous to the state-space case: Commencing from a transfer function \((l, k)\) with Hankel matrix \( \mathcal{H} \) and Kronecker index \( \alpha \), we express the rows \( h(n_i + 1, i), i = 1, \ldots, s \) as linear combinations of all preceding basis rows:

\[
-h(n_i + 1, i) = \sum_{j=1}^{s} \sum_{u=1}^{n_{ij}} \tilde{a}_{ij}(u-1)h(u, j) \quad i = 1, \ldots, s 
\]

where again \( n_{ij} \) is given in (4.8), defines unique coefficients \( \tilde{a}_{ij}(u) \). Writing

\[
\tilde{a}(z) = \tilde{A}(0) + \tilde{A}(1)z + \cdots + \tilde{A}(\tilde{p})z^{\tilde{p}}
\]

with \( \tilde{a}_{ij}(u) = \tilde{A}(u)_{ij}, \tilde{A}(n_i)_{ii} = 1 \) (all other elements are equal to zero) and \( \tilde{p} = \max(n_1, \ldots, n_s) \), this is a special relation of the form

\[
(\tilde{A}(0), \tilde{A}(1), \ldots, \tilde{A}(\tilde{p}), 0, \ldots)\mathcal{H} = 0_{s \times \infty}
\]

Note that (4.28) corresponds to a comparison of coefficients corresponding to all negative powers of \( z \) in equation (4.24). Given the unique polynomial matrix \( \tilde{a}(z) \), the right hand side of

\[
(\tilde{A}(0), \tilde{A}(1), \ldots, \tilde{A}(\tilde{p}), 0, \ldots) \begin{pmatrix} 0 & 0 \\ (L_1, K_1) & 0 \\ (L_2, K_2) & (L_1, K_1) \\ (L_3, K_3) & (L_2, K_2) & (L_1, K_1) \\ \vdots & \end{pmatrix} = \left((\tilde{D}(0), \tilde{B}(0)), (\tilde{D}(1), \tilde{B}(1)), \ldots \right)
\]

will of course be unique, too. Note that (4.29) corresponds to a comparison of coefficients corresponding to all non-negative powers of \( z \) in equation (4.24). Moreover, \( (\tilde{D}(i), \tilde{B}(i)) = (0, 0) \) for \( i \geq \tilde{p} \), i.e.

\[
\tilde{a}(z) = \tilde{D}(0) + \tilde{D}(1)z + \cdots + \tilde{D}(\tilde{p}-1)z^{\tilde{p}-1} \\
\tilde{b}(z) = \tilde{B}(0) + \tilde{B}(1)z + \cdots + \tilde{B}(\tilde{p}-1)z^{\tilde{p}-1}
\]

are polynomial matrices. Clearly, \( \tilde{a}^{-1}(\tilde{d}, \tilde{b}) = (\tilde{i}, \tilde{k}) \) because (4.28) and (4.29) correspond to a comparison of coefficients corresponding to (4.24) for all \( \tilde{z}^i, i = 0, \pm 1, \pm 2, \ldots \). In other words, for given \((l, k)\), we have constructed a unique MFD. This MFD has some additional properties:

(i) First note that

\[
[a]_r = \begin{pmatrix} 1 & & \\ x & 1 & \\ x & x & \ddots & \\ x & x & \cdots & 1 \end{pmatrix}
\]
where $x$ denotes arbitrary entries, i.e. $\tilde{a}(z)$ is row reduced. The row degrees of $\tilde{a}(z)$ are exactly $n_i$ for each row $i = 1, \ldots, s$ and also coincide with the corresponding column degrees. The column end matrix has an even simpler structure: $[\tilde{a}]_c = I_s$, i.e. $\tilde{a}(z)$ is also column reduced.

(ii) As $\tilde{a}(z)$ is row (column) reduced, the degree of $\det(\tilde{a}(z))$ equals the sum of the row degrees $\sum_{i=1}^s n_i = n = |\tilde{a}|$. There is no MFD $\tilde{a}^{-1}(\tilde{d}, \tilde{b})$ of $(\tilde{l}, \tilde{k})$ where the determinant of $\tilde{a}(z)$ has smaller degree. Hence, $\tilde{a}^{-1}(\tilde{d}, \tilde{b})$ is irreducible.

(iii) Each row of $(\tilde{a}(z), \tilde{b}(z))$ must have degree strictly less than $n_i$ by construction.

We obtain the following

**Theorem 4.4.1 (Echelon ARMAX canonical form).** Let a transfer function $(l, k) \in \mathbb{M}(n)$ with Kronecker index $\alpha = (n_1, \ldots, n_s)$ be given and let us denote the whole set of transfer functions with this Kronecker index by $V_\alpha$. Then we obtain a unique ARMAX realization in the forward shift $z^{-1}$ of the form

$$
(l(z), k(z)) = \tilde{a}^{-1}(z^{-1})(\tilde{a}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1}))
$$

(4.31)

by (4.26), (4.29) and $D = l(0)$. The MFD $(\tilde{l}, \tilde{k}) = \tilde{a}^{-1}(\tilde{d}, \tilde{b})$ has the properties (i), (ii) and (iii) as discussed above.

Conversely, let a parameter vector $\tilde{\tau}_\alpha$ – consisting of the entries in the matrices $\tilde{A}(i), \tilde{D}(i), \tilde{B}(i)$ and $D$ that are not a priori restricted to zero or one – be given. Let $\tilde{T}_\alpha$ be the parameter space containing all vectors $\tilde{\tau}_\alpha$ such that the MFD $\tilde{a}^{-1}(\tilde{d}, \tilde{b})$ satisfies (i), (ii) and (iii). Then (4.31) is an echelon ARMAX realization in the forward shift $z^{-1}$ of a transfer function in $V_\alpha$.

**Proof.** For a detailed proof, see the proof of theorem 2.5.1 in (Hannan and Deistler, 1988).

Usually, one prefers to deal with ARMAX parametrizations in the backward shift $z$. The properties (i), (ii) and (iii) discussed above allow us to easily construct unique reversed echelon ARMAX parametrizations with similar properties:

**Theorem 4.4.2 (Reversed echelon ARMAX canonical form).** Let a transfer function $(l, k) \in \mathbb{M}(n)$ with Kronecker index $\alpha = (n_1, \ldots, n_s)$ be given and let us denote the whole set of transfer functions with this Kronecker index by $V_\alpha$. Then we obtain a unique ARMAX realization in the backward shift $z$ of the form

$$
(l(z), k(z)) = a^{-1}(z)(d(z), b(z))
$$

(4.32)

by (4.26), (4.29), $D = l(0)$ and (4.23). This MFD has the following properties:

(i) $A(0) = B(0)$ is lower triangular and all its diagonal entries are equal to one; $D(0) = A(0)D$.

(ii) $a^{-1}(d, b)$ is irreducible.

(iii) The degree of the $i$th row of $(a, d, b)$ is $n_i$.

Conversely, let a parameter vector $\tilde{\tau}_\alpha$ – consisting of the entries in the matrices $\tilde{A}(i), \tilde{D}(i), \tilde{B}(i)$ and $D$ that are not a priori restricted to zero or one – be given. Let $\tilde{T}_\alpha$ be the parameter space containing all vectors $\tilde{\tau}_\alpha$ such that the MFD $a^{-1}(d, b)$ satisfies (i), (ii) and (iii). Then (4.32) is a reversed echelon ARMAX realization in the backward shift $z$ of a transfer function in $V_\alpha$.

**Proof.** Property (i) follows from the fact that $A(0) = [\tilde{a}]$, which is immediately clear from (4.23). As $a^{-1}(d, b)$ is irreducible according to (ii) and $A(0)$ is nonsingular by (i), use of (4.25) shows that $a^{-1}(d, b)$ is also irreducible, i.e. (ii) holds true. Finally, by remark (1.5.5), irreducibility of $a^{-1}(d, b)$ implies that $(\tilde{a}(0), d(0), b(0))$ has rank $s$, and thus no row can be identically zero. It therefore follows from (4.23) that the degree of the $i$th row of $(a, d, b)$ must be $n_i$, i.e. (iii) holds true. The converse statement follows from theorem (4.4.1) above.

We conclude with a few remarks:
Remark 4.4.2. Note that the parameter vectors $\tilde{\tau}_\alpha$ and the parameter spaces $\tilde{T}_\alpha$ are exactly the same for echelon and reversed echelon ARMAX parametrizations. In the echelon ARMAX case, $\tilde{\tau}_\alpha$ is mapped to $(\tilde{a}, \tilde{d}, \tilde{b})$ by (4.26), (4.29) and $D = I(0)$, whereas in the reversed echelon ARMAX case $\tilde{\tau}_\alpha$ is mapped to $(a, d, b)$ by additionally applying transformation (4.23).

Remark 4.4.3. It is clear from the construction of the polynomial matrix $\tilde{a}(z)$ that the entries $\tilde{a}_{ij}(u)$ coincide with the corresponding echelon state-space parameters $-a_{ij}(k)$ for $k = u$. This is because in (4.3) the row $h(n_i + 1, i)$ is explained as a linear combination of the preceding basis rows whereas in (4.26) the same holds true for $-h(n_i + 1, i)$. Evidently, the parameters comprising $I(0) = D$ are identical. For the rest of the parameters it is not difficult to see that they are also linked homeomorphically: There is a homeomorphic (one-to-one) relation between the free echelon ARMAX parameters and the free echelon state-space parameters.

Again, for $|\alpha| = \sum_{i=1}^{s} n_i = n$, the number of free parameters $d_\alpha$ will be different for different $\alpha$. By

\[
\tilde{\phi}_\alpha : V_\alpha \rightarrow \tilde{T}_\alpha \\
(l, k) \rightarrow \tilde{\tau}_\alpha
\]

we denote the (reversed) echelon ARMAX parametrization attaching the free parameters $\tilde{\tau}_\alpha$ to the transfer functions $(l, k) \in V_\alpha$. Note that by construction the domain of definition of $\psi_\alpha$ (see (4.9)) and $\tilde{\phi}_\alpha$ is the same, i.e. echelon (and reversed echelon) ARMAX and echelon state-space parametrizations slice up the original model class $U_A$ in the same way.

An illustrative example

We give an illustrative example analogous to the one presented in section (4.2) above. For $m = 2$, $s = 2$, $n = 4$ and $\alpha = (2, 2)$, 28 free parameters appear in the echelon ARMAX parametrization (4.33) and the reversed echelon ARMAX parametrization (4.34),

\[
\begin{align*}
\tilde{a}(z) &= \begin{pmatrix} \tilde{a}_{11}(0) & \tilde{a}_{12}(0) \\ \tilde{a}_{21}(0) & \tilde{a}_{22}(0) \end{pmatrix} + \begin{pmatrix} \tilde{a}_{11}(1) & \tilde{a}_{12}(1) \\ \tilde{a}_{21}(1) & \tilde{a}_{22}(1) \end{pmatrix} z + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} z^2 \\
\tilde{d}(z) &= \begin{pmatrix} \tilde{d}_{11}(0) & \tilde{d}_{12}(0) \\ \tilde{d}_{21}(0) & \tilde{d}_{22}(0) \end{pmatrix} + \begin{pmatrix} \tilde{d}_{11}(1) & \tilde{d}_{12}(1) \\ \tilde{d}_{21}(1) & \tilde{d}_{22}(1) \end{pmatrix} z \\
\tilde{b}(z) &= \begin{pmatrix} \tilde{b}_{11}(0) & \tilde{b}_{12}(0) \\ \tilde{b}_{21}(0) & \tilde{b}_{22}(0) \end{pmatrix} + \begin{pmatrix} \tilde{b}_{11}(1) & \tilde{b}_{12}(1) \\ \tilde{b}_{21}(1) & \tilde{b}_{22}(1) \end{pmatrix} z \\
D &= \begin{pmatrix} \tilde{d}_{11} & \tilde{d}_{12} \\ \tilde{d}_{21} & \tilde{d}_{22} \end{pmatrix} \\
(l(z), k(z)) &= \tilde{a}^{-1}(z^{-1})(\tilde{d}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1})) \\
(l(z), k(z)) &= \begin{bmatrix} z^2 & 0 \\ 0 & z^2 \end{bmatrix} \tilde{a}(z^{-1})^{-1} \left[ \begin{bmatrix} z^2 & 0 \\ 0 & z^2 \end{bmatrix} \tilde{a}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1}) \right]
\end{align*}
\]

whereas for $\alpha = (4, 0)$, for instance, we get 25 free parameters in

\[
\begin{align*}
\tilde{a}(z) &= \begin{pmatrix} \tilde{a}_{11}(0) & 0 \\ \tilde{a}_{21}(0) & \tilde{a}_{22}(0) \end{pmatrix} + \begin{pmatrix} \tilde{a}_{11}(1) & 0 \\ \tilde{a}_{21}(1) & \tilde{a}_{22}(1) \end{pmatrix} z + \begin{pmatrix} \tilde{a}_{11}(2) & 0 \\ \tilde{a}_{21}(2) & \tilde{a}_{22}(2) \end{pmatrix} z^2 + \begin{pmatrix} \tilde{a}_{11}(3) & 0 \\ \tilde{a}_{21}(3) & \tilde{a}_{22}(3) \end{pmatrix} z^3 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} z^4 \\
\tilde{d}(z) &= \begin{pmatrix} \tilde{d}_{11}(0) & \tilde{d}_{12}(0) \\ \tilde{d}_{21}(0) & \tilde{d}_{22}(0) \end{pmatrix} + \begin{pmatrix} \tilde{d}_{11}(1) & \tilde{d}_{12}(1) \\ \tilde{d}_{21}(1) & \tilde{d}_{22}(1) \end{pmatrix} z + \begin{pmatrix} \tilde{d}_{11}(2) & \tilde{d}_{12}(2) \\ \tilde{d}_{21}(2) & \tilde{d}_{22}(2) \end{pmatrix} z^2 + \begin{pmatrix} \tilde{d}_{11}(3) & \tilde{d}_{12}(3) \\ \tilde{d}_{21}(3) & \tilde{d}_{22}(3) \end{pmatrix} z^3 \\
\tilde{b}(z) &= \begin{pmatrix} \tilde{b}_{11}(0) & \tilde{b}_{12}(0) \\ \tilde{b}_{21}(0) & \tilde{b}_{22}(0) \end{pmatrix} + \begin{pmatrix} \tilde{b}_{11}(1) & \tilde{b}_{12}(1) \\ \tilde{b}_{21}(1) & \tilde{b}_{22}(1) \end{pmatrix} z + \begin{pmatrix} \tilde{b}_{11}(2) & \tilde{b}_{12}(2) \\ \tilde{b}_{21}(2) & \tilde{b}_{22}(2) \end{pmatrix} z^2 + \begin{pmatrix} \tilde{b}_{11}(3) & \tilde{b}_{12}(3) \\ \tilde{b}_{21}(3) & \tilde{b}_{22}(3) \end{pmatrix} z^3 \\
D &= \begin{pmatrix} \tilde{d}_{11} & \tilde{d}_{12} \\ \tilde{d}_{21} & \tilde{d}_{22} \end{pmatrix} \\
(l(z), k(z)) &= \tilde{a}^{-1}(z^{-1})(\tilde{d}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1})) \\
(l(z), k(z)) &= \begin{bmatrix} z^4 & 0 \\ 0 & z^4 \end{bmatrix} \tilde{a}(z^{-1})^{-1} \left[ \begin{bmatrix} z^4 & 0 \\ 0 & z^4 \end{bmatrix} \tilde{a}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1}) \right]
\end{align*}
\]

Note that in the first case $A(0)$ will be the identity, whereas in the second case the entry $\tilde{a}_{21}(0)$ appears in the $(2, 1)$ position of $A(0)$. 
4.4.2 Topological and geometrical properties

It is clear from remark (4.4.3) above that theorem (4.2.2) is also applicable to the ARMAX case:

**Theorem 4.4.3.** The echelon and the reversed echelon ARMAX parametrizations

\[
\varphi_\alpha : V_\alpha \rightarrow \mathcal{T}_\alpha \\
(l, k) \mapsto \tau_\alpha
\]

where \( \alpha = (n_1, \ldots, n_s) \), have the properties (i) - (vii) described in theorem (4.2.2).

*Proof.* See the proof of theorem 2.5.3 in (Hannan and Deistler, 1988). \( \square \)

4.5 Overlapping ARMAX parametrizations in \( z^{-1} \) and \( z \)

4.5.1 Introduction

In deriving an overlapping ARMAX parametrization, we commence from a transfer function \((l, k) \in U_\alpha \)

where \( U_\alpha \) again denotes the set of transfer functions where \( \alpha \) is a structural index; see theorem (4.3.1). As we did not necessarily choose the first \( n \) rows that span the row space of the Hankel matrix \( \mathcal{H}_h \) all rows of \( \mathcal{H}_h^\infty \) have to be used to express the rows \( h(n_i + 1, i) \), \( i = 1, \ldots, s \), i.e. (4.26) becomes

\[
-h(n_i + 1, i) = \sum_{j=1}^{s} \sum_{u=1}^{n_j} \hat{a}_{ij}(u-1)h(u,j) \quad i = 1, \ldots, s \quad (4.35)
\]

This equation system defines unique coefficients \( \hat{a}_{ij}(u) \). Now, \( \hat{a}(z) \) is defined as in (4.27), and \((\hat{d}(z), \hat{b}(z))\) is obtained from (4.29) and (4.30), yielding an MFD \((\hat{l}(z), \hat{k}(z))\) and an ARMAX realization in the forward shift \( z^{-1} \) from (4.31) and \( l(0) = D \). Clearly, this construction is unique for a given structure index \( \alpha \), but in general the structure index \( \alpha \) is not unique for a given transfer function as has been discussed in section (4.3) already. We get the following

**Theorem 4.5.1 (Overlapping ARMAX parametrization in the forward shift \( z^{-1} \)).** Let a transfer function \((l, k) \in \mathbb{M}(n) \) be given, let \( \alpha = (n_1, \ldots, n_s) \) be a structural index for the Hankel matrix corresponding to \((l, k) \) and let us denote the whole set of transfer functions where \( \alpha \) is a structural index by \( U_\alpha \). Then we obtain an ARMAX realization in the forward shift \( z^{-1} \), unique with respect to \( \alpha \), of the form

\[
(l(z), k(z)) = \hat{a}(z^{-1})(\hat{d}(z^{-1}) + \hat{a}(z^{-1})D, \hat{b}(z^{-1}) + \hat{a}(z^{-1}))
\]

by (4.35), (4.29) and \( D = l(0) \). The MFD \((\hat{l}(z), \hat{k}(z)) = \hat{a}(z^{-1})(\hat{d}(z), \hat{b}(z))\) has the following properties:

(i) \( [\hat{a}] = I_s \) and \( \nu(\hat{a}_u) = n_i \), implying that the polynomials \( \hat{a}_u(z) \) are monic and \( \hat{a}(z) \) is column reduced.

(ii) \( \hat{a}(z^{-1})(\hat{d}, \hat{b}) \) is irreducible.

(iii) Each row of \((\hat{d}(z), \hat{b}(z))\) has degree strictly less than the degree of the corresponding row of \( \hat{a}(z) \). However, the latter degree need not be \( n_i \).

Conversely, let a parameter vector \( \tau^{(1)}_\alpha \) be given where \( \tau^{(1)}_\alpha \) consists of the elements

\[
\hat{a}_{ij}(u), \quad u = 0, 1, \ldots, n_j - 1; \quad j, i = 1, \ldots, s \quad (4.37)
\]

\[
\hat{d}_{ij}(u), \quad u = 0, 1, \ldots, n_i - 1; \quad i = 1, \ldots, s, j = 1, \ldots, m \quad (4.38)
\]

\[
\hat{b}_{ij}(u), \quad u = 0, 1, \ldots, n_i - 1; \quad i, j = 1, \ldots, s \quad (4.39)
\]

\[
\hat{d}_{ij}, \quad i = 1, \ldots, s, j = 1, \ldots, m \quad (4.40)
\]
where $\hat{a}_{ij}$ are the entries of $D$. Let $\hat{T}_{\alpha}^{(1)}(z)$ be the parameter space containing all vectors $\hat{\tau}_{\alpha}^{(1)}$ such that the MFD $\hat{a}_{ij}^{-1}(\hat{d}, \hat{b})$ satisfies (i), (ii) and (iii) above where $(\hat{a}, \hat{d}, \hat{b})$ is obtained by inserting the parameters in (4.37), (4.38) and (4.39) and the zero and one entries resulting from (i), (ii) and (iii) in the correct position. The remaining entries are given by (nonlinear) functions of the free parameters in $\hat{\tau}_{\alpha}^{(1)}$ which are not specified here. Then (4.36) is an ARMAX realization in the forward shift $z^{-1}$ of a transfer function in $U_\alpha$.

Proof. For a detailed proof, see the proofs of theorems 2.6.1 and 2.6.4 in (Hannan and Deistler, 1988).

A few remarks are in place:

Remark 4.5.1. Note that the column degrees of $\hat{a}$ are $n_i$, but the row degrees of $\hat{a}$ may be larger. This is clear from the construction of $\hat{a}$. The matrix $\hat{a}$ need not be row reduced.

Remark 4.5.2. Despite the fact that $\hat{a}$ need not be be row reduced, (iii) is clear from (4.29). Hence, (4.23) indeed yields polynomial matrices $(a, d, b)$ if $n_i$ in (4.23) is replaced by the "true" row degrees of $\hat{a}$.

Remark 4.5.3. It is clear from the converse statement in the theorem above that not all entries in the matrices $A(i), D(i)$ and $B(i)$ corresponding to $(\hat{a}, \hat{d}, \hat{b})$ which are not explicitly restricted by (i), (ii) and (iii) are free in general, as there are nonlinear restrictions between them. For an illustrative example, see (4.44) below. The fact that the vector $\hat{\tau}_{\alpha}^{(1)}$ as defined above is indeed a vector of free parameters can be shown by establishing a homeomorphism between the parameter vector $\tau_{\alpha}^{(1)}$ for the overlapping state-space parametrization and $\hat{\tau}_{\alpha}^{(1)}$ for the overlapping ARMAX parametrization.

We might again be interested primarily in ARMAX realizations in the backward shift $z$:

Theorem 4.5.2 (Overlapping ARMAX parametrization in the backward shift $z$). Let a transfer function $(l, k) \in M(n)$ be given, let $\alpha = (n_1, \ldots, n_s)$ be a structural index for the Hankel matrix corresponding to $(l, k)$ and let us denote the whole set of transfer functions where $\alpha$ is a structural index by $U_\alpha$. Then we obtain an ARMAX realization in the backward shift $z$, unique with respect to $\alpha$, of the form

$$
(l(z), k(z)) = a^{-1}(z) (d(z), b(z))
$$

by (4.35), (4.29) and $D = l(0)$ and (4.23) where the integers $n_i$ in (4.23) refer to the true row degrees of $\hat{a}$ which may be larger than the integers $n_i$ appearing in $\alpha$.

Conversely, let a parameter vector $\hat{\tau}_{\alpha}^{(1)}$ and the parameter space $\hat{T}_{\alpha}^{(1)}$ be given as in theorem (4.5.1). Then (4.41) is an ARMAX realization in the backward shift $z$ of a transfer function in $U_\alpha$.

Proof. This theorem follows immediately from theorem (4.5.1) and remark (4.5.2).

Again, we give a few remarks:

Remark 4.5.4. Note that the properties (i), (ii) and (iii) of reversed echelon ARMAX realizations of transfer functions in $V_\alpha$ can no longer be guaranteed: As $\hat{a}$ need not necessarily be row reduced any more, $A(0)$ (corresponding to $a(z)$) may be singular and $a^{-1}(d, b)$ may not be irreducible as a direct consequence of (4.25). Moreover, the degree of the ith row of $(a, d, b)$ equals the true degree of the ith row of $\hat{a}$ which may be larger than the corresponding column degree $n_i$; see remark (4.5.1).

Remark 4.5.5. Of course, the fact that some entries in $(\hat{a}, \hat{d}, \hat{b})$ are nonlinear functions of the entries in $\hat{\tau}_{\alpha}^{(1)}$ is a shortcome of overlapping ARMAX parametrizations as compared to overlapping state-space parametrizations discussed in (4.3) above.

As can be seen from theorems (4.5.1) and (4.5.2), the number of free parameters for $|\alpha| = \sum_{i=1}^{s} n_i = n$ is the same for different $\alpha$. By

$$
\hat{\phi}_{\alpha}^{(1)} : U_\alpha \rightarrow \hat{T}_{\alpha}^{(1)}
$$
we denote the overlapping ARMAX parametrizations (in the forward shift \( z^{-1} \) and the backward shift \( z \), respectively) attaching the free parameters \( \tilde{\psi}_{\alpha}^{(1)} \) to the transfer functions \((l, k) \in U_\alpha\). Clearly, the domain of definition of \( \psi_{\alpha}^{(1)} \) is the same as for the state-space case discussed in section (4.3).

**An illustrative example**

We give an illustrative example analogous to the one presented in section (4.3) above. For \( m = 2, s = 2, n = 4 \) and \( \alpha = (2, 2) \), 28 free parameters appear in the overlapping ARMAX realizations (4.42) and (4.43). These realizations evidently coincide with the echelon ARMAX and reversed echelon ARMAX realizations in (4.33) and (4.34):

\[
\begin{align*}
\tilde{a}(z) &= \left( \begin{array}{cc} \tilde{a}_{11}(0) & 0 \\ \tilde{a}_{21}(0) & \tilde{a}_{22}(0) \end{array} \right) + \left( \begin{array}{cc} \tilde{a}_{11}(1) & \tilde{a}_{12}(1) \\ \tilde{a}_{21}(1) & \tilde{a}_{22}(1) \end{array} \right) z + \left( \begin{array}{c} 1 \\ 0 \\ 0 \\ 1 \end{array} \right) z^2, \\
\tilde{d}(z) &= \left( \begin{array}{cc} \tilde{d}_{11}(0) & \tilde{d}_{12}(0) \\ \tilde{d}_{21}(0) & \tilde{d}_{22}(0) \end{array} \right) + \left( \begin{array}{cc} \tilde{d}_{11}(1) & \tilde{d}_{12}(1) \\ \tilde{d}_{21}(1) & \tilde{d}_{22}(1) \end{array} \right) z, \\
\tilde{b}(z) &= \left( \begin{array}{cc} \tilde{b}_{11}(0) & \tilde{b}_{12}(0) \\ \tilde{b}_{21}(0) & \tilde{b}_{22}(0) \end{array} \right) + \left( \begin{array}{cc} \tilde{b}_{11}(1) & \tilde{b}_{12}(1) \\ \tilde{b}_{21}(1) & \tilde{b}_{22}(1) \end{array} \right) z, \\
D &= \left( \begin{array}{cc} \tilde{d}_{11} & \tilde{d}_{12} \\ \tilde{d}_{21} & \tilde{d}_{22} \end{array} \right)
\end{align*}
\]

\[
\begin{align*}
(l(z), k(z)) &= \tilde{a}^{-1}(z^{-1})(\tilde{d}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1})) \\
(l(z), k(z)) &= \left[ \begin{array}{c} z^2 \\ 0 \\ 0 \\ z^2 \\ z^2 \end{array} \right]^{-1} \left[ \begin{array}{c} z^3 \\ 0 \\ 0 \\ z^3 \\ z^3 \end{array} \right] \left( \tilde{d}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1}) \right) \right] \quad (4.42)
\end{align*}
\]

However, for \( \alpha = (4, 0) \), we also obtain 28 free parameters in

\[
\begin{align*}
\tilde{a}(z) &= \left( \begin{array}{cc} \tilde{a}_{11}(0) & 0 \\ \tilde{a}_{21}(0) & \tilde{a}_{22}(0) \end{array} \right) + \left( \begin{array}{cc} \tilde{a}_{11}(1) & \tilde{a}_{12}(1) \\ \tilde{a}_{21}(1) & \tilde{a}_{22}(1) \end{array} \right) z + \left( \begin{array}{c} 1 \\ 0 \\ 0 \\ 0 \end{array} \right) z^4, \\
\tilde{d}(z) &= \left( \begin{array}{cc} \tilde{d}_{11}(0) & \tilde{d}_{12}(0) \\ \tilde{d}_{21}(0) & \tilde{d}_{22}(0) \end{array} \right) + \left( \begin{array}{cc} \tilde{d}_{11}(1) & \tilde{d}_{12}(1) \\ \tilde{d}_{21}(1) & \tilde{d}_{22}(1) \end{array} \right) z + \left( \begin{array}{c} 0 \\ 1 \\ 0 \\ 0 \end{array} \right) z^3, \\
\tilde{b}(z) &= \left( \begin{array}{cc} \tilde{b}_{11}(0) & \tilde{b}_{12}(0) \\ \tilde{b}_{21}(0) & \tilde{b}_{22}(0) \end{array} \right) + \left( \begin{array}{cc} \tilde{b}_{11}(1) & \tilde{b}_{12}(1) \\ \tilde{b}_{21}(1) & \tilde{b}_{22}(1) \end{array} \right) z + \left( \begin{array}{c} 0 \\ 0 \\ 1 \\ 0 \end{array} \right) z^2, \\
D &= \left( \begin{array}{cc} \tilde{d}_{11} & \tilde{d}_{12} \\ \tilde{d}_{21} & \tilde{d}_{22} \end{array} \right)
\end{align*}
\]

\[
\begin{align*}
(l(z), k(z)) &= \tilde{a}^{-1}(z^{-1})(\tilde{d}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1})) \\
(l(z), k(z)) &= \left[ \begin{array}{c} z^4 \\ 0 \\ 0 \\ 0 \end{array} \right]^{-1} \left[ \begin{array}{c} z^5 \\ 0 \\ 0 \\ 0 \end{array} \right] \left( \tilde{d}(z^{-1}) + \tilde{a}(z^{-1})D, \tilde{b}(z^{-1}) + \tilde{a}(z^{-1}) \right) \right] \quad (4.44)
\end{align*}
\]

where \( \chi \) denotes some (varying) nonlinear function of the free parameter. Note that in general \( \tilde{a} \) is not row reduced, the degree of the second row of \( \tilde{a} \) is three (in general) and thus strictly greater than the second column degree. Hence, \( A(0) \) in (4.44) is singular and (4.44) is not irreducible.

### 4.5.2 Topological and geometrical properties

It is clear from remark (4.5.3) above that theorem (4.3.2) is also applicable to the ARMAX case:

**Theorem 4.5.3.** The overlapping ARMAX parametrizations in the forward shift \( z^{-1} \) and the backward shift \( z \)

\[
\bar{\psi}_{\alpha}^{(1)} : U_\alpha \rightarrow \tilde{T}_{\alpha}^{(1)}
\]

\[
(l, k) \mapsto \bar{T}_{\alpha}^{(1)}
\]

where \( \alpha = (n_1, \ldots, n_s) \), have the properties (i) - (vii) described in theorem (4.3.2).

**Proof.** See the proofs of theorems 2.6.3 and 2.6.5 in (Hannan and Deistler, 1988).

Of course, remarks completely analogous to those made below theorem (4.3.2) can also be made at this point.
4.6 Lyapunov balanced canonical form (Ober)

Balanced canonical forms have originally been introduced for different classes of continuous time state-space representations. For a brief introduction of continuous time state-space systems see the appendix (A.1). The aim of this section and the following two sections is to introduce continuous time balanced canonical state-space forms and to then show how according balanced canonical forms can be obtained for the discrete time case.

4.6.1 Introduction

Lyapunov balancing for continuous time state-space realizations is introduced in the appendix (A.3). There, only square transfer functions are considered for ease of a unified presentation. As we are now aiming at a parametrization of the pair \((l, k)\), we will consider \((l_c, k_c)\) to be one transfer function:

\[
(l_c(s), k_c(s)) = \left[ \begin{array}{c|c} A_c & \tilde{B}_c \\ \hline C_c & D_c \end{array} \right] \quad \text{where} \quad \tilde{B}_c = [B_c, K_c] \quad \text{and} \quad \tilde{D}_c = [D_c, E_c] \quad (4.45)
\]

In the appendices (A.3) and (A.7), the symbol \(\mathcal{M}_c(n)\) was used to denote the set of all rational, proper and stable \(s \times s\) continuous time transfer functions of order \(n\) where\(^2\) \(k_c(\infty) = D\) for an arbitrary \(D \in \mathbb{R}^{s \times s}\). 

In this section we will, with slight abuse of notation, use the same symbol for the non square case, i.e. for the set of all rational, proper and stable \(s \times (m + s)\) continuous time transfer functions of order \(n\) where \((l_c(\infty), k_c(\infty)) = \tilde{D}_c\) as in (4.45).

We now give a short sketch of the construction of Ober’s Lyapunov balanced canonical form; see, e.g., (Maciejewski and Ober, 1988). For details we refer to (Ober, 1991) as well as to (Ober, 1987) and (Ober, 1996). Note, however, that the canonical forms in the latter two papers differ in (some) details from the one we introduce below.

Assume we are given a transfer function \((l_c, k_c) \in \mathcal{M}_c(n)\). Let the structure of its Hankel singular values be described by a unique structural index,

\[
\alpha = (n_1, \ldots, n_k) \quad (4.46)
\]

say, specifying the multiplicities of the Hankel singular values. Clearly, \(|\alpha| = \sum_{i=1}^{k} n_i = n\) must hold true. As the class of balanced state-space realizations is determined up to state transformations of the form (A.28), a canonical form can only be obtained if additional restrictions are imposed. First, it is shown in (Ober, 1991) that one can always apply a transformation \(T_{bal}\) of the form (A.28) that makes the matrices \(\tilde{B}_i\) nonnegative upper triangular. Here, \(\tilde{B}_i\) is a \(n_i \times (m + s)\) submatrix of \(\tilde{B}_c\), i.e.

\[
\tilde{B}_c = \begin{pmatrix}
\tilde{B}_1 \\
\vdots \\
\tilde{B}_k
\end{pmatrix}
\]

and nonnegative upper triangular has the following meaning:

\[
\tilde{B}_i = \begin{pmatrix}
0 & \ldots & 0 & \tilde{B}_{1,t_i,1} & x & \ldots & x & x & \ldots & x \\
0 & \ldots & 0 & 0 & 0 & \ldots & 0 & \tilde{B}_{2,t_i,2} & \ldots & x \\
\vdots & & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & \tilde{B}_{r_i,t_i,r_i} & x
\end{pmatrix} \in \mathbb{R}^{n_i \times (m + s)}
\]

\[
\tilde{B}_{i,t_i,l} > 0 \quad \text{for} \quad l = 1, \ldots, r_i \quad (4.47)
\]

\(^2\)Note that \(\lim_{t \to \infty} (l_c(s), k_c(s))\) will be abbreviated by \((l_c(\infty), k_c(\infty))\) in the sequel.
and $x$ denotes arbitrary entries. Such a transformation fixes each orthogonal matrix $T_i$ in $T_{01t}$ in the first $r_i$ columns. Note that the indices

$$r_1, \ldots, r_k \text{ with } 1 \leq r_i \leq \min(n_i, s) \quad \text{rank of the matrices } B_i$$

$$1 \leq t_{i,1} < t_{i,2} < \cdots < t_{i,r_i} \leq (m + s) \quad \text{column position of first nonzero entry in } 1, \ldots, r_i, \text{ row of } B_i$$

are uniquely determined for a given transfer function, i.e. they are invariants within the class of observationally equivalent minimal systems. In order to obtain a canonical form, the orthogonal transformations $T_i$ have to be fixed in the remaining $(n_i - r_i)$ columns, too.

Up to now we have transformed our original state-space realization such that it is balanced and $B_c$ consists of nonnegative upper triangular submatrices $B_i$. From the two Lyapunov equations (A.24) and (A.25) it is apparent that each subsystem $(A_{ii}, B_i, C_i, D_c)$ is also balanced with Gramians $\sigma_i I_{n_i}$. Here, $A_{ii}$ denotes the $i$. diagonal block of $A_c$, $B_i$ is the $i$. nonnegative upper triangular submatrix of $B_c$, $C_i$ is the $i$. block column of $C_c$ and the partitioning is according to the multiplicities of the Hankel singular values given by $\alpha$:

$$\sigma_i A_{ii} I_{n_i} + \sigma_i I_{n_i} A_{ii}' + B_i B_i' = 0$$

$$\sigma_i A_{ii}' I_{n_i} + \sigma_i I_{n_i} A_{ii} + C_i' C_i = 0$$

It follows directly from the first equation that the symmetric part $A_{ii}'$ of

$$A_{ii} = \frac{1}{2} (A_{ii} + A_{ii}') + \frac{1}{2} (A_{ii} - A_{ii}')$$

is uniquely determined from the entries in $B_i$ that are not restricted to zero:

$$A_{ii}' = -\frac{1}{2\sigma_i} \tilde{B}_i B_i' = -\frac{1}{2\sigma_i} \begin{pmatrix} \tilde{B}_i B_i' & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n_i \times n_i}$$

where $\tilde{B}_i = \begin{pmatrix} B_i & \tilde{B}_i' \\ \tilde{B}_i & B_i' \end{pmatrix}$ denotes the matrix consisting of the first $r_i$ rows of $B_i$ in (4.47). Moreover, taking both equations, we see that $\tilde{B}_i B_i' = C_i' C_i$, i.e. $\tilde{B}_i$ and $C_i'$ are square roots of the same positive semidefinite matrix (therefore, the restriction $r_i \leq \min(n_i, s)$ was imposed in (4.48) above). Thus, we have

$$C_i = \begin{pmatrix} U_i \Delta_i, 0_{s \times (n_i - r_i)} \end{pmatrix} \in \mathbb{R}^{s \times n_i}, \quad \Delta_i = (\tilde{B}_i B_i')^{1/2}$$

where the columns of $U_i \in \mathbb{R}^{s \times r_i}$ are a set of orthonormal vectors, i.e. $U_i^t U_i = I_{r_i}$ and $\Delta_i$ is the uniquely defined symmetric positive definite square root of $(\tilde{B}_i B_i')$.

As already mentioned, if $(n_i - r_i) > 0$, we still have to fix the orthogonal transformations $T_i$ in the last $(n_i - r_i)$ columns. This can be done by bringing the skew symmetric part $A_{ii}'$ into $r$-balanced form. For a given skew symmetric matrix, there is a unique orthogonal transformation of the form $\text{diag}(I_{r_i}, Q_i)$, $Q_i \in O(n_i - r_i)$ making $A_{ii}'$ r-balanced, where $r$-balanced has the following meaning:

$$A_{ii}' = \begin{pmatrix} A_{ii}'_{11} & A_{ii}'_{12} \\ A_{ii}'_{21} & A_{ii}'_{22} \end{pmatrix} \in \mathbb{R}^{n_i \times n_i}$$

where

(i) $A_{ii}'_{11} \in \mathbb{R}^{r_i \times r_i}$ is skew symmetric. Hence, $A_{ii}'_{11}$ contains only zeros in its diagonal. Furthermore, note that in case $r_i = n_i \geq 1$ the matrices $A_{ii}'_{11}, A_{ii}'_{21}$ and $A_{ii}'_{22}$ do not appear and $A_{ii}' = A_{ii}'_{11}$. If $r_i = n_i = 1$, then $A_{ii}'_{2k} = 0$: scalars do not have a nontrivial skew symmetric component.
(ii) $A_{ii}^{21} = -A_{ii}^{22}$ and $A_{ii} \in \mathbb{R}^{(n_i-r_i) \times (n_i-r_i)}$ are determined by a unique set of indices

$$
q_1, \ldots, q_i \text{ with } q_i \leq (n_i - r_i) \quad \text{number of jumps in } A_{ii}^{12}
$$

$$
1 = h_{i,1} < h_{i,2} < \cdots < h_{i,q_i} \leq (n_i - r_i) \quad \text{column indices for jumps in } A_{ii}^{12}
$$

$$
1 \leq g_{i,q_i} < g_{i,q_i-1} < \cdots < g_{i,1} \leq r_i \quad \text{row indices for jumps in } A_{ii}^{12}
$$

(4.49)

and have the following form

$$
A_{ii}^{12} = \begin{pmatrix}
1 & \cdots & 1 & \cdots & 1 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
x_{1,1} & \cdots & x_{1,q_i} & 0 & \cdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
x_{1,1} & \cdots & x_{q_i,1} & a_{g_{1,1},h_{i,1}} & x_{q_i,q_i} \\
0 & \cdots & 0 & a_{g_{2,1},h_{i,2}} & 0 \\
\vdots & \ddots & \ddots & \cdots & \vdots \\
x_{q_i,1} & \cdots & x_{q_i,q_i} & a_{g_{q_i,1},h_{i,q_i}} & 0
\end{pmatrix} \in \mathbb{R}^{n_i \times (n_i-r_i)}
$$

$$
A_{ii}^{22} = \begin{pmatrix}
0 & \alpha_{i,2} & \cdots & \cdots & \cdots \\
-\alpha_{i,2} & 0 & \alpha_{i,3} & \cdots & \cdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & 0 & \alpha_{i,n_i-r_i}
\end{pmatrix} \in \mathbb{R}^{(n_i-r_i) \times (n_i-r_i)}.
$$

$$
\begin{align*}
\alpha_{i,l} & = 0 & \text{if } l = h_{i,j} \text{ for some } 1 \leq j \leq q_i, \text{ i.e. } l \text{ is a column index for a jump in } A_{ii}^{12} \\
\alpha_{i,l} & > 0 & \text{otherwise} \\
a_{g_{i,1},h_{i,l}} & > 0 & \text{for } l = 1, \ldots, q_i
\end{align*}
$$

We will gather all structural information for $\tilde{B}_r$ in (4.48) and all structural information for $A_r$, i.e. the information for the skew symmetric r-balanced parts $A_{ii}^{2k}$ in (4.49), in one structural index which we shall denote by $\beta$.

Finally, the block off diagonal elements $A_{ij}, i \neq j$ of $A_r$ are uniquely determined by the Lyapunov equations (A.24) and (A.25) because the two (matrix) equations corresponding to block $(i,j)$ are given by

$$
\begin{align*}
\sigma_j A_{ij} + \sigma_i A_{ji}^{t} & = -\tilde{B}_i \tilde{B}_j^t \\
\sigma_j A_{ji}^{t} + \sigma_i A_{ij} & = -C_i^t C_j
\end{align*}
$$

and adding $(-\sigma_j)$ times the first and $\sigma_i$ times the second equation yields the block off diagonal element $A_{ij}$:

$$
A_{ij} = \frac{\sigma_j \tilde{B}_i \tilde{B}_j^t - \sigma_i C_i^t C_j}{\sigma_i^2 - \sigma_j^2} = \frac{1}{\sigma_i^2 - \sigma_j^2} \begin{pmatrix}
\sigma_j \tilde{B}_i \tilde{B}_j^t - \sigma_i \Delta_i U_i U_j^t \Delta_j & 0 \\
0 & 0
\end{pmatrix} \in \mathbb{R}^{n_i \times n_j}
$$

To sum it up, we have now constructed a unique balanced continuous time state-space representation for a given transfer function $(l, k) \in M_0^r(n)$. In order to get a unique discrete time state-space representation for a given transfer function $(l, k) \in M_0(n)$ we have to make use of the homeomorphism $\rho$ given in (A.44) in the appendix (A.4). Due to the properties of (A.44), the discrete time state-space representation will also be Lyapunov balanced; see definition (3.4.1) in section (3.4) for a definition of discrete time Lyapunov balancing.
The only remaining problem is the (homeomorphic) parametrization of the matrices \( U_i \in \mathbb{R}^{q \times r_i} \). It is easy to see that each \( U_i \) cannot be continuously parametrized using one coordinate mapping. This is the reason for introducing another structural index, \( \gamma \) say, specifying a particular set out of a disjoint partition of the set of all possible \( U_i, i = 1, \ldots, k \). For this subset, the \( U_i, i = 1, \ldots, k \), can then be parametrized continuously. This will be discussed in more detail below for special values of \( s \) and \( r_i \). We will collect all structural information in one multi index \( \delta = (\alpha, \beta, \gamma) \). For given \( n \), specification of this multi index \( \delta \) will give a decomposition of \( M_\delta(n) \) into disjoint pieces\(^3\).

Referring to (A.42) for the bilinear transformation \( \rho_l \) between continuous time and discrete time transfer functions, we are now ready to state the following theorem. Note that by \( \tau > 0 \) we mean that each component of the vector \( \tau \) to be positive:

**Theorem 4.6.1 (Ober’s Lyapunov balanced canonical form).** Let a transfer function \((l, k) \in M_\delta(n)\) be given and let the corresponding \((l_\delta(s), k_\delta(s)) = (l(\rho_l^{-1}(s)), k(\rho_l^{-1}(s))) \in M_\delta(n)\) have the structural index \( \delta = (\alpha, \beta, \gamma) \) where \( \alpha \) is given in (4.46), \( \beta \) is given by (4.48) and (4.49) and \( \gamma \) specifies some subset of the set of all possible \( U_i \in \mathbb{R}^{q \times r_i}, i = 1, \ldots, k \) in (4.50) below. Moreover, let us denote the whole set of stable transfer functions with this multi index by \( V_\delta \). Then there exists a unique minimal, stable and (discrete time) Lyapunov balanced state-space representation with matrices

\[
A = (I + A_c)(I - A_c)^{-1}
\]

\[
(B, K) = \sqrt{2}(I - A_c)^{-1} B_c
\]

\[
C = \sqrt{2}C_c(I - A_c)^{-1}
\]

\[
(D, I) = D_c + C_c(I - A_c)^{-1} B_c
\]

where \((A_c, B_c, C_c, D_c)\) have the following form:

\[
\begin{align*}
\hat{B}_c &= [B_c, K_c] = \begin{pmatrix} \hat{B}_1 \\ \vdots \\ \hat{B}_k \end{pmatrix} \quad \text{where } \hat{B}_i \in \mathbb{R}^{n_i \times (m + s)} \text{ are nonnegative upper triangular as given in (4.47)}
\end{align*}
\]

\[
C_c = \begin{pmatrix} C_1 & \cdots & C_k \end{pmatrix} \quad \text{where } C_i = [U_i \Delta_i, 0_{s \times (n_i - r_i)}] \text{ with } \Delta_i = (\hat{B}_i^r \hat{B}_i^r)^{1/2} \text{ and } U_i^r U_i = I_{r_i}
\]

\[
A_c = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kk} \end{pmatrix} \quad \text{where } A_{ij} \in \mathbb{R}^{n_i \times n_j} \text{ have the following form:}
\]

\[
A_{ij} = \frac{1}{\sigma_i^2 - \sigma_j^2} \begin{pmatrix} \sigma_j \hat{B}_i^r \hat{B}_j^r - \sigma_i \Delta_i U_i^r U_j \Delta_j & 0 \\ 0 & \Delta_i^2 \end{pmatrix}_{(n_i-r_i) \times (n_j-r_j)}
\]

\[
\Delta_i^2 = \begin{pmatrix} A_{ii}^1 & A_{ii}^2 \\ A_{ii}^1 & A_{ii}^3 \end{pmatrix}
\]

\[
\hat{D}_c = [D_c, E_c]
\]

\[
D_c = \begin{pmatrix} d_{c,11} & \cdots & d_{c,1m} \\ \vdots & \ddots & \vdots \\ d_{c,s1} & \cdots & d_{c,sm} \end{pmatrix}, \quad E_c = I - C_c(I - A_c)^{-1} K_c
\]

where \( A_{ii}^1 \) is in \( r \)-balanced form and thus \( A_{ii}^1, A_{ii}^2, A_{ii}^3 \) and \( A_{ii}^2 \) are given as in (i) and (ii) above.

Conversely, let a parameter vector \( \tau_\delta = (\tau_{\Sigma}, \tau_{B^+}, \tau_{B^r}, \tau_{A^+}, \tau_{A^r}, \tau_{\Phi}, \tau_D) \) be given,

\(^3\)There are also ways of defining *overlapping parametrizations* for \( \mathbb{M}_\delta(n) \) making use of the idea of balancing; see (Hanzon and Ober, 1997) for the stable SISO case, for instance.
\[
\tau_\Sigma = (\sigma_1 - \sigma_2, \sigma_2 - \sigma_3, \ldots, \sigma_{k-1} - \sigma_k, \sigma_k) > 0
\]
\[
\tau_{B^+} = (\bar{B}_{i_1,i_1}, \ldots, \bar{B}_{i_k,i_k}, \bar{B}_{i_1,i_k}, \ldots, \bar{B}_{i_k,i_k}) > 0
\]
\[
\tau_{B^r} = \text{unrestricted entries } x \text{ in the matrices } \bar{B}_i \text{ in (4.47) for } i = 1, \ldots, k
\]
\[
\tau_{A^+} = (a_{i_1,i_1}, a_{i_2,i_2}, \ldots, a_{i_k,i_k}, \ldots, a_{g_k,a_k}, a_{g_k,a_k}, \alpha_{1,2}, \ldots, \alpha_{1,m-r_k}, \ldots, \alpha_{k,2}, \ldots, \alpha_{k,n_k} > 0
\]
where only nonzero \( \alpha_{ij} \) are included
\[
\tau_{A^r} = \text{unrestricted entries in the strictly lower triangular parts of the skew symmetric matrices } A_{ik}^{11}
\]
\[
\tau_{D^r} = \text{free parameters for the matrices } U_i \in \mathbb{R}^{s \times n} \text{ for } i = 1, \ldots, k
\]

(4.51)

where \( \tau_\delta \) consists of the differences of the distinct Hankel singular values \( \tau_\Sigma \), the vector \( \tau_{B^+} \) of positive entries for \( \bar{B}_c \), the vector \( \tau_{B^r} \) of the remaining (unrestricted) entries for \( \bar{B}_c \), the vector \( \tau_{A^+} \) of positive entries for the skew symmetric diagonal blocks \( \bar{A}^{11}_i \), the vector \( \tau_{A^r} \) of the remaining (unrestricted) entries for the same blocks \( \bar{A}^{11}_i \) and the vector \( \tau_D \) containing the free parameters for the \( k \) matrices \( U_i \in \mathbb{R}^{s \times n} \) which appear in the construction of the \( k \) block columns of \( C_c \). Let \( T_\delta \) be the parameter space containing all vectors \( \tau_\delta \) satisfying the positivity restrictions above and let

\[
\varphi_\delta : T_\delta \to S_m(n)
\]
\[
\tau_\delta \to (A(\tau_\delta), B(\tau_\delta), C(\tau_\delta), D(\tau_\delta), K(\tau_\delta))
\]

be the mapping attaching the discrete time system matrices to the parameters as given in the first part of this theorem. Then \( \varphi_\delta(\tau_\delta) \in T_\delta \) is a discrete time, minimal, stable and Lyapunov balanced state-space realization of a transfer function in \( V_\delta \).

**Proof.** The fact that every transfer function in \( V_\delta \) has a state-space realization of the given form has been discussed above. For a more detailed proof of this statement, see the proof of theorem 2.1 in (Ober, 1991). The proof of the converse statement follows along the same lines as the proof of theorem 7.1 in (Ober, 1987). Note that we chose \( E_c \) in such a way that it does not contain any extra parameters and that \( k(0) = I \). This is straightforward to see from (A.44).

\( \Box \)

**Remark 4.6.1.** In theorem 2.1 in (Ober, 1991) \( A_{ik}^{12} \) is given in a slightly different way: \( A_{ik}^{11} \) in our theorem above differs from the corresponding \( A_{ik}^{11} \) in (Ober, 1991) by a constant skew symmetric matrix, which is not relevant as \( A_{ik}^{11} \) does not contain any restricted entries (apart from the restrictions of skew symmetric, of course).

**Remark 4.6.2.** Theorems 6.1 and 7.1 in (Ober, 1987) yield a slightly different canonical form.

**Remark 4.6.3.** Note that in the case \( r_i = 1 \) for \( i = 1, \ldots, k \), we only have to parametrize unit vectors in \( \mathbb{R}^m \). In particular, this will be the case in the following

**Remark 4.6.4 (The case of distinct second order modes, i.e. \( \alpha = (1, \ldots, 1) \)).** If all \( n_i \) are equal to one, i.e. in the case of distinct Hankel singular values where \( \alpha = (1, \ldots, 1) \), the matrices \( \bar{B}_i \) become row vectors. Note that each row vector \( \bar{B}_i \neq 0 \); if one row of \( \bar{B}_c \) were zero, we would obtain a contradiction to minimality by considering the Lyapunov equations (A.24) and (A.25). Hence, \( n_i = 1 \) implies \( r_i = 1 \) for \( i = 1, \ldots, n \). The \( n \) unit vectors \( U_i \in \mathbb{R}^m \) can be parametrized homeomorphically using the following parameter (sub-) vector \( \tau_\phi \):
Here, \( \phi_{i,1}, \ldots, \phi_{i,s-1} \) are the polar coordinates needed to describe the unit vector \( U_i \in \mathbb{R}^s \); see also (Chou and Maciejowski, 1997). Note that, of course, the unit sphere \( S_{s-1} \) cannot be described continuously by one coordinate chart: The unit vectors of the form \( (0, U_{i,2}, \ldots, U_{i,s})' \) with \( U_{i,2} \leq 0 \) cannot be described by (4.52). However, using a structural index \( \gamma \) which specifies the parametrization (4.52), only a thin set out of the set of all possible \( U_i, i = 1, \ldots, n \) is neglected.

**Remark 4.6.5 (The generic piece: \( \delta_g = ((1, \ldots, 1), (1, \ldots, 1), \gamma_g) \) and \( s > 1 \).** In the case of distinct Hankel singular values, i.e. \( \alpha = (1, \ldots, 1) \), the situation simplifies considerably. As already mentioned in remark (4.6.4) above, the rows \( \tilde{B}_i \) must satisfy \( \tilde{B}_i \neq 0 \). Consequently, \( r_i = 1 \) for all \( i = 1, \ldots, n \) and all structural information in (4.48) can be summarized in a vector of length \( n \) containing the positions \( t_{i,1} \) of the first positive entries in each row \( \tilde{B}_i \) for \( i = 1, \ldots, n \). Furthermore, we do not need the structural information for \( A^{sk}_{i} \) in (4.49) now, because \( n_i = r_i = 1 \) implies that \( A^{sk}_{i} = A^{11}_{i} = 0 \) as scalars do not have a nontrivial skew symmetric component. It can be shown that the case \( t_{i,1} = 1 \) for all \( i = 1, \ldots, n \), i.e. \( \beta = (1, \ldots, 1) \), and \( \gamma = \gamma_g \), where \( \gamma_g \) refers to the coordinate chart given in (4.52), yields a generic piece to which we shall call \( V_{\delta_g} \); this is shown in statement (vi) of theorem (4.6.2) below. However, it is important to note that this only holds true for \( s > 1 \). Generally speaking, this is because if \( r_j = s \) for some \( j \in \{1, \ldots, k\} \), then by lemma (3.1.1) the set of orthogonal matrices \( U_j \in O(s) \) consists of two compact disconnected components of the same topological dimension and thus one omits more than just a thin set by considering only one component of \( O(s) \). In our special case, this means that for \( s = 1 \), there is no generic \( V_{\delta_g} \). For \( s > 1 \) and \( \delta = \delta_g \), the vector of free parameters reduces to \( \tau = (\tau_\Sigma, \tau_{B^+}, \tau_{B^r}, \tau_\phi, \tau_D) \) with

\[
\begin{align*}
\tau_\Sigma &= (\sigma_1 - \sigma_2, \sigma_2 - \sigma_3, \ldots, \sigma_{n-1} - \sigma_n, \sigma_n) > 0 \\
\tau_{B^+} &= (\tilde{B}_{c,11}, \tilde{B}_{c,21}, \ldots, \tilde{B}_{c,m}) > 0 \\
\tau_{B^r} &= (\tilde{B}_{c,12}, \tilde{B}_{c,13}, \tilde{B}_{c,14}, \ldots, \tilde{B}_{c,m1}, \tilde{B}_{c,m2}, \tilde{B}_{c,m3}, \tilde{B}_{c,m4}, \ldots, \tilde{B}_{c,n1}, \tilde{B}_{c,n2}, \tilde{B}_{c,n3}, \tilde{B}_{c,n4}, \ldots, \tilde{B}_{c,nm}) \\
\tau_\phi &= (\phi_{1,1}, \phi_{1,2}, \phi_{1,3}, \dotsc, \phi_{1,s-1}, \phi_{2,1}, \phi_{2,2}, \phi_{2,3}, \dotsc, \phi_{2,s-1}, \phi_{3,1}, \phi_{3,2}, \phi_{3,3}, \dotsc, \phi_{3,s-1}, \dotsc, \phi_{n,1}, \phi_{n,2}, \phi_{n,3}, \dotsc, \phi_{n,s-1}) \\
\tau_D &= (d_{c,11}, \ldots, d_{c,sm})
\end{align*}
\]

and \((A_c, \tilde{B}_c, C_c, \tilde{D}_c)\) is given by

\[
\begin{align*}
A_{i,j,c} &= -\frac{\tilde{B}_i \tilde{B}_j'}{2\sigma_i} \\
A_{ij,c} &= \frac{\sigma_j \tilde{B}_i \tilde{B}_j' - \sigma_i \tilde{B}_j \tilde{B}_i'}{\sigma_i^2 - \sigma_j^2}, i,j = 1, \ldots, n \\
\tilde{B}_c &= [B_c, K_c] = \begin{pmatrix} \tilde{B}_1 \\ \vdots \\ \tilde{B}_n \end{pmatrix} = (B^{+}, B^{r}) \\
C_c &= [C_1, \ldots, C_n] = (U_1 (\tilde{B}_1 \tilde{B}_1')^{1/2}, \ldots, U_n (\tilde{B}_n \tilde{B}_n')^{1/2}) \\
\tilde{D}_c &= [D_c, E_c]; \quad D_c = \begin{pmatrix} d_{c,11} & \cdots & d_{c,1m} \\ \vdots & \ddots & \vdots \\ d_{c,s1} & \cdots & d_{c,sm} \end{pmatrix}, \quad E_c = I - C_c (I - A_c)^{-1} K_c
\end{align*}
\]
4.6. LYAPUNOV BALANCED CANONICAL FORM (OBER)

where \( U_i = U_i(\phi_{i,1}, \ldots, \phi_{i,s-1}) \), \( i = 1, \ldots, n \). Of course, the corresponding discrete time system matrices are again given by (A.44).

Remark 4.6.6 (The single output case: \( s = 1 \)). In the single output case, \( r_i = 1 \) for all \( i = 1, \ldots, k \). In particular, this is independent of the structure of \( \alpha \). Therefore, the unit vectors \( U_i \) degenerate to \( U_i = \pm 1 \), i.e., \( U_i \in O(1) \) (see remark (4.6.5) above): \( \tau_\delta \) is no longer a vector of free parameters. In this case the structural index \( \gamma \) becomes a (row) index \( \gamma = (\gamma_1, \ldots, \gamma_k) = (\pm 1, \ldots, \pm 1) \) of length \( k \), corresponding to the scalars \( U_i, i = 1, \ldots, k \). Note that there are \( 2^k \) different structural indices \( \gamma \) now. For fixed \( \alpha \) and \( \beta \), each of these pieces has the same topological dimension.

Clearly, the number of free parameters \( d_\delta \) may be different for different structural indices \( \delta \); see theorem (4.6.2) below. By

\[
\psi_\delta : V_\delta \to T_\delta
\]

we denote Ober’s Lyapunov balanced state-space parametrization attaching the free parameters \( \tau_\delta \) to the transfer functions \((l, k) \in V_\delta\).

An illustrative example

We give an illustrative example for Ober’s Lyapunov balanced state-space parametrization in the MIMO case: For \( m = 2, s = 2, n = 4 \) and \( \delta = (\alpha, \beta, \gamma) \) with

(i) \( \alpha = (3,1) \),
(ii) \( \beta \) determining the structure of \( \tilde{B}_c \) through \( r_1 = 2, r_2 = 1, t_{1,1} = 2, t_{1,2} = 3, t_{2,1} = 1 \) and the structure of \( A_c \) through \( q_1 = 1, h_{1,1} = 1, g_{1,1} = 1 \) and
(iii) \( \gamma \) corresponding to the parametrization of a generic subset of one connected piece of \( U_1 \in O(2) \subset \mathbb{R}^{2 \times 2} \) given in the proof of lemma (3.1.1) in chapter 3 and the usual unit vector parametrization of \( U_2 \in \mathbb{R}^2 \) through (4.52),

the continuous time state-space matrices have the following form:

\[
A_c = \begin{pmatrix}
0 -x & x & 0 & 0 \\
-x & 0 & x & 0 \\
0 & x & 0 & x \\
-x & 0 & x & 0
\end{pmatrix}, \quad B_c = [B_c, K_c] = \begin{pmatrix}
0 & \tilde{B}_{c,12} & x & x \\
0 & 0 & \tilde{B}_{c,23} & x \\
0 & 0 & 0 & 0 \\
\tilde{B}_{c,41} & x & x & x
\end{pmatrix}
\]

\[
C_c = \begin{pmatrix}
x & x & 0 & x \\
x & x & 0 & x
\end{pmatrix}, \quad \tilde{D}_c = [D_c, E_c], \quad D_c = \begin{pmatrix}
x & x & x \\
x & x & x
\end{pmatrix}
\]

and \( E_c = I - C_c(I - A_c)^{-1} K_c \). Here, \( x \) denotes unrestricted free parameters, \( x \) is a (nontrivial) function of the free parameters and the other nonzero entries denote positive free parameters. Note that we need 2 parameters for the Hankel singular values, 9 parameters for \( \tilde{B}_c \), 2 parameters for \( A_{11} \), 1 parameter for \( U_1 \in O(2) \), 1 parameter for \( U_2 \in \mathbb{R}^2 \) and 4 parameters for \( D_c \), which makes a total of 19 free parameters.

For the generic piece with

(i) \( \alpha = (1,1,1,1) \),
(ii) \( \beta = (1,1,1,1) \) and
(iii) \( \gamma = \gamma_\delta \) according to (4.52)

we get

\[
A_c = \begin{pmatrix}
x & x & x & x \\
x & x & x & x \\
x & x & x & x \\
x & x & x & x
\end{pmatrix}, \quad B_c = [B_c, K_c] = \begin{pmatrix}
\tilde{B}_{c,11} & x & x & x \\
\tilde{B}_{c,21} & x & x & x \\
\tilde{B}_{c,31} & x & x & x \\
\tilde{B}_{c,41} & x & x & x
\end{pmatrix}
\]

\[
C_c = \begin{pmatrix}
x & x & x & x \\
x & x & x & x \\
x & x & x & x \\
x & x & x & x
\end{pmatrix}, \quad \tilde{D}_c = [D_c, E_c], \quad D_c = \begin{pmatrix}
x & x & x \\
x & x & x
\end{pmatrix}, \quad E_c = I - C_c(I - A_c)^{-1} K_c
\]
Now we need 4 parameters for the Hankel singular values, 16 parameters for \( \hat{B}_c \), 4 parameters for the \( U_i \in \mathbb{R}^2, i = 1, \ldots, 4 \) and 4 parameters for \( D_c \), which makes a total of 28 free parameters. This, of course, coincides with the number of free parameters for the generic echelon piece on the left hand side of (4.10).

Finally, in order to motivate statement (v) of theorem (4.6.2) below, we consider an illustrative example showing some peculiarities which may occur at the boundary of the parameter space \( T_\delta \). We deal with the special case where \( m = 0, s = 1 \) and \( n = 2 \) and \( \delta = (\alpha, \beta, \gamma) \) with

(i) \( \alpha = (1, 1) \),
(ii) \( \beta = (1, 1) \) and
(iii) \( \gamma = (1, -1) \)

Let us fix two out of four parameters in (4.53):

\[ \tau_\delta = (\tau_\Sigma, \tau_{B^+}) = (\sigma_1 - \sigma_1, b) \]

Note that the first Hankel singular value is fixed to one. Moreover, note that the off-diagonals of \( A_c \) in (4.54) can be decomposed as follows:

\[ A_{ij,c} = A_{ij}^g + A_{ij}^s, \quad A_{ij}^s = -\frac{B_i B_j^T}{\sigma_i + \sigma_j}, \quad A_{ij}^g = \frac{\sigma_i \hat{B}_i \hat{B}_j^T - C_i^T C_j}{\sigma_i - \sigma_j} \]

(4.55)

It can be seen from (4.55) that for \( \sigma_i \to \sigma_j > 0 \) the part \( A_{ij}^g \) remains a continuous function of the parameters whereas the denominator \( \sigma_i - \sigma_j \) in \( A_{ij}^s \) tends to zero and the limiting behavior of \( A_{ij}^g \) therefore depends on the rate of convergence of \( \hat{B}_i \hat{B}_j^T - C_i^T C_j \to 0 \) in relation to \( \sigma_i - \sigma_j \). Note that for \( \sigma_i = \sigma_j \), it can be seen from the Lyapunov equations that \( \hat{B}_i \hat{B}_j^T = C_i^T C_j \) must hold true. The reason for the notation used in (4.55) is that \( A_{ij}^g = A_{ij}^s \) (symmetric part) and for \( \sigma_i \approx \sigma_j > 0 \) we get \( A_{ij}^g \approx -A_{ij}^g \) (skew symmetric part). In the special case considered here, the continuous time matrices have the following form:

\[ A_c = \begin{pmatrix} \frac{-1}{2} & \frac{b}{\sigma} \\ \frac{b}{\sigma} & \frac{-1}{2(1-\sigma)} \end{pmatrix} \quad \frac{-1}{2} \quad \frac{b}{2(1-\sigma)} \quad \frac{-b}{2-\sigma} \quad \frac{2b}{2-\sigma} \quad -b \quad -\frac{b}{1-\sigma} \]

\[ K_c = \begin{pmatrix} 1 \\ b \end{pmatrix} \quad C_c = \begin{pmatrix} 1 & -b \end{pmatrix} \quad E_c = 1 - C_c(I - A_c)^{-1} K_c \]

(4.56)

(4.57)

Note that the corresponding continuous time Markov parameters are easily obtained:

\[ (K_j)_{j \in \mathbb{N}} = 
\begin{pmatrix}
1 - \frac{1 + \frac{b^2}{2(1-\sigma)} + \frac{2b^2}{3} - \frac{b^2}{3(1-\sigma)}}{\frac{3}{2} + \frac{3b^2}{4(1-\sigma)} + \frac{b^2}{\sigma^2} - 1 - \sigma} - \frac{b^2}{2(1-\sigma)} + \frac{b^4}{\sigma^2} \left(1 - \sigma^2\right) - \frac{b^4}{\sigma(1-\sigma)} - \frac{b^4}{(2(1-\sigma))^2} \ldots
\end{pmatrix} \]

(4.58)

Let us now consider the following cases; see figure (4.1):

(1a) \( \sigma \to 0, b \to 0 \) and \( \frac{b}{\sigma} \to 0 \) results in

\[ \tau_\delta \to (0, 1, 1, 0) \quad A_c \to \begin{pmatrix} \frac{-1}{2} & 0 \\ 0 & 0 \end{pmatrix} \quad K_c \to \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad C_c \to \begin{pmatrix} 1 & 0 \end{pmatrix} \quad E_c \to \frac{1}{3} \]

and the corresponding sequence of transfer functions converges in \( T_{pl} \) to the transfer function

\[ k_c(s) = \frac{1}{s + 0.5} + \frac{1}{3} \]

(4.59)

of order one because

\[ (K_j)_{j \in \mathbb{N}} \to \left( \frac{1}{3}, 1, -\frac{1}{2}, \frac{1}{4}, \ldots \right) \]
4.6. LYAPUNOV BALANCED CANONICAL FORM (OBER)

Figure 4.1: Approaching the boundary of the parameter space $T_δ, δ = ((1, 1), (1, 1), (1, -1))$ in various ways: The $s$-axis shows the difference of the two Hankel singular values.

(a) In the cases (1a), (1b) and (1c) both Hankel singular values become equal and $τ_{B+} > 0$ is also violated (in the limit). In the cases (1d) and (2), only a single restriction is violated (in the limit).

(b) In the cases (3a), (3b) and (3c) the smallest Hankel singular value tends to zero, and $τ_{B+} > 0$ is also violated (in the limit). In case (3d), only a single restriction is violated: the smallest Hankel singular value tends to zero.

(1b) $σ → 0, b → 0$ and $\frac{δ}{σ} → x_0$ results in

$$τ_δ → (0, 1, 1, 0) \quad A_c → \begin{pmatrix} -\frac{1}{x_0^2} & x_0 \\ 0 & 0 \end{pmatrix} \quad K_c → \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad C_c → (1, 0) \quad E_c → \frac{1 + 2x_0^2}{3 + 2x_0^2} (4.59)$$

This is also a minimal realization of the second order limiting transfer function corresponding to

$$(K_j)_{j \in \mathbb{N}} → (\frac{1 + 2x_0^2}{3 + 2x_0^2}, 1, -\frac{1}{2}, -x_0^2, \ldots)$$

This limiting transfer function is given by

$$k_c(s) = \frac{s}{s^2 + 0.5s + x_0^2} + \frac{1 + 2x_0^2}{3 + 2x_0^2} \quad \text{corresponding to} \quad k(z) = \frac{2 - 2x_0^2}{(2x_0^2 + 3) + (4x_0^2 - 4)z + (2x_0^2 + 1)z^2} + \frac{1 + 2x_0^2}{3 + 2x_0^2}$$

and is contained in $V_{x_0}$ where $δ = (\alpha, \beta, γ)$ and $\alpha = (2), \beta$ only specifies $τ_1 = 1 (t_{1,1} = 1, q_1 = 1$ and $h_{1,1} = 1)$ follow automatically) and $γ = γ_o$ from (4.52). In fact, the realization in (4.59) is in Ober’s Lyapunov balanced canonical form. Note that different values for $x_0 ∈ \mathbb{R} \setminus \{0\}$ trivially yield different transfer functions of degree two.

(1c) $σ → 0, b → 0$ and $\frac{δ}{σ} → \infty$ results in

$$τ_δ → (0, 1, 1, 0) \quad A_c → \begin{pmatrix} -\frac{1}{x_0^2} & \infty \\ -\infty & 0 \end{pmatrix} \quad K_c → \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad C_c → (1, 0) \quad E_c → 1$$

and the corresponding sequence of transfer functions diverges in $T_p$ as

$$(K_j)_{j \in \mathbb{N}} → (1, 1, -\frac{1}{2} + 2 \lim \frac{b^2}{σ}, -\infty, \ldots)$$
Note that we have divergence in $T_{pl}$ despite the fact that both poles and zeros of $k_c(s)$ cancel at $s = -\frac{1}{2} \pm i \cdot \infty$ and $s = -\frac{3}{4} \pm i \cdot \infty$ in the limit, yielding the trivial transfer function $k_c(s) = 1$; see (Ribarits, 2000).

(1d) $\sigma \to 0$, $b \to 0$, and $\frac{b}{\sigma} \to \infty$ results in

$$
\tau_\delta \to (0, 1, 1, b) \quad A_c \to \left( \begin{array}{cc} -\frac{1}{2} & 0 \\ 0 & 0 \end{array} \right) \quad K_c \to \left( \begin{array}{c} 1 \\ b \end{array} \right) \quad C_c \to \left( \begin{array}{cc} 1 & 0 \end{array} \right) \quad E_c \to \frac{1}{3}
$$

and the corresponding sequence of transfer functions diverges in $T_{pl}$ as

$$(K_j)_{j \in \mathbb{N}} \to (1, 1 - b^2, \infty, \ldots)$$

Again, we have divergence in $T_{pl}$ despite the fact that both poles and zeros cancel in the limit, yielding $k_c(s) = 1$ independent of the behaviour of $b$ (as long as the other two assumptions are satisfied); see (Ribarits, 2000).

If both Hankel singular values become equal, but $\tau_{H+} > 0$ is not violated, the corresponding sequence of transfer functions diverges in $T_{pl}$; see (1d). If $b$ tends to zero very fast, we have $T_{pl}$ convergence to a lower degree transfer function; see (1a). If $b$ tends to zero a bit slower, but at a specific rate, we have $T_{pl}$ convergence to a degree two transfer function with equal Hankel singular values (which is what one might intuitively expect); see (1b). Finally, if $b$ tends to zero very slowly, then the corresponding sequence of transfer functions again diverges in $T_{pl}$; see (1c).

(2) $0 < \sigma < 1$ and $b \to 0$ results in

$$
\tau_\delta \to (\sigma, 1 - \sigma, 1, 0) \quad A_c \to \left( \begin{array}{cc} -\frac{1}{2} & 0 \\ 0 & 0 \end{array} \right) \quad K_c \to \left( \begin{array}{c} 1 \\ 0 \end{array} \right) \quad C_c \to \left( \begin{array}{cc} 1 & 0 \end{array} \right) \quad E_c \to \frac{1}{3}
$$

and the corresponding sequence of transfer functions converges in $T_{pl}$ to

$$
k_c(s) = \frac{1}{s + 0.5} + \frac{1}{3} \quad \text{corresponding to } k(z) = \frac{2 + 2z}{3 - z} + \frac{1}{3}
$$

of order one because

$$(K_j)_{j \in \mathbb{N}} \to \left( \frac{1}{3}, 1, -\frac{1}{2}, \frac{1}{4}, \ldots \right)$$

Without violating the restrictions $\tau_\Sigma > 0$, we get a limiting transfer function of lower degree by violating $\tau_{H+} > 0$ only! Note, moreover, that the limiting transfer function is independent of $\sigma$, i.e. infinitely many points at this part of the boundary of $T_\delta$ correspond to the same transfer function.

(3a) $\sigma \to 1$, $b \to 0$ and $\frac{b^2}{\pi^2 (1 - \sigma)} \to 0$ results in

$$
\tau_\delta \to (1, 0, 1, 0) \quad A_c \to \left( \begin{array}{cc} -\frac{1}{2} & 0 \\ 0 & 0 \end{array} \right) \quad K_c \to \left( \begin{array}{c} 1 \\ 0 \end{array} \right) \quad C_c \to \left( \begin{array}{cc} 1 & 0 \end{array} \right) \quad E_c \to \frac{1}{3}
$$

and the corresponding sequence of transfer functions converges in $T_{pl}$ to the transfer function of degree one in (4.60) above because

$$(K_j)_{j \in \mathbb{N}} \to \left( \frac{1}{3}, 1, -\frac{1}{2}, \frac{1}{4}, \ldots \right)$$
(3b) \( \sigma \to 1, b \to 0 \) and \( \frac{\sigma^2}{1-\sigma} \to x_0 \) results in

\[
\tau_{\delta} \to (1, 0, 1, 0) \quad A_c \to \begin{pmatrix} -\frac{1}{2} & 0 \\ 0 & -x_0 \end{pmatrix} \quad K_c \to \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad C_c \to \begin{pmatrix} 1 & 0 \end{pmatrix} \quad E_c \to \frac{1}{3}
\]

and the corresponding sequence of transfer functions again converges in \( T_{pl} \) to the transfer function of degree one in (4.60) above because

\[
(K_j)_{j \in \mathbb{N}} \to \left( \frac{1}{3}, 1, -\frac{1}{2}, \frac{1}{4}, \ldots \right)
\]

(3c) \( \sigma \to 1, b \to 0 \) and \( \frac{\sigma^2}{1-\sigma} \to \infty \) results in

\[
\tau_{\delta} \to (1, 0, 1, 0) \quad A_c \to \begin{pmatrix} -\frac{1}{2} & 0 \\ 0 & -\infty \end{pmatrix} \quad K_c \to \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad C_c \to \begin{pmatrix} 1 & 0 \end{pmatrix} \quad E_c \to \frac{1}{3}
\]

The corresponding sequence of transfer functions diverges in \( T_{pl} \) which can be seen from the fact that the highest terms of \( \frac{\sigma^2}{1-\sigma} \) in \( C_c A_c^l K_c \) are of the form \( \left( \frac{\sigma^2}{1-\sigma} \right)^l b^2 \) and these terms diverge for \( l \) large enough. This occurs in spite of the fact that there is a pole-zero cancellation at \( s = -\infty \) in the limit, yielding the transfer function of degree one in (4.60) above.

(3d) \( \sigma \to 1 \) and \( b \to 0 \) results in

\[
\tau_{\delta} \to (1, 0, 1, b) \quad A_c \to \begin{pmatrix} -\frac{1}{2} & b \\ -b & -\infty \end{pmatrix} \quad K_c \to \begin{pmatrix} 1 \\ b \end{pmatrix} \quad C_c \to \begin{pmatrix} 1 & -b \end{pmatrix} \quad E_c \to \frac{1}{3}
\]

However, the corresponding sequence of transfer functions diverges in \( T_{pl} \) because

\[
(K_j)_{j \in \mathbb{N}} \to \left( \frac{1}{3}, 1 - b^2, \infty, \ldots \right)
\]

Again, this holds true despite the fact that there is a pole-zero cancellation at \( s = -\infty \) in the limit, yielding the transfer function of degree one in (4.60) above.

If the second Hankel singular value tends to zero, then \( b \) must also tend to zero fast enough in order to have \( T_{pl} \) convergence to a transfer function of order one; see (3a) and (3b). If \( b \) tends to zero slowly or does not tend to zero at all, the sequence of Markov parameters diverges (and thus yields divergence in \( T_{pl} \)) although there is a pole-zero cancellation at \( s = -\infty \) in the limit; see (3c) and (3d).

### 4.6.2 Topological properties

The following theorem summarizes results in (Bauer, 1998) and states the main topological properties of Ober’s Lyapunov balanced state-space parametrization:

**Theorem 4.6.2.** Ober’s Lyapunov balanced state-space parametrization

\[
\psi_{\delta} : V_{\delta} \to T_{\delta} \\
(l, k) \mapsto \tau_{\delta}
\]

where \( \delta = (\alpha, \beta, \gamma) \) is described in theorem (4.6.1), has the following properties:

(i) \( T_{\delta} \) is an open subset of \( \mathbb{R}^{d_{\delta}} \), where the dimension \( d_{\delta} \) depends on the multi index \( \delta \).

(ii) \( \psi_{\delta} : V_{\delta} \to T_{\delta} \) is a \( (T_{pl},-) \) homeomorphism.

(iii) \( \{ V_{\delta}, |\alpha| = n \} \) is a disjoint partition of \( M_{\delta}(n) \).
(iv) \( \pi(\varphi_\delta(T_{\delta_\beta})) = \bigcup_{\beta \leq n} M_\beta(i) \) where \( \delta_\beta = \left( (1, \ldots, 1), (1, \ldots, 1), \gamma_\beta \right) \) is the multi index corresponding to the generic piece \( V_{\delta_\beta} \) in the multi output case; see remark (4.6.5).

(v) There always exist parameter vectors \( \tau_\delta \in \tilde{T}_\delta \setminus T_\delta \) such that \( l > 1 \) different sequences \( \{\tau_{\delta,l}^{(i)}\}_{i \in \mathbb{N}} \), \( \tau_{\delta,l}^{(i)} \in \tilde{T}_\delta \), \( i = 1, \ldots, l \), converging to the same parameter vector \( \tau_\delta \) for \( t \to \infty \) can be found where \( \pi(\varphi_\delta(\tau_{\delta,l}^{(i)})) \to (l_i, k_i) \) and \( (l_i, k_i) \neq (l_j, k_j) \) for \( i \neq j \).

(vi) \( V_\delta \) is open in \( \tilde{V}_\delta \). \( \tilde{V}_{\delta_\beta} = \tilde{M}_\beta(n) \).

(vii) \( \pi(\varphi_\delta(T_{\delta_\beta})) = \tilde{V}_{\delta_\beta} \).

Proof. These results are collected from various theorems in section 2.2 in (Bauer, 1998); see also (Bauer and Dëstler, 1999). Note that for the case \( r_1 = 1 \), the unit vectors \( U_i \in \mathbb{R}^6 \) are parametrized using stereographic projections there, but it is easy to see that the results remain valid if we use polar coordinates as in (4.52).

Remark 4.6.7. We give a short discussion of these results:

(i) Again, this means that the parameters are really free in the sense that they are not restricted to a thin subset of \( \mathbb{R}^{d_\beta} \). Compared to echelon forms, the structure of the set \( T_\delta \) of admissible parameter values is less complicated: Some components of \( \tau_\delta \) lie in an open interval, others have to be positive and the rest is unrestricted. Consequently, \( T_\delta \) is not dense in \( \mathbb{R}^{d_\beta} \).

(ii) assures bijectivity and continuity of the parametrization on the pieces \( V_\delta \). Note that the inverse mapping \( \psi_\delta^{-1} = \pi \circ \varphi_\delta \) is known to be continuous because both \( \pi \) and \( \varphi_\alpha \) are continuous. An analogous remark on consistency as in (ii) in remark (4.2.3) for echelon canonical forms can also be made at this point.

(iii) implies that the set of mappings \( (\psi_\delta, \lvert \alpha \rvert = n) \) constitutes a canonical form for \( M_\beta(n) \). Note that even in the simplest case of SISO transfer functions of order one, this partition contains two pieces corresponding to \( \gamma = (+1) \) and \( \gamma = (-1) \). In general, Ober’s Lyapunov balanced canonical form needs many more pieces to cover \( M_\beta(n) \) than the echelon canonical form does for \( M_\beta(n) \).

(iv) If we take the closure of \( \varphi_\delta(T_{\delta_\beta}) \subseteq S_m(n) \) within the subset of \( S(n) \) corresponding to stable systems, the corresponding set of transfer functions is exactly the set of all stable transfer functions of equal and lower McMillan degree. Note that this is again a difference to echelon canonical forms where \( \pi(T_{\alpha_\beta}) = \bigcup_{\beta \leq \alpha} V_\beta \), and this set is a proper subset of \( \bigcup_{i \leq n} M(i) \) in the multi output case. Note, moreover, that statement (iv) is not in terms of the closure of the parameter space \( T_{\delta_\beta} \in \mathbb{R}^{d_{\delta_\beta}} \) because the mapping \( \varphi_\delta \) (and, hence also \( \pi \circ \varphi_\alpha \)) cannot be uniquely extended to the boundary of \( T_{\delta_\beta} \). This follows from statement (v) and the discussion of the last illustrative example above.

(v) states that the same parameter vector at the boundary of \( T_\delta \) can represent more than one transfer function: Different sequences \( \tau_{\delta,l}^{(i)} \to \tau_\delta \) where \( \tau_{\delta,l}^{(i)} \in T_\delta \) and \( \tau_\delta \in \tilde{T}_\delta \setminus T_\delta \), \( i = 1, \ldots, l \), converge to the same parameter vector \( \tau_\delta \), but the corresponding limits of the transfer functions \( \pi(\varphi_\delta(\tau_{\delta,l}^{(i)})) \to (l_i, k_i) \) are different. Clearly, this is in contrast to statement (v) in theorem (4.2.2) for the echelon canonical forms, where any point at the boundary of \( T_{\alpha_\beta} \) corresponds to one transfer function in \( V_\beta, \beta < \alpha \).

(vi) The set \( V_\delta \) is always open in its closure. This is important again in connection with the consistency result in section (2.3.4); see statement (v) in remark (4.1.2). The second statement involves the particular multi index \( \delta_\beta \) of statement (iv). Note that the dimension of the parameter vector is \( d_{\delta_\beta} = 2n + m(n + s) \) in this case, i.e., equal to the dimension of the manifold \( M_\beta(n) \) itself\(^4\). Here, we get \( \tilde{V}_{\delta_\beta} = \tilde{M}_\beta(n) = \bigcup_{i \leq n} M(i) \) where the last equality follows from (vii) together with (iv); the closure is taken within the set of stable transfer functions. This means that transfer functions having this particular multi index are not only open in their closure but also form a dense subset of all stable transfer functions of McMillan degree less than or equal to \( n \). This is the reason for calling this \( V_{\delta_\beta} \) a generic neighborhood of \( M_\beta(n) \).

(vii) states that the same transfer functions are described in the closure of \( \varphi_\delta(T_{\delta_\beta}) \in S_m(n) \) (within the subset of \( S(n) \) corresponding to stable systems) and in the closure of the corresponding space of transfer functions \( V_{\delta_\beta} \) (within the set of stable transfer functions). Note that this is clearly different in case of echelon canonical forms; see statement (vii) in theorem (4.2.2) and statement (iv) above.

\(^4\)The set \( M_\beta(n) \) can be shown to be a submanifold of \( M(n) \) of the same dimension as \( M(n) \).
4.7. STOCHASTICALLY BALANCED CANONICAL FORM (OBER)

4.7.1 Introduction

Stochastic balancing for continuous time state-space representations is introduced in the appendix (A.3). As it is the case there, we will first restrict ourselves to \( s \times s \) transfer functions \( k_c(s) \) (and \( k(z) \)) and then include the transfer function \( l_c(s) \) (and \( l(z) \)) in a second step. Hence, the same sets of transfer functions \( \mathbb{M}_c^+ \) and \( \mathbb{M}_{\text{mp}}^+ \) as introduced in the appendices (A.3) and (A.7) are used below.

We will not describe how to separately derive a canonical form for the new class of transfer functions from first principles, though this would be possible and is in fact done for the case of minimum phase balanced systems in section (4.8). Instead, we make use of the homeomorphic relations between various classes of transfer functions as given in the diagram at the end of appendix (A.3)\(^5\). This allows us to carry the canonical form over from one class of transfer functions to another without having to repeat the basic construction of finding a unique representative.

It is clear from the diagram in section (A.3.5) and the preceding discussion in appendix (A.3) that the model class \( \mathbb{M}_c^+ \) can be homeomorphically linked to the model class \( \mathbb{M}_{\text{mp}}^+ \) through the compository mapping

\[
Y = S_{h, h}^{\text{mp}} \circ S_{(I - PQ)^{-1/2}} \circ S_{h}^{s, \text{mp}} \circ S_{h}^{s, (0,1)} \circ S_{h}^{s, (0,1)}
\]  

(4.61)

Starting with Ober’s Lyapunov balanced state-space representation \((\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)\) of \( k_c(s) \in \mathbb{M}_c^+ \) with \( P = Q = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n) \), \( Y \) takes the following form (it is just the composition of the mappings described in (A.29), (A.32), a state transformation \( T = (I - PQ)^{-1/2} \) and (A.37)):

\(^5\)In fact, in (Ober, 1991) and (Ober, 1999), balanced canonical forms are also derived for allpass, bounded real, positive real and transfer functions in \( \mathbb{M}(n) \) in this manner.
\[ T : \mathbb{M}_n^1 \rightarrow \mathbb{M}_n^1 \]

\[
\begin{bmatrix}
A_c \\
C_c \\
D_c
\end{bmatrix} =
\begin{bmatrix}
\frac{W^{-1/2} A_c W^{1/2} - W^{-1/2} B_c (B_c^T \Sigma_c - C_c) W^{-1/2}}{C_c W^{-1/2} - B_c W^{1/2} \Sigma_c}
\end{bmatrix}
\]

Here, \((A_\rho, B_\rho, C_\rho, D_\rho) = (\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)\) is obtained from \((\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)\) by replacing \(\sigma_i\) by

\[ \rho_i = \frac{\sigma_i}{\sqrt{1 + \sigma_i^2}} \]

in Ober’ Lyapunov balanced canonical form of \((\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)\), i.e. \((A_\rho, B_\rho, C_\rho, D_\rho)\) is also in Ober’s Lyapunov balanced canonical form. Additionally, \(\Sigma_\rho = \text{diag}(\rho_1, \ldots, \rho_n)\) and \(W_\rho = (I - \Sigma_\rho^2)\).

It is also clear from section (4.3) that \(\Upsilon(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)\) is (left) stochastically balanced, and thus Ober’s Lyapunov balanced canonical form directly induces a (left) stochastically balanced canonical form for \(k_c(s) \in \mathbb{M}_n^1\) through \(\Upsilon\).

We can now set \(\hat{K}_c = \hat{K}_c E_c\) where \(\hat{K}_c = W_\rho^{-1/2} B_\rho\) has the same structure as \(B_\rho = \tilde{B}_c\) and is considered to contain the new free parameters: \(\hat{K}_c = (K'_1, \ldots, K'_k)\). This yields

\[
K_c = \hat{K}_c E_c = \begin{pmatrix} K_1 \\ \vdots \\ K_k \end{pmatrix} E_c \text{ where } K_i \in \mathbb{R}^{n \times s} \text{ are nonnegative upper triangular}
\]

\[
C_c = \begin{pmatrix} C_1 & \ldots & C_k \end{pmatrix} \text{ where } C_i = \left[U_i \Delta_i - \rho_i K_i' \Delta_i, 0_{s \times (s - i - 1)} \right] \text{ with } \Delta_i = (K_i' K_i')^{1/2} \text{ and } U_i' U_i = I_i
\]

\[ E_c \ldots \text{ invertible as } (D_c + D_c') > 0 \]

Clearly, also \(A_c\) inherits the block structure of \(A_\rho\) and can be calculated in terms of the new free parameters \(\rho_i\) and \(\hat{K}_c = (K'_1, \ldots, K'_k)\); see theorem (4.7.1) below. What remains in the first step is to choose \(E_c\) in such a way that the corresponding discrete time \(E\) obtained from (A.44) by

\[ E = E_c + C_c (I - A_c)^{-1} K_c = \left(I + C_c (I - A_c)^{-1} \hat{K}_c\right) E_c \]

is the identity matrix. Now, as \(E_c\) is guaranteed to be invertible and \(E_c + C_c (I - A_c)^{-1} K_c = k_c(1)\) must also be invertible because \(k_c(s)\) has no zeros in the right half plane, \((I + C_c (I - A_c)^{-1} \hat{K}_c)\) also has to be invertible and we can choose \(E_c = (I + C_c (I - A_c)^{-1} \hat{K}_c)^{-1}\). Note that this \(E_c\) is completely determined by the other parameters.

The second step consists of incorporating the transfer function \(l_c(s)\). This is done by simply adding the state-space matrices \(B_c\) and \(D_c\) to \((A_c, K_c, C_c, E_c)\) and considering the entries of \(B_c\) and \(D_c\) to be free. Of course, the pair \((A_c, B_c)\) may now no longer be controllable implying that the McMillan degree of \(l_c(s)\) on its own may be less than \(n\).

The procedure outlined above yields a (left) stochastically balanced canonical form for \((l_c, k_c) \in \mathbb{M}_n^1\), where we also use the symbol \(\mathbb{M}_n^1\) for the non square case to denote the set of all rational, proper and stable \(s \times (m + s)\) continuous time transfer functions of order \(n\) where \((l_c(\infty), k_c(\infty) = D_c\) for an arbitrary \(D_c \in \mathbb{R}^{s \times (m + s)}\) and \(k_c\) additionally satisfies the strict minimum phase property.

Applying (A.44) yields a (left) stochastically balanced canonical form for \((l, k) \in \mathbb{M}_n^1\):

**Definition 4.7.1 (Right and Left Stochastic Balancing).** A minimal state-space representation \((A, B, C, D, K)\) of \((l, k) \in \mathbb{M}_n^1\) is called right stochastically balanced if the minimal solutions to the positive real Riccati equations (see (A.18) and (A.21))
\[ P - \bar{A}P\bar{A}' - (\bar{B} - \bar{A}P\bar{C}')(\bar{D} + \bar{D}' - \bar{C}P\bar{C}')^{-1}(\bar{B} - \bar{A}P\bar{C})' = 0 \]  
(4.64)

\[ Q - \bar{A}'Q\bar{A} - (\bar{C}' - \bar{A}'Q\bar{B})(\bar{D} + \bar{D}' - \bar{B}'Q\bar{B})^{-1}(\bar{C}' - \bar{A}'Q\bar{B})' = 0 \]  
(4.65)

satisfy \( P = Q = \Sigma_p = \text{diag}(\rho_1, \ldots, \rho_n) \), where \((\bar{A}, \bar{B}, \bar{C}, \bar{D}) = (A, APC' + KK', C, \frac{1}{2}(CPC' + I'))\). Note that the first Riccati equation is in fact a Lyapunov equation of the form \( P - AP\bar{A} - KK' = 0 \). It is called left stochastically balanced if the minimal solutions to the same equations (4.64) and (4.65) satisfy \( P = Q = \Sigma_p = \text{diag}(\rho_1, \ldots, \rho_n) \), where now \((\bar{A}, \bar{B}, \bar{C}, \bar{D}) = (A, K'QA + FC', \frac{1}{2}(K'QK + I'))\). Note that the second Riccati equation is in fact a Lyapunov equation of the form \( Q - A'QA - C'C = 0 \). The \( \rho_i \) satisfy \( 1 > \rho_1 \geq \cdots \geq \rho_n > 0 \) and are called canonical correlation coefficients.

**Remark 4.7.1.** The calculation of \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) from \((A, K, C, I)\) in the definition above is given by the inverse of the mapping attaching strictly minimum phase systems \((A, K, C, E)\) or \((A, B, L, E)\) to strictly positive real spectral summands \((A, B, C, D)\) as described in theorem (A.2.2) in the appendix (A.2).

We are now ready to state the following

**Theorem 4.7.1 (Ober's stochastically balanced canonical form).** Let a transfer function \((l, k) \in \mathbb{M}_{\text{amp}}(n)\) be given and let the transfer function \(k_c(s) = \Upsilon^{-1}(k(\rho_1^{-1}(s))) \in \mathbb{M}_{\text{amp}}^+(n)\) have the structural index \(\delta = (\alpha, \beta, \gamma)\) where \(\alpha\) is given in (4.46), \(\beta\) is given by (4.48) and (4.49) and \(\gamma\) specifies some subset of the set of all possible \(U_i \in \mathbb{R}^{n_i \times n_i}, i = 1, \ldots, k\) in (4.66) below. Moreover, let us denote the whole set of stable and strictly minimum phase transfer functions \((l, k)\) with this multi index by \(V_{\delta}^{(1)}\). Then there exists a unique minimal, stable, strictly minimum phase and (discrete time left) stochastically balanced state-space representation with matrices

\[
A = (I + A_c)(I - A_c)^{-1} \\
(B, K) = \sqrt{2}(I - A_c)^{-1} \bar{B}_c \\
C = \sqrt{2}C_c(I - A_c)^{-1} \\
(D, I) = \bar{D}_c + C_c(I - A_c)^{-1} \bar{B}_c
\]

where \((A_c, \bar{B}_c, C_c, \bar{D}_c)\) have the following form:

\[
\bar{B}_c = [B_c, K_c] \\
B_c = \begin{pmatrix} b_{c,11} & \cdots & b_{c,1m} \\
\vdots & \ddots & \vdots \\
b_{c,n1} & \cdots & b_{c,nm} \end{pmatrix} \\
K_c = \bar{K}_c E_c = \begin{pmatrix} K_1 \\
\vdots \\
K_k \end{pmatrix} E_c \text{ where } K_i \in \mathbb{R}^{n_i \times n_i} \text{ are nonnegative upper triangular as given in (4.47)}
\]

\[
C_c = \begin{pmatrix} C_1 & \cdots & C_k \end{pmatrix} \text{ where } C_i = \begin{pmatrix} U_i \Delta_i - \rho_i K_i' \otimes 0_{n_i \times (n_i - r_i)} \end{pmatrix} \text{ with } \Delta_i = (K_i' K_i)^{1/2} \text{ and } U_i' U_i = I_{r_i}
\]

\[
A_c = \begin{pmatrix} A_{11} & \cdots & A_{1h} \\
\vdots & \ddots & \vdots \\
A_{h1} & \cdots & A_{hh} \end{pmatrix} \text{ where } A_{ij} \in \mathbb{R}^{n_i \times n_j} \text{ have the following form:}
\]

\[
A_{ij} = \frac{1}{\rho_i - \rho_j} \begin{pmatrix} \rho_j (1 - \rho_j^2) K_i' K_j' & 0 & \rho_i (1 - \rho_j^2) \Delta_i U_i U_j \Delta_j & 0 \\
0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} K_i' U_i \Delta_j & 0 \\
0 & 0 \end{pmatrix} \text{ for } i \neq j
\]

\[
A_{ii} = A_{ii}^* = \begin{pmatrix} A_{11} & \cdots & A_{1h} \\
\vdots & \ddots & \vdots \\
A_{h1} & \cdots & A_{hh} \end{pmatrix} = \begin{pmatrix} \Delta_i^2 & 0 & 0 \\
0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} K_i' U_i \Delta_j & 0 \\
0 & 0 \end{pmatrix}
\]

\[
\bar{D}_c = [D_c, E_c] \\
D_c = \begin{pmatrix} d_{c,11} & \cdots & d_{c,1m} \\
\vdots & \ddots & \vdots \\
d_{c,n1} & \cdots & d_{c,nm} \end{pmatrix}, \quad E_c = (I + C_c(I - A_c)^{-1} \bar{K}_c)^{-1}
\]

(4.66)
where the notation is the same as in theorem (4.6.1).

Conversely, let a parameter vector \( \tau_{\delta}^{(1)} = (\tau_\rho, \tau_{B^+}, \tau_{B^*}, \tau_{A^+}, \tau_{A^*}, \tau_\phi, \tau_D) \) be given,

\[
\begin{align*}
\tau_\rho &= (\rho_1 - \rho_2, \rho_2 - \rho_3, \ldots, \rho_{k-1} - \rho_k, \rho_k) > 0 \text{ and } \tau_\rho < 1 \\
\tau_{B^+} &= (K_1, t_1, \ldots, K_{t_1}, \ldots, K_{t_{k-1}}, \ldots, K_{t_k}, t_k) > 0 \\
\tau_{B^*} &= \text{unrestricted entries } x \text{ in the matrices } K_i \text{ for } i = 1, \ldots, k \\
\text{unrestricted entries in the matrix } B_e \text{ in (4.66) above} \\
\tau_{A^+} &= (a_{g_1, 1, h_1, 1}, \ldots, a_{g_1, a_1, h_1, a_1}, \ldots, a_{g_k, 1, h_k, 1}, \ldots, a_{g_k, a_k, h_k, a_k}, \alpha_1, 2, \ldots, \alpha_1, m - r_1, \ldots, \alpha_k, 2, \ldots, \alpha_k, n - r_k) > 0 \\
\text{where only nonzero } \alpha_{ij} \text{ are included} \\
\tau_{A^*} &= \text{unrestricted entries in the strictly lower triangular parts of the skew symmetric matrices } A_{11}^{1,2} \text{ and } \text{unrestricted entries } x \text{ in the matrices } A_{11}^{1,2} \text{ for } i = 1, \ldots, k; \text{ see (i) and (ii) above and below (4.49)} \\
\tau_\phi &= \text{free parameters for the matrices } U_i \in \mathbb{R}^{n \times n}, i = 1, \ldots, k \\
\tau_D &= (d_{c,11}, \ldots, d_{c,nn}) \\
\end{align*}
\]

(4.67)

where \( \tau_\rho \) consists of the differences of the distinct canonical correlations and the rest is given as in theorem (4.6.1). Let \( T_{\delta}^{(1)} \) be the parameter space containing all vectors \( \tau_{\delta}^{(1)} \) satisfying the restrictions above and let

\[
\begin{align*}
\varphi_{\delta}^{(1)} : T_{\delta}^{(1)} &\rightarrow S_m(n) \\
\tau_{\delta}^{(1)} &\mapsto (A(\tau_{\delta}^{(1)}), B(\tau_{\delta}^{(1)}), C(\tau_{\delta}^{(1)}), D(\tau_{\delta}^{(1)}), K(\tau_{\delta}^{(1)}))
\end{align*}
\]

be the mapping attaching the discrete time system matrices to the parameters as given in the first part of this theorem. Then \( \varphi_{\delta}^{(1)}(\tau_{\delta}^{(1)}) = T_{\delta}^{(1)} \) is a discrete time, minimal, stable, strictly minimum phase and (left) stochastically balanced state-space realization of a transfer function in \( V_{\delta}^{(1)} \).

Proof. The proof follows immediately from the discussion above. Note that the formula for the construction of \( A_e \) in (4.66) can be derived in the same manner as has been shown for the other system matrices in (4.63), i.e., by using the state-space representation of the mapping \( T \) in (4.62).

\[ \square \]

Remark 4.7.2. In theorem 7.1 in (Ober, 1991), the matrices \( A_{11} \) in (4.66) are defined in a slightly different way; see also remark (4.6.1) below theorem (4.6.1). Moreover, only the parametrization of the square transfer function \( k_c(s) \) (and \( k(z) \)) is considered and the continuous time \( E_c \)-matrix is not required to be such that the corresponding discrete time matrix becomes the identity. Finally, the parametrization is called minimum phase balanced canonical form in (Ober, 1991). The reason for this is that Ober uses a different notion of minimum phase balancing in definition 7.2 in (Ober, 1991) which is easily shown to coincide with our definition of stochastic balancing.

Of course, remarks analogous to those in remarks (4.6.3), (4.6.4), (4.6.5) and (4.6.6) below theorem (4.6.1) can also be made at this point.

An illustrative example

Illustrative examples analogous to those in section (4.6) for Ober’s Lyapunov balanced canonical form can also be given at this point.

4.7.2 Topological properties

The following theorem states the main topological properties of Ober’s stochastically balanced state-space parametrization:
Theorem 4.7.2. Ober’s (left) stochastically balanced state-space parametrization

\[
\psi_{\delta}^{(1)} : \psi_{\delta}^{(1)} \rightarrow T_{\delta}^{(1)} \\
(l, k) \rightarrow \tau_{\delta}^{(1)}
\]

where \( \delta = (\alpha, \beta, \gamma) \) is described in theorem (4.7.1), has the properties (i) - (vii) described in theorem (4.6.2). Clearly, \( V_{\delta}, T_{\delta}, \varphi_{\delta} \) etc. have to be replaced by \( V_{\delta}^{(1)}, T_{\delta}^{(1)}, \varphi_{\delta}^{(1)} \) etc.

Proof. The theorem follows from the fact that the mapping \( T \) in (4.61) as well as its state-space representation in (4.62) and the state-space representation of the bilinear transformation in (A.44) are homeomorphic. \( \square \)

Remarks analogous to those in remarks (4.6.7), (4.6.8) and (4.6.8) below theorem (4.6.2) can also be made at this point. Note that now also the strict minimum phase assumption is built in without any need for additional restrictions on the parameters.

Remark 4.7.3. As a general remark, we want to mention the role of stochastic balancing in connection with model reduction. If one aims at a balanced truncation method for state-space models where the criterion to be minimized is the asymptotic likelihood function, then the author in (Scherrer, 2002) has shown that stochastically balanced realizations (and minimum phase balanced realizations; see section (4.8) below) are (locally) optimal and therefore preferable to Lyapunov balanced ones.

4.8 Minimum phase balanced canonical form (McGinnie)

4.8.1 Introduction

It is shown in appendix (A.3.4) that right (left) minimum phase balanced and right (left) stochastically balanced state-space realizations of a continuous time transfer function in \( \mathbb{M}_{\text{amp}}(n) \) are connected through a diagonal state transformation: If \((A_c, K_c, C_c, E_c)\) is right (left) stochastically balanced with \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n), 0 < \rho_i < 1 \), then a (homeomorphic) state transformation \( T = (I - PQ)^{-1/4} \) (or \( T = (I - PQ)^{-1/4} \), respectively) makes the realization right (left) minimum phase balanced. Of course, this could be used to directly obtain a left minimum phase balanced canonical form for systems \((l, k) \in \mathbb{M}_{\text{amp}}(n)\) from theorem (4.7.1). Discrete time minimum phase balancing is defined as follows:

Definition 4.8.1 (Right and left minimum phase balancing). A minimal state-space representation \((A, B, C, D, K)\) of \((l, k) \in \mathbb{M}_{\text{amp}}(n)\) is called right minimum phase balanced, if the (unique) solutions to the Lyapunov equations

\[
P - APA' - KK' = 0
\]

\[
Q - \bar{A}'Q\bar{A} - C'C = 0
\]

with \( \bar{A} = (A - KC) \) satisfy \( P = Q = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n) \). It is called left minimum phase balanced, if the (unique) solutions to the Lyapunov equations

\[
P - \bar{A}P\bar{A}' - KK' = 0
\]

\[
Q - A'Q A - C'C = 0
\]

satisfy \( P = Q = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n) \). The \( \sigma_i \) satisfy \( \sigma_1 \geq \cdots \geq \sigma_n > 0 \) and are called minimum phase singular values.
In (McGinnie, 1993), however, a separate canonical form for \((l, k) \in \mathbb{M}_{n+m}(n)\) is derived from first principles. As again minimum phase balanced realizations always exist, are never unique and are related to each other through the same state transformations (A.28) as in the case of Lyapunov balanced realizations, the construction is exactly the same as in the Lyapunov balanced case. The only difference in the form of the parametrization is due to the different continuous time Lyapunov equations: (A.33) and (A.34) are now used in the right minimum phase balanced case and (A.35) and (A.36) are used in the left minimum phase balanced case. As in section (4.7) above, we also use the symbol \(\mathbb{M}_{n+m}(n)\) for the non square case to denote the set of all rational, proper and stable \(s \times (m + s)\) continuous time transfer functions of order \(n\) where \((l, k, \infty, k_c, \infty) = D_c\) for an arbitrary \(D_c \in \mathbb{R}^{s \times (m + s)}\) and \(k_c\) additionally satisfies the strict minimum phase property:

**Theorem 4.8.1 (McGinnie’s minimum phase balanced canonical form).** Let a transfer function \((l, k) \in \mathbb{M}_{n+m}(n)\) be given and let the transfer function \(k_c(s) = k(\rho_i^{-1}(s)) \in \mathbb{M}_{n+m}(n)\) have the structural index \(\delta = (\alpha, \beta, \gamma)\) where \(\alpha\) is given in (4.46), \(\beta\) is given by (4.48) and (4.49) and \(\gamma\) specifies some subset of the set of all possible \(U_i \in \mathbb{R}^{s \times r_i}, i = 1, \ldots, k\) in (4.72) below. Moreover, let us denote the whole set of stable and strictly minimum phase transfer functions \((l, k)\) with this multi index by \(Y_\delta^{(2)}\). Then there exists a unique minimal, stable, strictly minimum phase and (discrete time right) minimum phase balanced state-space representation with matrices

\[
\begin{align*}
A &= (I + A_c)(I - A_c)^{-1} \\
(B, K) &= \sqrt{2}(I - A_c)^{-1} \tilde{B}_c \\
C &= \sqrt{2}C_c(I - A_c)^{-1} \\
(D, I) &= \tilde{D}_c + C_c(I - A_c)^{-1} \tilde{B}_c
\end{align*}
\]

where \((A_c, \tilde{B}_c, C_c, \tilde{D}_c)\) have the following form:

\[
\begin{align*}
\tilde{B}_c &= [B_c, K_c] \\
B_c &= \begin{pmatrix} b_{c,11} & \cdots & b_{c,1m} \\ \vdots & \ddots & \vdots \\ b_{c,n1} & \cdots & b_{c,nm} \end{pmatrix} \\
K_c &= \begin{pmatrix} K_1 \\ \vdots \\ K_k \end{pmatrix} \text{ where } K_i \in \mathbb{R}^{n_i \times s} \text{ are nonnegative upper triangular as given in (4.47)}
\end{align*}
\]

\[
C_c = E_c \tilde{C}_c = E_c \begin{pmatrix} C_1, \ldots, C_k \end{pmatrix}
\]

where \(C_i = \left[ \sigma_i K_i^{\alpha_i} + \sqrt{1 + \sigma_i^2} \Delta_i U_i 0_{s \times (m_n - r_i)} \right] \) with \(\Delta_i = (K_i^{\alpha_i} K_i^{\beta_i})^{1/2} \) and \(U_i U_i = I_{r_i}\)

\[
A_c = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kk} \end{pmatrix} \text{ where } A_{ij} \in \mathbb{R}^{n_i \times n_j} \text{ have the following form:}
\]

\[
A_{ij} = \frac{1}{\sigma_i - \sigma_j} \begin{pmatrix} \sigma_j (1 + \sigma_i^2) K_i^{\alpha_i} K_j^{\beta_j} - \sigma_i \sqrt{1 + \sigma_i^2} \Delta_i U_i U_j \Delta_j & 0 \\ 0 & 0_{(m_n - r_i) \times (m_n - r_j)} \end{pmatrix}
\]

for \(i \neq j\)

\[
A_{ii} = A_{ii}^{\alpha_i} + A_{ii}^{\beta_i} = \begin{pmatrix} A_{ii}^{11} & A_{ii}^{12} \\ A_{ii}^{21} & A_{ii}^{22} \end{pmatrix} - \frac{1}{2\sigma_i} \begin{pmatrix} \Delta_i^2 & 0 \\ 0 & 0_{(m_n - r_i) \times (m_n - r_i)} \end{pmatrix}
\]

\[
\tilde{D}_c = [D_c, E_c]
\]

\[
D_c = \begin{pmatrix} d_{c,11} & \cdots & d_{c,1m} \\ \vdots & \ddots & \vdots \\ d_{c,s1} & \cdots & d_{c,sm} \end{pmatrix}, \quad E_c = \left( I + \tilde{C}_c(I - A_c)^{-1} K_c \right)^{-1}
\]

(4.72)

where the notation is the same as in theorem (4.6.1).
4.8. MINIMUM PHASE BALANCED CANONICAL FORM (MCGINNIE)

Conversely, let a parameter vector \( \tau_\delta^{(2)} = (\tau_\Sigma, \tau_{B+}, \tau_{B'}, \tau_{A+}, \tau_{A'}, \tau_\phi, \tau_D) \) be given,

\[
\begin{align*}
\tau_\Sigma &= (\sigma_1 - \sigma_2, \sigma_2 - \sigma_3, \ldots, \sigma_k - \sigma_k) > 0 \\
\tau_{B+} &= (K_1, t_{t_1,1}, \ldots, K_{t_k, t_{t_1,1}}, \ldots, K_{t_k, t_{t_k, t_{t_1,1}}}) > 0 \\
\tau_{B'} &= \text{unrestricted entries } x \text{ in the matrices } K_i \text{ for } i = 1, \ldots, k \text{ and} \\
\tau_{A+} &= (a_{g_{1,i}}, h_{1,i}, \ldots, a_{g_{k,i}}, h_{k,i}, a_{g_{k,a_k}}, h_{k,a_k}, a_{1,2}, \ldots, a_{i_1, n_i - r_1}, \ldots, a_{k, 2}, \ldots, a_{k, n_k - r_k}) > 0 \\
\tau_\phi &= \text{free parameters for the matrices } U_i \in \mathbb{R}^{n \times n} \text{ for } i = 1, \ldots, k \\
\tau_D &= (d_{c,11}, \ldots, d_{c, mn}) \quad (4.73)
\end{align*}
\]

where \( \tau_\Sigma \) consists of the differences of the distinct minimum phase singular values and the rest is given as in theorem (4.6.1). Let \( T_\delta^{(2)} \) be the parameter space containing all vectors \( \tau_\delta^{(2)} \) satisfying the restrictions above and let

\[
\varphi_\delta^{(2)} : T_\delta^{(2)} \to S_m(n) \\
\tau_\delta^{(2)} \mapsto (A(\tau_\delta^{(2)}), B(\tau_\delta^{(2)}), C(\tau_\delta^{(2)}), D(\tau_\delta^{(2)}), K(\tau_\delta^{(2)}))
\]

be the mapping attaching the discrete time system matrices to the parameters as given in the first part of this theorem. Then \( \varphi_\delta^{(2)}(\tau_\delta^{(2)}), \tau_\delta^{(2)} \in T_\delta^{(2)} \) is a discrete time, minimal, stable, strictly minimum phase and (right) minimum phase balanced state-space realization of a transfer function in \( T_\delta^{(2)} \).

Proof. As mentioned above, this canonical form can be constructed from first principles in the same way as has been shown for the Lyapunov balanced case. For a detailed proof, see theorem 4.32 in (McGinnie, 1993), where only the parametrization of the square transfer function \( k_c(s) \) (and \( k(z) \)) is considered and the continuous time \( E_c \)-matrix is not required to be such that the corresponding discrete time matrix becomes the identity. It is easy to see that \( I + C_c(I - A_c)^{-1}K_c \) is indeed invertible and that it is mapped to the identity by the homeomorphism (A.44). Finally, note that \( U_i \in \mathbb{R}^{n \times n} \) and that \( U_i = \begin{bmatrix} \tilde{U}_i & 0 \end{bmatrix} \in \mathbb{R}^{n \times m} \) in theorem 4.32 in (McGinnie, 1993).

Of course, remarks analogous to those in remarks (4.6.3), (4.6.4), (4.6.5) and (4.6.6) below theorem (4.6.1) can also be made at this point.

An illustrative example

Illustrative examples analogous to those in section (4.6) for Ober’s Lyapunov balanced canonical form can also be given at this point.

4.8.2 Topological properties

The following theorem states the main topological properties of McGinnie’s minimum phase balanced state-space parametrization:

Theorem 4.8.2. McGinnie’s (right) minimum phase balanced state-space parametrization
where \( \delta = (\alpha, \beta, \gamma) \) is described in theorem (4.8.1), has the properties (i) - (vii) described in theorem (4.6.2). Clearly, \( V_\delta, T_\delta, \varphi_\delta \) etc. have to be replaced by \( V_\delta^{[2]}, T_\delta^{[2]}, \varphi_\delta^{[2]} \) etc.

**Proof.** The theorem follows from the fact that the construction of this canonical form from first principles is analogous to the construction of Ober's Lyapunov balanced canonical form. \( \square \)

Remarks analogous to those in remarks (4.6.7), (4.6.8) and (4.6.9) below theorem (4.6.2) can also be made at this point. Note that now also the strict minimum phase assumption is built in without any need for additional restrictions on the parameters.

### 4.9 Data driven local coordinates (DDLC)

#### 4.9.1 Data driven local parametrizations

Data driven local coordinates can be considered to be one parametrization out of the class of *data driven local parametrizations*. The philosophy behind the use of data driven local parametrizations in general (and DDLC in particular) could be described as follows:

- **Try to avoid the use of canonical forms in parametrization problems.** One argument for such a viewpoint may be that some canonical forms are rather complicated (see, e.g. Ober’s Lyapunov and stochastically balanced or McGinnis’s minimum phase balanced canonical form as given in sections (4.6), (4.7) and (4.8), respectively), another argument relates to the necessity to additionally estimate some structural indices in order to determine which ”chart” of the canonical form one should use. From a practical point of view, the most relevant argument, however, is the fact that canonical forms are data-independent: The parametrization is ”frozen” a priori and this may potentially (but, of course, not necessarily) lead to numerically ill-conditioned estimation problems.

- **Try to keep the number of parameters to a minimum.** This is done in order to reduce the computational burden on the one hand, and on the other hand to still get an identifiable parametrization. The common feature of data driven local parametrizations is that the number of parameters is kept to a minimum by constructing ”local” parameter spaces (around the current system estimate) which are adapted if another system estimate becomes available.

Of course, there are many ways of constructing such local parametrizations: Assume, for instance, that we are given an atlas of coordinate charts for the manifold \( \mathbb{M}(n) \) where the corresponding coordinate neighborhoods overlap; see, e.g. the overlapping state-space parametrization in section (4.3). Given a transfer function \((l, k)\) in the intersection of (various) coordinate neighborhoods, we are free to choose which chart we want to use for the description of \((l, k)\), and, clearly, this choice can (and should reasonably) be based on data-dependent information, e.g. to make the estimation problem numerically well conditioned. For the overlapping parametrization in section (4.3), such a *structure selection procedure* was indeed proposed e.g. in (Overbeek and Ljung, 1982); see also (Peeters, 1994). We will not further discuss this approach, but treat the DDLC and *structuredDDLC* parametrization and the *orthogonalDDLC* algorithm in the sequel.

In a sense, DDLC uses uncountably many coordinate charts, \( (V_D^{loc}, \psi_D^{loc}) \), say. Nevertheless, the data-driven choice between the coordinate neighborhoods is particularly easy to implement. This choice, despite its intuitively appealing interpretation and its simplicity in implementation, will be shown to be problematic in two respects: On the one hand we will see in this section that it is not easy to guarantee that the particular data-driven choice of coordinate neighborhoods for DDLC makes it possible to reach every transfer function \((l, k)\) in (a connected component of) \( \mathbb{M}(n) \). On the other hand, the choice may
also be problematic from a numerical point of view as will be seen in section (5.6) in chapter 5. As far as orthoDDLc is concerned, we will show that the latter problem can be eliminated. In fact, we also conjecture that the first problem is solved by using orthoDDLc, but we still have no proof for this conjecture.

4.9.2 Introduction

We will now briefly introduce DDLC: As a starting point, consider the full state-space parametrization of section (4.1). As has been shown in lemma (3.2.1), the $(l,k)$-equivalence class $\mathcal{E}(A, B, C, D)$ in $S_m(n)$ is a real analytic manifold of dimension $n^2$. This means that there are $n^2$ essentially unnecessary coordinates when using the full state-space parametrization.

The idea now is to avoid this drawback by only considering the $2ns + m(n + s)$ dimensional ortho-complement $Q_{(A, B, C, D)}$ to the tangent space to $\mathcal{E}(A, B, C, D)$ at a given $(A, B, C, D)$ as a parameter space; see lemma (3.2.3) for a description of $Q_{(A, B, C, D)}$. Here, $(A, B, C, D)$ is obtained by some initial estimate, and this is the reason for calling the parametrization data driven local coordinates\(^6\). Clearly, the parameter space will then be of dimension $2ns + m(n + s)$ rather than $n^2 + 2ns + m(n + s)$ and thus has no unnecessary coordinates.

Note that the tangent space can also be described in terms of the vectorization of the state-space matrices: see (3.8). The matrix $Q \in \mathbb{R}^{n^2+2ns+m(n+s)\times n^2}$ has full column rank $n^2$ for any minimal $(A, B, C, D, K)$. In fact, the rank of $Q$ is less than $n^2$ if and only if the realization $(A, B, C, D, K)$ has a mode that is both uncontrollable and unobservable, i.e. there exists a $\lambda \in \mathbb{C}$ such that both $(\lambda I - A, B)$ and $(\lambda I - A', C')$ are rank deficient; see (McKelvey and Helmersson, 1999). By $Q^\perp$ we denote a matrix the columns of which span the orthogonal complement to the tangent space given above. $Q^\perp$ can be obtained e.g. from a singular value decomposition of $Q$, and the parametrization is then obtained as follows:

**Definition 4.9.1 (Data driven local coordinates (DDLC)).** Let a minimal $(A, B, C, D, K)$ be given. The DDLC are given by the mapping

$$\varphi_D : T_D \rightarrow S_m(n)$$

$$\tau_D \mapsto \begin{pmatrix}
\text{vec}(A(\tau_D)) \\
\text{vec}(B(\tau_D)) \\
\text{vec}(C(\tau_D)) \\
\text{vec}(D(\tau_D))
\end{pmatrix} = \begin{pmatrix}
\text{vec}(A) \\
\text{vec}(B) \\
\text{vec}(C) \\
\text{vec}(D)
\end{pmatrix} + Q^\perp \tau_D$$

Here, $T_D \subset \mathbb{R}^{2ns+m(n+s)}$ denotes the parameter space for DDLC, i.e. the set of all $\tau_D \in \mathbb{R}^{2ns+m(n+s)}$ such that $\varphi_D(\tau_D)$ is minimal. Let $\tilde{V}_D = \pi(\varphi_D(T_D))$.

**Remark 4.9.1.** For any fixed minimal $(A, B, C, D, K)$, the mapping $\varphi_D$ from the parameter vectors $\tau_D$ to the state-space matrices is affine (and therefore continuous and analytic) as can be seen from (4.74). Clearly, $\varphi_D(T_D) = \varphi_D(T\tilde{D})$, and in the sequel we will use the symbol $\pi(T_D) = \pi(\varphi_D(T_D))$ with slight abuse of notation.

4.9.3 An illustrative example

Before we state the main theorem concerning topological and geometrical properties of DDLC, we want to discuss a special case. This is done in order to motivate and (hopefully) clarify the results of the main theorem in the next section.

We consider the case where $n = s = 1$ and $m = 0$: no exogenous inputs are present and thus $I(z)$ vanishes. Clearly, $S(1) = \mathbb{R}^3$ and $S_m(1) = \{(a, k, c) \in \mathbb{R}^3 | (a, k, c)$ is minimal $\}$; see the illustrative example in section (4.1).

\(^6\)The change of the parameter spaces is, of course, only determined by the structure of the equivalence classes $\mathcal{E}(A, B, C, D)$, but the "sequence of initial estimates" $(A, B, C, D)$ is obtained in the course of the (data-driven) iterative search algorithm.
Let the minimal system \((a,k,c) \in S_m(1)\) be given. Then we have

\[
Q = \begin{pmatrix}
  0 \\
  k \\
  -c
\end{pmatrix}
\]

and thus we can choose \(Q^\perp = \begin{pmatrix}
  1 & 0 & 0 \\
  0 & k & \sqrt{k+c} \\
  0 & \sqrt{k+c} & k+c
\end{pmatrix}\)

Using this \(Q^\perp\), the DDLC parametrization is given by (note that \(\tau_D^i\) denotes the \(i\) component of the vector \(\tau_D\))

\[
\begin{pmatrix}
  a(\tau_D^1, \tau_D^2) \\
  k(\tau_D^1, \tau_D^2) \\
  c(\tau_D^1, \tau_D^2)
\end{pmatrix} = \begin{pmatrix}
  a \\
  k \\
  c
\end{pmatrix} + \begin{pmatrix}
  1 & 0 & 0 \\
  0 & \frac{c}{\sqrt{k+c}} & 0 \\
  0 & 0 & \frac{k}{\sqrt{k+c}}
\end{pmatrix}\begin{pmatrix}
  \tau_D^1 \\
  \tau_D^2
\end{pmatrix}
\]

(4.75)

Here, \(T_D = \{(\tau_D^1, \tau_D^2) \in \mathbb{R}^2 | (a(\tau_D^1, \tau_D^2), k(\tau_D^1, \tau_D^2), c(\tau_D^1, \tau_D^2))\) is minimal \} and \(\varphi_D(T_D)\) is a subset of the whole affine plane in \(S(1)\) given by \((a, k, c) = (a, k, \pm k)\), \(\varphi_D(T_D)\) becomes a subset of the whole plane given by \(k = \pm c\). In the sequel, we will call the affine subspace (plane) containing \(\varphi_D(T_D)\) the affine subspace (plane) corresponding to \(T_D\).

Making use of figure (4.3) in section (4.13), we will now discuss geometrical and topological properties of DDLC:

(i) The affine subspace corresponding to \(T_D\) intersects the planes given by \(k = 0\) and \(c = 0\) yielding two straight lines: \(T_D = \mathbb{R}^2 \setminus \{(x, -\frac{1}{2}\sqrt{k^2 + c^2}), x \in \mathbb{R}\} \cup \{(x, -\frac{1}{2}\sqrt{k^2 + c^2}), x \in \mathbb{R}\}. \) In case of an initial system of the form \((a, k, \pm k)\), this intersection becomes the \(a\)-axis only: \(T_D = \mathbb{R}^2 \setminus \{(x, +\sqrt{2}k), x \in \mathbb{R}\}. \) In any case, \(T_D\) is seen to be an open and dense subset of \(\mathbb{R}^2\).

(ii) There exists a neighborhood \(T_D^0\) of \((0, 0) \in T_D\) containing the initial system \((a, k, c)\), such that each hyperbola entering a (sufficiently small) neighborhood of \((a, k, c)\) in \(\mathbb{R}^2\) also intersects the affine plane corresponding to \(T_D\) and the intersection yields one single point in \(T_D^0\). Moreover, no non minimal system is described in \(T_D^0\).

(iii) The boundary points of \(T_D\) (which do not belong to \(T_D\)) represent the trivial transfer function \(k(z) = 1\): \(\pi(\bar{T}_D)\) contains lower degree transfer functions.

(iv) Within \(T_D\), the \(k\)-equivalence classes consist of two elements except for the points where \(T_D\) touches a hyperbola which gives a singleton. These touching points – see figure (4.3) in section (4.13) – constitute a straight line and are given by:

\[
\begin{pmatrix}
  a(\xi) \\
  k(\xi) \\
  c(\xi)
\end{pmatrix} = \begin{pmatrix}
  a \\
  k \\
  c
\end{pmatrix} + \begin{pmatrix}
  1 & 0 & 0 \\
  0 & 1 & 0 \\
  0 & 0 & 1
\end{pmatrix}\cdot \begin{pmatrix}
  \xi \\
  \xi^2 + k^2
\end{pmatrix}, \quad \xi \in \mathbb{R}
\]

(4.76)

The touching points do not occur if the initial system is of the form \((a, k, \pm k)\). Note that in this case the DDLC parametrization is locally identifiable at any point \(\tau_D \in T_D\), but not globally identifiable.

In any case, we have a lack of global identifiability of \(T_D\), but every transfer function in \(V_D\) has just a finite number of representations within \(T_D\).

(v) \(V_D\) is clearly open in \(\pi(\bar{T}_D)\). However, note that \(V_D\) is not necessarily open in \(\mathbb{M}(1)\): The transfer functions corresponding to the points where hyperbolae touch the affine plane corresponding to \(T_D\) are boundary points of \(V_D\) which belong to \(V_D\). This is a direct consequence of the fact that in any neighborhood of such a touching point there are hyperbolae which do not touch or intersect the affine plane corresponding to \(T_D\). Hence, we can find a sequence of minimal realizations \((a_n, k_n, c_n) \notin \varphi_D(T_D), n \in \mathbb{N}\) converging to the touching point and continuity of \(\pi\) shows the statement. Note that an inner point of \(T_D\) in this case corresponds to a boundary points of \(V_D\) and that at these points local identifiability is also violated.

(vi) We now discuss a case where \(\pi(\bar{T}_D) = \tilde{V}_D\) and a case where \(\pi(\bar{T}_D) \subset \tilde{V}_D\) where the inclusion is strict:
(a) First, we consider the special case depicted in figure (4.3) in section (4.13) where \((a,k,c) = (0,3,1)\). It is straightforward to see that the Markov parameters corresponding to the parametrized transfer functions are given by

\[
(K_j)_{j \in \mathbb{N}} = \left(1 + \frac{1}{10} \tau_D^2 (1 + \frac{3}{10} \tau_D^2), \tau_D (1 + \frac{1}{10} \tau_D^2 (1 + \frac{3}{10} \tau_D^2)) \right)
\]  

(4.77)

and that \(T_D = \mathbb{R}^2 \setminus \{(x, -3\sqrt{10}), x \in \mathbb{R}\} \cup \{(x, -\sqrt{10}/3), x \in \mathbb{R}\}\). Clearly,

\[
\pi(\mathcal{T}_D) = \pi(\mathbb{R}^2) = V_D \cup \mathcal{M}(0)
\]

where \(\mathcal{M}(0)\) only contains the trivial transfer function \(k(z) = 1\) of McMillan degree zero. We have to show that \(\pi(\mathcal{T}_D) = \mathring{V}_D\), i.e. there must not exist a sequence \((\tau_{D,1}, \tau_{D,2})\) with the property that the corresponding sequence of transfer functions converges in \(T_{pl}\) (see (4.77) for the sequence of Markov parameters) and the limiting transfer function has no representation within \(\pi(\mathcal{T}_D)\). The only "candidate sequences" in this respect are sequences where at least one parameter becomes unbounded. If, however, \(\tau_D^2 \rightarrow \pm \infty\), then the second Markov parameter \(K_1\) in (4.77) goes to infinity, i.e. the corresponding sequence of transfer functions diverges in \(T_{pl}\). This is independent of the choice of \(\tau_{D,1}\). On the other hand, if \(\tau_D^2 \rightarrow \pm \infty\), then for the \((i+2)\) Markov parameter to remain bounded we must have \((1 + \frac{1}{10} \tau_D^2 (1 + \frac{3}{10} \tau_D^2)) = O((\tau_D^2)^{-i})\). This, however, implies that the sequence of Markov parameters must converge to \((1,0,0,\ldots)\), i.e. the limiting transfer function is the trivial \(k(z) = 1\) which is obtained in \(\pi(\mathcal{T}_D)\) anyway.

(b) In order to show that \(\pi(\mathcal{T}_D) \neq \mathring{V}_D\) is also possible, we consider an example where \(n = 1, s = 2\) and \(m = 0\):

\[
(a,K,C) = \left(1, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)
\]

Then we have

\[
Q = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}
\]

and thus we can choose

\[
Q^\perp = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}
\]

Note that the columns of \(Q^\perp\) are not chosen to be orthogonal in this case, which, however, does not affect the subsequent analysis. The DDLC parametrization is given by

\[
\begin{pmatrix}
vec(a(\mathcal{T}_D)) \\
vec(K(\mathcal{T}_D)) \\
vec(C(\mathcal{T}_D))
\end{pmatrix} = \begin{pmatrix}
1 + \tau_D^2 \\
1 + \tau_D^2 \\
1 + \tau_D^2 + \tau_D^4 \\
-1 + \tau_D^2
\end{pmatrix}
\]

(4.78)

Note that the Markov parameters are then easily obtained:

\[
(K_j)_{j \in \mathbb{N}} = \left(\frac{1 + \tau_D^2 + \tau_D^4}{(1 + \tau_D^2)(1 + \tau_D^2), (1 + \tau_D^2)\tau_D^4}{(1 + \tau_D^2)(1 + \tau_D^2), (1 + \tau_D^2)\tau_D^4}{(1 + \tau_D^2)(1 + \tau_D^2), (1 + \tau_D^2)\tau_D^4}{(1 + \tau_D^2)(1 + \tau_D^2), (1 + \tau_D^2)\tau_D^4}\right)
\]

(4.77)

It is easy to see that \(T_D = \mathbb{R}^4 \setminus \{(x,y,0,-1), x \in \mathbb{R}, y \in \mathbb{R}\} \cup \{(x,1,y,-2), x \in \mathbb{R}, y \in \mathbb{R}\}\). Clearly,

\[
\pi(\mathcal{T}_D) = \pi(\mathbb{R}^4) = V_D \cup \mathcal{M}(0)
\]

where \(\mathcal{M}(0)\) contains only the trivial transfer function \(k(z) = 1\) of degree zero. However, in this case, \(\mathring{V}_D\) contains an additional transfer function of degree one. To see this, consider a sequence \((\tau_{D,t,1}, \tau_{D,t,2}, \tau_{D,t,3}, \tau_{D,t,4}) = (0,\pm t, \pm \frac{1}{t} + 1), t \in \mathbb{N}\). Then we have...
(a(\tau_D,t), K(\tau_D,t), C(\tau_D,t)) \rightarrow \left( 1, \begin{pmatrix} 0 & 0 \\ 0 & \pm \infty \end{pmatrix} \right)

but in terms of the Markov parameters we have

\[(K_j)_{j \in \mathbb{N}} \rightarrow \left( \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \ldots \right)\]

We have convergence in \( T_{pl} \) to a degree one transfer function in \( \bar{V}_D \) which cannot be represented in \( \pi(T_D) \).

To sum it up, it is possible that the same transfer functions are described in the closure of the parameter space \( T_D \) and in the closure of the corresponding transfer function space \( V_D \), but it is also possible that the closure of the transfer function space contains more transfer functions than can be represented in the closure of the parameter space.

**Remark 4.9.2.** In the example considered in (4.75), iterating the DLDC construction commencing from an initial \((a,k,\pm k)\), e.g. in course of a gradient-type search procedure) leaves \( V_D \) unchanged. This is because for any minimal system in the affine plane corresponding to \( T_D \), the ortho-complement to its equivalence class coincides with the "old" affine plane. In other words: We cannot reach transfer functions outside the "first" \( V_D \) by an iteration of the DLDC construction, where iteration means the outlined procedure of constructing a new DLDC parameter space at some estimated system within the "old" \( T_D \). Note that the sets \( V_D \) corresponding to initial systems of the form \((a,k,k)\) and \((a,k,-k)\), respectively, coincide with the two connected components of \( \mathbb{M}(1) \) in this case; see statement (iii) in theorem (4.3.2).

If, however, we start from an initial \((a,k,c)\) where \(|k| \neq |c|\), then, by choosing an appropriate parameter vector \( \tau_D \in T_D \), say, and by applying the DLDC construction a second time at the system \((a(\tau_D), k(\tau_D), c(\tau_D))\), we can reach any hyperbola in the affine plane corresponding to the new DLDC parameter space which has not been reached in the affine plane corresponding to the old parameter space.

Summing up, we see that in this special case, we can guarantee to reach any transfer function \( k(z) \) in the same pathwise connected component of \( \mathbb{M}(1) \) as the initial \( \pi(a,k,c) \) by only one iteration of the DLDC construction. Recall that \( \mathbb{M}(n) \) consists of \( n + 1 \) disconnected components in the single-input, single-output case; see statement (ii) in theorem (4.3.2). It is easy to see that the two components in the special case considered here correspond to the first and third and the second an fourth quadrant in plot (b) of figure (4.3) in section (4.13), respectively.

**Remark 4.9.3.** It is evident that the affine plane corresponding to any \( T_D \) in (4.75) does not intersect "almost all" hyperboles. Hence, the corresponding \( V_D \) "leaves out" more than just a thin subset of \( \mathbb{M}(1) \), i.e. \( \mathbb{M}(1) \setminus V_D \) always contains an open set: No \( V_D \) is generic in \( \mathbb{M}(1) \) in the example considered in (4.75) above.

**Remark 4.9.4.** Note that \( V_D \) is strongly dependent on the particular initial system in the equivalence class where the DLDC construction is performed; \( V_D \) is smallest for the case of choosing a representative of the form \((a,k,\pm k)\) because we miss all hyperboles corresponding to systems \((a,k,c)\) with the opposite sign of \( kc \) in this case.

### 4.9.4 Topological and geometrical properties

Let us start with a definition of **local identifiability**.

**Definition 4.9.2 (Local identifiability of a state-space parametrization).** Let \( T_g \subset \mathbb{R}^d \) denote an open subset of \( \mathbb{R}^d \) and let \( \varphi_g : T_g \rightarrow S(\mathbb{R}) \) denote the mapping attaching to every parameter vector \( \tau_g \in T_g \) the corresponding state-space system \((A(\tau_g), B(\tau_g), C(\tau_g), D(\tau_g), K(\tau_g))\). Then the state-space parametrization is said to be **locally identifiable at the point** \( \tau_0^g \) if there exists an \( \varepsilon > 0 \) such that

(i) \( \|\tau_1^g - \tau_0^g\| < \varepsilon \), \( \|\tau_2^g - \tau_0^g\| < \varepsilon \) and
(ii) \( \pi(\varphi_g(\tau_1^g)) = \pi(\varphi_g(\tau_2^g)) \)

imply that \( \tau_1^g = \tau_2^g \).
This definition means that there exists an open neighborhood of the parameter vector $\tau_0^D$ such that any two distinct parameter vectors in this neighborhood correspond to distinct transfer functions.

As already mentioned, DDLC was proposed in (McKelvey and Helmersson, 1999). In (Ribarits et al., 2002), (Deistler and Ribarits, 2001) and (Ribarits and Deistler, 2002) it has been further investigated, and many of the results of the following theorem, which holds true for arbitrary $n$, $s$ and $m$, can be found in the latter references:

**Theorem 4.9.1.** Assume that the initial system $(A, B, C, D, K)$ is minimal. The parametrization by DDLC as given in (4.74) has the following properties:

(i) $T_D$ is an open and dense subset of $\mathbb{R}^{2n+pm(n+s)}$.

(ii) There exist open neighborhoods $T_D^{loc}$ of $0 \in T_D$ and $V_D^{loc}$ of $\pi(A, B, C, D, K)$ in $\mathbb{M}(n)$ such that $T_D^{loc}$ is identifiable, $V_D^{loc} = \pi(T_D^{loc})$ and the mapping $\psi_D^{loc} : V_D^{loc} \to T_D^{loc}$ defined by $\psi_D^{loc}(\pi(\tau_D)) = \tau_D$ is a homeomorphism.

(iii) $\pi(T_D)$ contains transfer functions of lower McMillan degree.

(iv) For "almost every" $(l, k) \in V_D$, the $(l, k)$-equivalence class in $T_D$ consists of a finite number of isolated points.

(v) $V_D$ is open (and trivially dense) in $\pi(T_D)$, but not necessarily open in $\mathbb{M}(n)$.

(vi) $\pi(T_D) \subseteq V_D$ where equality can hold, but the inclusion can also be strict.

**Remark 4.9.5.** We give a short discussion of these results:

(i) Again, this means that the parameters are really free and that almost any point in $\mathbb{R}^{2n+pm(n+s)}$ corresponds to a transfer function in $\mathbb{M}(n)$.

(ii) assures local bijectivity and continuity (in both directions) of the parametrization in the open neighborhood $V_D^{loc}$. As consistency is preserved under continuous transformations, consistency in terms of transfer functions (see section (2.3.4)) implies consistency in terms of the parameter estimates: The estimation problem is locally well posed.

(iii) The closure of the parameter space $T_D$ – note that $T_D = \mathbb{R}^{2n+pm(n+s)}$ by (i) – corresponds to transfer functions of equal and lower McMillan degrees. Note that we always have $\bar{T}_D \cap T_D \neq \emptyset$ and that the $(l, k)$-equivalence classes in $T_D \cap T_D$ are generally given by nonlinear restrictions and are thus difficult to describe.

(iv) This means that at least for almost any transfer function in $V_D$, there are only finitely many isolated points in $T_D$ representing this transfer function. Note that this is again important for gradient type algorithms and for asymptotic theory.

(v) This is a rather technical result which can again be important in connection with coordinate free consistency as discussed in section (2.3.4); cf. (v) in remark (4.1.2).

(vi) The fact that $\bar{V}_D$ may contain more transfer functions than can be described in the closure of the parameter space $T_D$ can clearly effect the actual estimation procedure. As has been discussed in the illustrative examples above, in that case the parameter vector diverges to infinity in spite of the fact that the corresponding sequence of transfer functions converges to a well defined transfer function estimate.

**Remark 4.9.6.** The conjecture that equality holds in the single output case in statement (vi), whereas the inclusion is strict in the case $s > 1$, is still left for future research.

**Remark 4.9.7.** In analogy to remark (4.9.2), one could ask whether it is possible in the general case that any transfer function within the same connected component as the initial transfer function can be reached by iterating the DDLC construction finitely many times. This is still an open question, though there will be some further related discussion in chapter 5.

**Remark 4.9.8.** The conjecture that $V_D$ is never generic in $\mathbb{M}(n)$ – see remark (4.9.3) for a simple example – is still to be proved.

Before theorem 4.9.1 is proved, a result from (Glover and Willems, 1974) is stated and a few remarks are given:
**Theorem 4.9.2.** Let $T_g \subset \mathbb{R}^d$ denote an open subset of $\mathbb{R}^d$ and let $\varphi_g: T_g \to S_m(n)$ be a continuously differentiable mapping attaching (the vectorization of the minimal) system $(A(\tau_g), B(\tau_g), C(\tau_g), D(\tau_g))$ to $\tau_g \in T_g$. Then the state-space parametrization

(i) is locally identifiable at $\tau_g = \tau_g^0$ if and only if

$$ F: GL(n) \times T_g \to S_m(n) $$

$$(T, \tau_g) \mapsto \begin{pmatrix} \text{vec}(TA(\tau_g)T^{-1}) \\ \text{vec}(TB(\tau_g)) \\ \text{vec}(C(\tau_g)T^{-1}) \\ \text{vec}(D(\tau_g)) \end{pmatrix} = T_{vec} \begin{pmatrix} \text{vec}(A(\tau_g)) \\ \text{vec}(B(\tau_g)) \\ \text{vec}(C(\tau_g)) \\ \text{vec}(D(\tau_g)) \end{pmatrix} $$

is locally injective at $(T, \tau_g) = (I, \tau_g^0)$; $T_{vec}$ is given in (3.4).

(ii) If the rank of $X(\tau_g)$ equals $r$ for all $\tau_g \in U(\tau_g^0)$, where $X(\tau_g) = \begin{pmatrix} \frac{\partial F}{\partial T}(I, \tau_g); \frac{\partial F}{\partial \tau_g}(I, \tau_g) \end{pmatrix}$ and $U(\tau_g^0)$ is some open neighborhood of $\tau_g^0$ in $T_g$, then the state-space parametrization is locally identifiable at the point $\tau_g = \tau_g^0$ if and only if $r = n^2 + d$, or, equivalently, if and only if $\det(X(\tau_g^0)X(\tau_g^0)) \neq 0$.

**Proof.** (i): "⇒" The fact that local identifiability translates to local injectivity of $F$ is clear: In a suitably chosen open neighborhood $O$ of $\varphi_g^0$ all parameters correspond to minimal systems ($\varphi_g$ is continuous and $S_m(n)$ is open in $S(n)$ by statement (i) in theorem (4.1.2)). Therefore, all observationally equivalent systems are related by the similarity transformation in (3.1). From the proof of lemma (3.2.1) we know that the mapping $\phi$ in (3.2) is a homeomorphism, in particular it is injective: Different $T \in GL(n)$ yield different points in $S_m(n) \subset S(n)$. Hence, local identifiability of the state-space parametrization implies the existence of an open set $T_{loc} \subset O$ such that $F|_{GL(n) \times T_{loc}}$ is injective. In particular, this also means that $F|_{O_T \times T_{loc}}$ is injective where $O_T$ from now on always denotes some open neighborhood of the identity matrix $I_n$ in $GL(n)$.

(i): "⇐" Let us assume that the state-space parametrization were not locally identifiable at $\tau_g = \tau_g^0$. Taking $\tau_g,i,1 = 1,2$ with $\tau_g,1 \neq \tau_g,2$ in an arbitrary small neighborhood $T_{loc} \subset T_g$ of $\tau_g^0$ with $\pi(\varphi_g(\tau_g,1)) = \pi(\varphi_g(\tau_g,2))$, one can easily calculate the unique state transformation $T(\tau_g,1, \tau_g,2) \in GL(n)$ depending continuously on $\tau_g,i,1 = 1,2$ and satisfying $T(\tau_g,1, \tau_g,2) \to I$ for $\tau_g,1 \to \tau_g,2$ such that $F(T(\tau_g,1, \tau_g,2), \tau_g,1) = F(I, \tau_g,2)$. Thus, $F$ restricted to $O_{T} \times T_{loc}$ cannot be injective.

(ii) Clearly, $n^2 + d \leq n^2 + 2ns + m(n + s)$ has to be fulfilled for local identifiability to hold. Under the assumption that the rank of $\frac{dF}{dT}(T, \tau_g)$ is constant in some neighborhood $O_T \times U(\tau_g^0)$, it is clear that such an $F$ is locally injective (i.e. $F|_{O_T \times T_{loc}}$ is injective with $O_T \times T_{loc} \subset O_T \times U(\tau_g^0)$) if and only if the rank of this matrix is equal to $n^2 + d$. This is because $F$ can be locally approximated by the linear function $DF$ the injectivity of which is directly determined by the Jacobian. However, note that the only if part is due to the constant rank assumption which excludes cases like $f(x) = x^3$ in a neighborhood of $x = 0$, for instance. What remains to be shown is that $rk(\frac{dF}{dT}(T, \tau_g)) = \frac{dF}{dT}(I, \tau_g) = rk(X(\tau_g))$ for all $T \in GL(n)$. To see this, first note that $dF(T, \tau_g)$ describes the tangent space to the equivalence class $E(A(\tau_g), B(\tau_g), C(\tau_g), D(\tau_g))$ and has already been computed in (3.7). Hence, we get

$$ \frac{dF}{dT}(T, \tau_g) = T_{vec} \cdot \begin{pmatrix} A(\tau_g)' \otimes I_n - I_n \otimes A(\tau_g) \\ B(\tau_g)' \otimes I_n - I_n \otimes C(\tau_g) \\ 0_{sm \times n^2} \end{pmatrix} \begin{pmatrix} \text{vec}(A(\tau_g)) \\ \text{vec}(B(\tau_g)) \\ \text{vec}(C(\tau_g)) \end{pmatrix} $$

$$ = T_{vec} \begin{pmatrix} \text{vec}(A(\tau_g)) \\ \text{vec}(B(\tau_g)) \\ \text{vec}(C(\tau_g)) \end{pmatrix} \cdot \begin{pmatrix} I_n \otimes T^{-1} \\ 0 \\ I_d \end{pmatrix} $$

(4.80)

and it is evident that $rk(\frac{dF}{dT}(T, \tau_g)) = rk(X(\tau_g))$ for all $T \in GL(n)$. \qed
We continue with a number of remarks:

**Remark 4.9.9.** Note that the theorem is also applicable for parametrizations resulting from some a priori system knowledge so that \( d \) need not be \( 2ns + m(n + s) \). However, for \( Tg = TD \) (and \( \tau_g = \tau_D^0 \)), \( X(\tau_D^0) \) will itself be a square matrix.

**Remark 4.9.10.** The theorem also deals with parametrizations where the parameter space \( Tg \) cannot be identified with an affine subspace in \( S(n) \). Balanced canonical forms such as Ober’s stochastically balanced and McGinie’s minimum phase balanced parametrizations correspond to **curved manifolds** even in the simplest case \( n = s = 1 \) and \( m = 0 \) and may therefore serve as an example at this point; see figure (4.2) in section (4.13).

**Remark 4.9.11.** The **constant and full rank assumption** on \( X(\tau_g) \) (and, equivalently, on \( \frac{dF}{dt(\tau_g)}(T, \tau_g) \)) has a nice general interpretation in case of \( d = 2ns + m(n + s) \). As long as it is valid, for any fixed \( F(T, \tau_g) \), there will be no direction in the tangent space to the parametrized manifold in \( S(n) \) (the structure of an affine subspace like \( TD \) is a special case) that locally coincides with any direction in the tangent space to the equivalence class \( E(\tau_g), B(\tau_g), C(\tau_g), D(\tau_g) \). Moreover, \( (T, \tau_g) \) can serve as a **local coordinate system** in \( S(n) \) around \( \varphi_g(\tau_g) \), i.e. \( F \) is a local homeomorphism: Note that \( O = GL(n) \times U(\mathfrak{n}^0) \) is an open subset of \( \mathbb{R}^{n^2 + 2n + m(n + s)} \) and \( F : O \rightarrow \mathbb{R}^{n^2 + 2n + m(n + s)} \) is continuously differentiable in \( O \) with nonsingular Jacobian. By the inverse function theorem this implies **local invertibility** of \( F \) around each \( (T, \tau_g) \in O \), i.e. bijectivity of \( F|_{O_1(T, \tau_g)} : O_1(T, \tau_g) \rightarrow O_2(F(T, \tau_g)) \) where \( O_1(T, \tau_g) \subseteq O \) and \( O_2(F(T, \tau_g)) \) are open neighborhoods (in \( \mathbb{R}^{n^2 + 2n + m(n + s)} \)) of \( (T, \tau_g) \) and \( F(T, \tau_g) \), respectively. Moreover, \( F^{-1}|_{O_2(F(T, \tau_g))} : O_2(F(T, \tau_g)) \rightarrow O_1(T, \tau_g) \) is also continuously differentiable, turning \( F|_{O_1(T, \tau_g)} \) into a \( C^1 \) diffeomorphism and thus trivially into a homeomorphism. Note the local aspect of this result: \( F \rceil_O \) need not be injective although the Jacobian is nonsingular for all \( (T, \tau_g) \in O \).

**Remark 4.9.12.** Theorem 4.9.2 also gives some information about the structure of the equivalence classes. If the **constant rank assumption** in (ii) in the theorem above holds true and \( r < n^2 + d \), then one cannot have local identifiability and thus certain shapes of the equivalence classes can be excluded a priori.

**Remark 4.9.13.** Assume that \( \varphi_g : T_g \rightarrow S_m(n) \) is a polynomial in the entries of \( \tau_g \) (in case of DDLC it would be linear in the entries of \( \tau_D \)). As \( F \) is rational in \( T \), it is an analytic function. In fact, it is an analytic mapping between **real analytic manifolds**: Its domain of definition is an open subset of \( \mathbb{R}^{n \times n} \times \mathbb{R}^d \equiv \mathbb{R}^{n^2 + d} \) because the set \( GL(n) \) is open (and dense) in \( \mathbb{R}^{n \times n} \) and \( T_g \) is open in \( \mathbb{R}^d \). The image space \( S_m(n) \) is also open (and dense) in \( S(n) = \mathbb{R}^{n^2 + 2n + m(n + s)} \). Trivially, open subsets of Euclidean spaces are real analytic manifolds. Note, however, that not every point in \( S_m(n) \) need to be an image point of some \( (T, \tau_g) \) and that the **image set** \( F(GL(n) \times T_g) \) need not be open in \( S_m(n) \) (see figure (4.6) for the case of DDLC) where \( n = s = 1 \) and \( m = 0 \).

Let us call a point \((\tilde{T}, \tilde{\tau}_g) \in GL(n) \times T_g \) a **regular point** if the Jacobian \( \frac{dF}{d(\tilde{T}, \tilde{\tau}_g)}(\tilde{T}, \tilde{\tau}_g) \) in (4.80) has full rank. We will call \( b = (\text{vec}(A)', \text{vec}(B)', \text{vec}(C)', \text{vec}(D)')' \in S_m(n) \) a **regular value** of \( F \) if every point in the set \( F^{-1}(b) \) is regular (and, in particular, if \( F^{-1}(b) = \emptyset \)). Non regular points and non regular values are said to be singular points and singular values, respectively.

We will also need a few basics in real algebraic geometry:

**Definition 4.9.3 (Algebraic and Semi-algebraic sets).** An **algebraic subset** of \( \mathbb{R}^d \) is the set of zeros of some polynomial set \( B \subseteq \mathbb{R}[x_1, \ldots, x_d] \), where \( \mathbb{R}[x_1, \ldots, x_d] \) denotes the polynomial ring with coefficients in \( \mathbb{R} \). **Semi-algebraic subsets** of \( \mathbb{R}^d \) are subsets of \( \mathbb{R}^d \) of the form \( \bigcup_{i=1}^k \cap_{j=1}^m \{ x \in \mathbb{R}^d | f_{i,j} \ast_{i,j} 0 \} \) where \( f_{i,j} \in \mathbb{R}[x_1, \ldots, x_d] \) and \( \ast_{i,j} \) is either < or =, for all \( i, j \).

Clearly, every algebraic subset is also semi-algebraic. The following result can be found in (Bochnak et al., 1998):

**Theorem 4.9.3.** Every semi-algebraic subset of \( \mathbb{R}^d \) is the disjoint union of a finite number of semi-algebraic sets, each of them being semi-algebraically homeomorphic to an open hypercube \( (0,1)^l \subset \mathbb{R}^l \) for some \( l \in \mathbb{N} \) (with \( (0,1)^0 \) being a point).

**Proof.** See theorem 2.3.6 in (Bochnak et al., 1998).
The dimension \( \text{dim}(A) \) of a semi-algebraic set \( A \) can be defined algebraically. From the theorem above we know that each semi-algebraic subset \( A \) of \( \mathbb{R}^d \) can be written as \( A = \bigcup_{i=1}^p A_i \) where each \( A_i \) is semi-algebraically homeomorphic to an open hypercube \((0,1)^d_i\). It can be shown that the dimension \( \text{dim}(A) \) coincides with \( \max\{l_1,\ldots,l_p\} \); see corollary 2.8.9 in (Bochnak et al., 1998). Moreover, if the semi-algebraic set is a real analytic submanifold of \( \mathbb{R}^d \) of dimension \( r \) (with the usual definition of dimension for real analytic manifolds), then the two dimension concepts agree: \( \text{dim}(A) = r \); see proposition 2.8.14 in (Bochnak et al., 1998).

We are now ready to prove theorem (4.9.1):

**Proof.** (i) We consider the mapping \( \Delta: \mathbb{R}^{n+m(n+s)} \to \mathbb{R} \) attaching \( \det(W_0^n(\tau D)W_c^n(\tau D)) \) to \( \tau D \) where \( W_0^n(\tau D) = O_n(\tau D)^tO_n(\tau D) \in \mathbb{R}^{n \times n} \) with

\[
O_n(\tau D) = O_n(A(\tau D), B(\tau D), C(\tau D), D(\tau D), K(\tau D)) = O_n(\varphi D(\tau D))
\]

being the corresponding (finite) observability matrix; see definition (1.4.2). The matrix \( W_c^n(\tau D) \) is given analogously. Note that \( W_0^n(\tau D) \) and \( W_c^n(\tau D) \) have full rank if and only if \( \varphi D(\tau D) \) is a minimal state-space realization. Due to the fact that \( \varphi D \) is affine, the determinant of \( W_0^n(\tau D)W_c^n(\tau D) \) is a polynomial in the parameters \( \tau D_i \), \( i = 1,\ldots, 2ns + m(n + s) \) and thus analytic (and trivially continuous). Openness of \( TD \) in \( \mathbb{R}^{2ns+m|n+s|} \) follows from the fact that \( TD = \Delta^{-1}(\mathbb{R} \setminus \{0\}) \) is the inverse image of an open set and \( \Delta \) is continuous. Denseness of \( TD \) in \( \mathbb{R}^{2ns+m|n+s|} \) follows from a well known result for analytic functions: \( \Delta(\tau D) = 0 \) can only hold true on a thin subset of \( \mathbb{R}^{2ns+m|n+s|} \) (\( \Delta \) cannot vanish everywhere in \( \mathbb{R}^{2ns+m|n+s|} \)).

(ii) We show that for a sufficiently small open neighborhood \( T_D^{\text{loc}} \) of \( 0 \in TD \) the corresponding \( \pi(T_D^{\text{loc}}) = V_D^{\text{loc}} \) is a nonvoid open subset of \( M(n) \) and the mapping \( \pi|_{T_D^{\text{loc}}} \) is a homeomorphism. This implies, of course, that \( \psi_D^{\text{loc}} = (\pi|_{T_D^{\text{loc}}})^{-1} \) is a homeomorphism from \( V_D^{\text{loc}} \) onto \( T_D^{\text{loc}} \), too.

First, the rank condition for \( X(\tau D) \) of theorem 4.9.2 (ii) is verified for \( \theta D_0 = 0 \). The first part \( \text{det}(I, \tau D) \) in (4.80) is independent of the parametrization and the second part \( \text{det}(I, \tau D) \) is particularly simple for DDLC because \( \varphi D \) is affine. Thus,

\[
X(\tau D) = \begin{bmatrix}
A(\tau D)^t \otimes I_n - I_n \otimes A(\tau D) \\
B(\tau D)^t \otimes I_n - I_n \otimes C(\tau D) \\
-I_n \otimes I_n \\
0_{n \times n^2} \\
Q(\tau D)
\end{bmatrix}
\]

and for \( \tau D = \tau D_0 = 0 \), one gets \( X(0) = [\dot{Q}; Q^\perp] \). Moreover, \( X(\tau D) \) has constant rank \( n^2 + 2ns + m(n + s) \) in a neighborhood \( O \) of \( \tau D_0 = 0 \) because \( X(\tau D) \) depends continuously on \( \tau D \) and the determinant is a continuous function of the entries in \( X(\tau D) \). This shows local identifiability at \( \tau D_0 = 0 \) by theorem (4.9.2), i.e. injectivity of the function \( \pi|_{T_D^{\text{loc}}} = \pi \circ F|_{I \times T_D^{\text{loc}}} \) where \( T_D^{\text{loc}} \subseteq O \) is chosen appropriately. Note that \( \pi \circ F|_{I \times T_D^{\text{loc}}} \) is clearly continuous because \( \pi \) and \( \varphi D \) are continuous. Set \( \pi(T_D^{\text{loc}}) = V_D^{\text{loc}} \).

Then \( \pi|_{T_D^{\text{loc}}} : T_D^{\text{loc}} \to V_D^{\text{loc}} \) is surjective by definition, it is injective by what was said above, and thus it is bijective and continuous. Note that \( \pi \circ F|_{I \times T_D^{\text{loc}}} \) is seen to be an open mapping because \( \pi|_{GL(n) \times T_D^{\text{loc}}} : GL(n) \times T_D^{\text{loc}} \to O_2(F(I, 0)) \) is a diffeomorphism and hence trivially an open mapping by remark (4.9.11) above and \( \pi \) is open by statement (ii) in theorem (4.1.2). This shows that \( V_D^{\text{loc}} \) is open in \( M(n) \) and that \( \pi|_{T_D^{\text{loc}}} \) is a homeomorphism.

(iii) Let us consider the \( n^2 + ns + m(n + s) \) dimensional linear subspace of \( S(n) \) given by \( C = 0 \) which will be denoted by \( T_0 \) in the sequel. This subspace is a subset of the \((l,k)\)-equivalence class corresponding to the trivial transfer function \((l,k) = (D,I)\), and we will show that it always has a nonvoid intersection with the \( 2ns + m(n + s) \) dimensional affine subspace of \( S(n) \) corresponding to the parameter space \( TD \); with slight abuse of notation, this affine subspace will also be denoted by \( TD \) in the rest of the proof of statement (iii).
Note that \( \dim T_0 + \dim T_D > n^2 + 2ns + m(n + s) \), such that the intersection can only be empty if \( T_0 \) and \( T_D \) are parallel. We distinguish between two cases:

First, let us assume that \( \dim T_D > \dim T_0 \), i.e. \( s > n \). If \( T_D \) and \( T_0 \) are indeed parallel, then we must have \( T_D^\perp \subset T_0^\perp \), where \( T_D^\perp \) denotes the ortho-complement to the linear subspace corresponding to \( T_D \) and is thus spanned by the columns of the matrix \( Q \) in (3.7). Note that \( T_0^\perp \) is spanned by the vectors \( e_{n^2 + n(m+s)+i}, i = 1, \ldots, ns \), where \( ej \) is a vector containing only zeros except for a one entry at position \( j \). Now, \( T_D^\perp \subset T_0^\perp \) implies that each column of \( Q \) in (3.7) can be written as a linear combination of the vectors \( e_{n^2 + n(m+s)+i}, i = 1, \ldots, ns \) which clearly yields a contradiction to the minimality of \( (A, \bar{B}, C, D) \) because it would mean that \( \bar{B}' \cap I_n = 0 \) and consequently \( \bar{B} = 0 \).

The second case is the case when \( \dim T_D \leq \dim T_0 \), i.e. \( s \leq n \). Parallelism of \( T_D \) and \( T_0 \) now implies \( T_0^\perp \subset T_D^\perp \), i.e. each vector \( e_{n^2 + n(m+s)+i}, i = 1, \ldots, ns \) can be written as a linear combination of the columns of \( Q \): there exist vectors \( v_i \in \mathbb{R}^2, i = 1, \ldots, ns \) such that \( Qv_i = e_{n^2 + n(m+s)+i}, i = 1, \ldots, ns \). Note that the vectors \( v_i \) must be linearly independent. From these equations and the structure of the matrix \( Q \) (see (3.7)), it follows immediately that \( \mathsf{rk}(I_n \otimes C) = ns \) (hence, all rows of \( C \) must be linearly independent) and that the rank of the matrix

\[
Q_{\text{red}} = \begin{pmatrix}
A' \otimes I_n - I_n \otimes A \\
\bar{B}' \otimes I_n \\
0_{sm \times n^2}
\end{pmatrix}
\]

is less than or equal to \( n^2 - ns \). In fact, its rank must be equal to \( n^2 - ns \) because the rank of \( Q \) as a whole is \( n^2 \) by assumption. It follows that each row of \( -(I_n \otimes C) \) must increase the rank of \( Q_{\text{red}} \) in (4.82), i.e. the rows of \( Q_{\text{red}} \) and the rows of \( (I_n \otimes C) \) must be linearly independent. To show that this cannot be the case, we note that each column of the \( \bar{B} \)-block in (4.82) must contain at least one non zero element as otherwise a row of \( \bar{B} \) would be zero and this would be a contradiction to minimality of \( (A, \bar{B}, C, D) \). Without restriction of generality, we can therefore assume that \( \tilde{b}_{11} \neq 0 \), and we get

\[
\begin{pmatrix}
0_{1 \times n^2} & c_{11} & c_{12} & \cdots & c_{1n} \\
\tilde{b}_{11} & 0_{1 \times (n-1)(m+s)} & 0_{1 \times sm}
\end{pmatrix}
\]

which yields a contradiction to the statement that the rows of \( Q_{\text{red}} \) and \( (I_n \otimes C) \) are linearly independent.

In total, we see that the intersection of \( T_0 \) and \( T_D \) always yields a non empty affine subspace, i.e. the trivial transfer function \((I, k) = (D, I)\) is always contained in \( V_D \).

(iv) Let us again consider the mapping \( F \) in (4.79) for the case of DDLC, i.e. \( F : GL(n) \times T_D \to S_m(n) \):

First, note the following: From a well known result (see, e.g. lemma 5.9 in (Bröcker and Jänich, 1990)) we know that for any regular value \( b \in S_m(n) \) of \( F \) the corresponding set \( F^{-1}(b) \) is a real
analytic submanifold of $GL(n) \times T_D$ of dimension zero. Note that this submanifold is given by the set of equations $F(T, T_D) = b = 0$ which can be transformed into a set of polynomial equations by multiplication by $\det(T)$, where $\det(T) \neq 0$ by assumption. Thus, $F^{-1}(b)$ is in fact a (semi-) algebraic set, and the dimension is zero. Clearly, from theorem (4.9.3) and the discussion below this theorem, $F^{-1}(b)$ consists of a finite number of points in $GL(n) \times T_D$.

Second, from Sard’s theorem (see, for instance, theorem 6.1 in (Bröcker and Jänich, 1990)) we know that the set of singular values of $F$ has Lebesgue measure zero in $S_m(n)$. It follows that for "almost all" points in $S_m(n)$ -- and therefore, because of continuity of $\pi$, for "almost all" $(l, k) \in M(n)$ -- the $(l, k)$-equivalence classes in $T_D$ consist of (at most) finitely many points in $T_D$. The set $V_D$ is known to contain an open subset of $M(n)$ by statement (ii), and thus it is also true that the $(l, k)$-equivalence classes of "almost all" $(l, k) \in V_D$ consist of (at most) finitely many points in $T_D$.

(v) Openness of $V_D$ in $\pi(T_D)$ follows from the definition of relative openness. Here $V_D = \pi(T_D) \cap M(n)$ and $M(n)$ is known to be open in $M(n)$; see (Hannan and Deistler, 1988), for instance. Denseness is trivial. The fact that $V_D$ need not necessarily be open in $M(n)$ has been discussed in detail for the special case $m = 0$ and $n = s = 1$ in (v) in section (4.9.3).

(vi) Note that $\pi(T_D) \subset V_D$ follows from the continuity of $\pi$. Every $\tilde{T}_D \in T_D$ can be written as $\tilde{T}_D = \lim_{t \to \infty} \gamma_{D,t}$ with $\gamma_{D,t} \in T_D$, such that $\pi(\tilde{T}_D) = \pi(\lim_{t \to \infty} \gamma_{D,t}) = \lim_{t \to \infty} \pi(\gamma_{D,t}) \in V_D$.

The fact that $\tilde{V}_D$ may or may not contain more transfer functions than can be represented in the closure of the parameter space has been discussed in detail in (vi), (a) and (b) in section (4.9.3).

\[ \square \]

## 4.10 DDLC combined with separable least squares methods

### 4.10.1 Introduction

One prerequisite for the usage of the slsDDLC parametrization in connection with ML estimation is that the concentration step described in section (2.3.3) can be performed. Considering the criterion function $L_T(Y_1^T; U_1^T, \tau^0)$ in (2.85), we could take $\tau^0$ to be a vector in $\mathbb{R}^{1+ns}$: $\tau^0 = \text{vec}(\tilde{A}', \tilde{C}')$. In other words, we could consider all matrix entries of observable pairs $(\tilde{C}, \tilde{A})$ to be free parameters. This has certain drawbacks, however: For any given observable pair $(\tilde{C}, \tilde{A})$, all pairs of the form $(T\tilde{A}T^{-1}, CT^{-1})$ correspond to the same original transfer function; this is a direct consequence of lemma (3.6.2).

Moreover, if we artificially add $B \in \mathbb{R}^{n \times s}$ and $D \in \mathbb{R}^{n \times s}$ such that $(\tilde{A}, B, \tilde{C}, D)$ becomes an observable square system -- see section (3.6) --, then the $L_f^k$ $k$-equivalence classes $E_{cc}(\tilde{A}, B, \tilde{C}, D)$ are real analytic manifolds of dimension $n^2 + ns + s^2$; see lemma (3.6.2). This in turn means that there are $n^2 + ns + s^2$ essentially unnecessary coordinates when using all entries in $(\tilde{A}, B, \tilde{C}, D)$ as free parameters.

Again, the idea now is to avoid this drawback by only considering the $ns$ dimensional ortho-complement $Q_{cc}^\perp(\tilde{A}, B, \tilde{C}, D)$ to the tangent space $Q_{cc}^c(\tilde{A}, B, \tilde{C}, D)$ to a certain $L_f^c k$-equivalence class at a given $(\tilde{A}, B, \tilde{C}, D)$ as a parameter space; see lemma (3.6.4) for a description of $Q_{cc}^\perp(\tilde{A}, B, \tilde{C}, D)$. Clearly, the parameter space will then be of dimension $ns$ and thus has no unnecessary coordinates.

Note the tangent space $Q_{cc}^c(\tilde{A}, B, \tilde{C}, D)$ can also be described in terms of the vectorization of the state-space matrices: see (3.34). The matrix $Q_{cc} \in \mathbb{R}^{n^2 + ns + s^2 \times (n^2 + ns + s^2)}$ has full column rank for any observable $(\tilde{A}, B, \tilde{C}, D)$. By $Q_{cc}^\perp$, we denote the matrix the columns of which span the orthogonal complement to the tangent space spanned by the columns of $Q_{cc}$. $Q_{cc}^\perp$ can again be obtained e.g. from a singular value decomposition of $Q_{cc}$. The slsDDLC parametrization is now defined as follows:

**Definition 4.10.1 (Separable least squares data driven local coordinates (slsDDLC)).** Let a minimal $(\tilde{A}, B, \tilde{C}, D, K)$ be given where $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K}) = (A - KC, B - KD, -C, -D, K)$ is the corresponding minimal inverse system and let **Assumption 3.6.1** be satisfied. The slsDDLC are given by the mapping
\[ \varphi_D : T_D^0 \rightarrow S_m(n) \quad (4.83) \]

\[ \tau_D^0 \rightarrow \Delta_U^V \left( \begin{array}{c}
 vec(\bar{A}(\tau_D^0)) \\
 vec(B(\tau_D^0)) \\
 vec(C(\tau_D^0)) \\
 vec(D(\tau_D^0))
\end{array} \right) = \Delta_U^V \left( \begin{array}{c}
 vec(\bar{A}) \\
 vec(B) \\
 vec(C) \\
 vec(D)
\end{array} \right) + Q_{cc}^\perp \tau_D^0 \]

Here, \( T_D^0 \subseteq \mathbb{R}^{ns} \) denotes the parameter space for \( \text{sisDDLc} \), i.e., the set of all \( \tau_D^0 \in \mathbb{R}^{ns} \) such that

\[ \Delta_U^V(\bar{A}(\tau_D^0), B(\tau_D^0), C(\tau_D^0), D(\tau_D^0)) \]

is well defined and minimal. The mapping \( \Delta_U^V \) is defined in (3.27) in lemma (3.6.1). Let \( V_D^0 = \pi(\varphi_D(T_D^0)) \).

**Remark 4.10.1.** Note that the matrices \( B \) and \( D \) are incorporated only for later use: There is no reason for considering \( B \) and \( D \), e.g., when implementing \( \text{sisDDLc} \) on a computer. In this case, one would only consider the rows of \( Q_{cc} \) corresponding to \( \bar{A} \) and \( \bar{C} \) and then use, e.g., a singular value decomposition of this matrix of size \( n^2 \times n^2 \) to determine its orthogonal complement to get the \( \text{sisDDLc} \) parametrization (4.83); the rows of \( Q_{cc}^\perp \) corresponding to \( B \) and \( D \) contain only zero entries.

**Remark 4.10.2.** Note that for any fixed minimal \( (\bar{A}, B, C, D, K) \), the mapping \( \varphi_D^0 \) from the parameter vectors \( \tau_D^0 \) to the state-space matrices of the original system is no longer affine but highly nonlinear due to the transformation \( \Delta_U^V \).

**Remark 4.10.3.** We will use the symbol \( \pi(T_D^0) = \pi(\varphi_D^0(T_D^0)) \) with slight abuse of notation.

### 4.10.2 An illustrative example

First, we will again consider the case where \( n = s = 1 \) and \( m = 0 \); see section (4.9.3).

We commence from a minimal system \( (a, k, c) \in S_m(1) \) where \( (\bar{a}, \bar{k}, \bar{c}) = (a - kc, k, -c) \) is the corresponding minimal inverse system. Let Assumption 3.6.1 be satisfied. Then we have

\[ Q_{cc} = \begin{pmatrix}
 0 & 0 & 0 \\
 0 & 1 & 0 \\
 -\bar{c} & 0 & 0 \\
 0 & 0 & 1
\end{pmatrix} \quad \text{and} \quad Q_{cc}^\perp = \begin{pmatrix}
 1 \\
 0 \\
 0 \\
 0
\end{pmatrix} \]

Using this \( Q_{cc}^\perp \), the \( \text{sisDDLc} \) parametrization is given by

\[ \Delta_U^V \left( \begin{array}{c}
 \bar{a}(\tau_D^0) \\
 B(\tau_D^0) \\
 C(\tau_D^0) \\
 D(\tau_D^0)
\end{array} \right) = \Delta_U^V \left( \begin{array}{c}
 \bar{a} \\
 B \\
 C \\
 D
\end{array} \right) + \tau_D^0 \]

\[ = \Delta_U^V \left( \begin{array}{c}
 \bar{a} + \tau_D^0 & B & C \\
 0 & 0 & D
\end{array} \right) \]

Here, \( T_D^0 = \{ \tau_D^0 \in \mathbb{R} | \Delta_U^V(\bar{a} + \tau_D^0, B, C, D) \text{ is well defined and minimal} \} \). Note that \( \bar{c} \neq 0 \) and that consequently

\[ \bar{X} = \bar{X}(y_1^{-1}, \bar{c}(\tau_D^0), \bar{a}(\tau_D^0), \bar{c}^2) = -\frac{1}{\bar{c}} \begin{pmatrix}
 0 \\
 \bar{c}y_1 \\
 \bar{c}a(\tau_D^0)y_1 + \bar{c}y_2 \\
 \vdots \\
 \bar{c}a(\tau_D^0)y_T - y_1 + \cdots + \bar{c}y_{T-1}
\end{pmatrix} \]

has full column rank for all \( \tau_D^0 \in \mathbb{R} \) unless \( y_1 = y_2 = \cdots = y_{T-1} = 0 \). The latter case is ruled out by Assumption 3.6.1, and therefore, using the notation of lemma (2.3.2), we have \( T^0 = \mathbb{R} \). Making use of figure (4.4), we will now discuss geometrical and topological properties of \( \text{sisDDLc} \), where the statements in italics refer to the general case discussed in theorem (4.10.1) below:
(i) It is immediately clear from figure (4.4) that \((a(\tau_D^0), B(\tau_D^0), c(\tau_D^0), D(\tau_D^0))\) is minimal for all \(\tau_D^0 \in \mathbb{R}\). Neglecting the scalar \(D(\tau_D^0)\), the straight line \((a + \tau_D^0, B, c)\) does not intersect the planes given by \(B = 0\) and \(c = 0\), respectively. Nevertheless, \(\Delta_V^U(a(\tau_D^0), B(\tau_D^0), c(\tau_D^0), D(\tau_D^0))\) may contain non minimal systems. Consider, for instance, the case \((a, B, c, D) = (0, 1, 1, 0)\) depicted in figure (4.4), where additionally \(Y^3 = (1, 1, 1)'\) and \(\sigma^2 = 1\), implying that

\[
\begin{pmatrix}
    \dot{a}(\tau_D^0) \\
    \dot{B}(\tau_D^0) \\
    \dot{c}(\tau_D^0) \\
    \dot{D}(\tau_D^0)
\end{pmatrix} = \begin{pmatrix}
    \tau_D^0 \\
    1 \\
    1 \\
    0
\end{pmatrix} \quad \begin{pmatrix}
    \dot{X} = \dot{X}(Y_1^3, c(\tau_D^0), \dot{a}(\tau_D^0), \sigma^2) = -
\end{pmatrix} \begin{pmatrix}
    0 \\
    1 \\
    \tau_D^0 + 1
\end{pmatrix}
\]

and

\[
\dot{\tau}^u = \tilde{k}(\tau_D^0) = (\dot{X}'\dot{X})^{-1} \dot{X}'Y_1^3 = -\frac{\tau_D^0 + 2}{1 + (\tau_D^0 + 1)^2}
\]

such that

\[
(a(\tau_D^0), \tilde{k}(\tau_D^0), c(\tau_D^0)) = (\tau_D^0, -\frac{(\tau_D^0 + 1)^2 + 2}{1 + (\tau_D^0 + 1)^2 + (1 + \tau_D^0 + (\tau_D^0)^2)^2}, 1)
\]

For \(\tau_D^0 = -2\), we get \((a(-2), \tilde{k}(-2), c(-2)) = (-2, 0, 1)\) and therefore \(\Delta_V^U(a(-2), B(-2), c(-2), D(-2)) = (-2, 0, -1)\) which is clearly non minimal. Hence, in this case, we have \(T_D^0 = \mathbb{R} \setminus \{-2\}\), which is obviously an open and dense subset of \(\mathbb{R}\).

(ii) Neglecting the scalar \(D\) again, it is straightforward to see from figure (4.4) that there exists an open interval \(T_D^{0,\text{loc}}\) containing 0 \(\in T_D^0\) such that each \(L^\infty_k\) \(k\)-equivalence class given by the plane described in remark (3.6.1) intersects the straight line \((a + \tau_D^0, B, c)\) only once and this intersection corresponds to a minimal original system. Considering the special case discussed in (i), the interval \(T_D^{0,\text{loc}}\) could be chosen to be \(T_D^{0,\text{loc}} = (-\infty, 0)\). In other words: \(T_D^0\) is locally identifiable at \(\tau_D^0 = 0\).

(iii) Clearly, the boundary points of \(T_D^0\) (which do not belong to \(T_D^0\) by statement (ii)) correspond to systems \((a(\tau_D^0), B(\tau_D^0), c(\tau_D^0), D(\tau_D^0))\) where \(\Delta_V^U(a(\tau_D^0), B(\tau_D^0), c(\tau_D^0), D(\tau_D^0))\) is either not defined or non minimal. In the special case considered in (i), the point \(\tau_D^0 = -2\) was the only boundary point where \(\Delta_V^U(a(-2), B(-2), c(-2), D(-2))\) was non minimal.

Note that \(T_D^0 = \mathbb{R}\) is also possible, implying that \(\pi(T_D^0) = \pi(\mathbb{R})\) does not contain any transfer functions of lower McMillan degree. To see this, consider the example in (i), but this time with \(Y_1^3 = (1, 1, 1)', i.e. with one more observation. We obtain:

\[
\begin{pmatrix}
    \dot{a}(\tau_D^0) \\
    \dot{B}(\tau_D^0) \\
    \dot{c}(\tau_D^0) \\
    \dot{D}(\tau_D^0)
\end{pmatrix} = \begin{pmatrix}
    \tau_D^0 \\
    1 \\
    1 \\
    0
\end{pmatrix} \quad \begin{pmatrix}
    \dot{X} = \dot{X}(Y_1^3, c(\tau_D^0), \dot{a}(\tau_D^0), \sigma^2) = -
\end{pmatrix} \begin{pmatrix}
    0 \\
    1 \\
    (\tau_D^0 + 1)^2 + (\tau_D^0 + 1)
\end{pmatrix}
\]

and

\[
\dot{\tau}^u = \tilde{k}(\tau_D^0) = (\dot{X}'\dot{X})^{-1} \dot{X}'Y_1^4 = -\frac{(\tau_D^0 + 1)^2 + 2}{1 + (1 + \tau_D^0)^2 + (1 + \tau_D^0 + (\tau_D^0)^2)^2}
\]

such that

\[
(a(\tau_D^0), \tilde{k}(\tau_D^0), c(\tau_D^0)) = (\tau_D^0, -\frac{(\tau_D^0 + 1)^2 + 2}{1 + (1 + \tau_D^0)^2 + (1 + \tau_D^0 + (\tau_D^0)^2)^2}, 1)
\]

which is minimal for all \(\tau_D^0 \in \mathbb{R}\). This is clearly different from the situation for DDLC in (iii) in section (4.9.3) where \(\pi(T_D)\) always contained transfer functions of lower order.
(iv) Within $T_D^0$, the $L_T^0$ $k$-equivalence classes (i.e., neglecting the scalar $D$ again, the intersection of the $L_T^0$ $k$-equivalence classes with the straight line $(\bar{a} + \tau_D^0 B, \bar{a})$ are seen to be single points: $T_D^0$ is globally identifiable in this case. Note again the difference to the simple case for DDLC discussed in (iv) in section 4.9.3. There, we had a lack of global identifiability of $T_D^0$.

To see that (local and) global identifiability may also be violated for the s1sDDLC parametrization, we have to consider another example where $n = s = 2$ and $m = 0$:

$$(A, B, C, D) = \left( \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right)$$

Then we have

$$Q_{cc} = \begin{pmatrix}
0 & -1 & 0 & 0 & \cdots & . \\
0 & 0 & 0 & 0 & \cdots & . \\
1 & 0 & 0 & -1 & \cdots & . \\
0 & 1 & 0 & 0 & \cdots & . \\
. & . & . & . & . & I_4 \\
. & . & . & . & . & . \\
\end{pmatrix} \quad \text{and} \quad Q_{cc}^\perp = \begin{pmatrix}
0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 \\
0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 \\
\end{pmatrix}$$

where the dots represent zero matrices of appropriate dimension (as will always be the case in the sequel). Note that the columns of $Q_{cc}^\perp$ are not orthogonal, which is not relevant for this investigation, however. Using this $Q_{cc}^\perp$, the s1sDDLC parametrization is given by

$$\Delta V^\perp \begin{pmatrix}
\text{vec}(\bar{A}(\tau_D^0)) \\
\text{vec}(\bar{B}(\tau_D^0)) \\
\text{vec}(\bar{C}(\tau_D^0)) \\
\text{vec}(\bar{D}(\tau_D^0))
\end{pmatrix} = \Delta V^\perp \begin{pmatrix}
\tau_D^{0,3} + \tau_D^{0,4} \\
\tau_D^{0,1} \\
1 + \tau_D^{0,4} \\
\tau_D^{0,2} + \tau_D^{0,3} \\
1 \\
0 \\
1 + \tau_D^{0,4} \\
\tau_D^{0,2} + \tau_D^{0,3} \\
0 \\
1 - \tau_D^{0,4} \\
0 \\
0 \\
0 \\
0 \\
\end{pmatrix}$$

Using the notation $Q_{cc}(\tau_D^0)$ as in (4.90) below, we thus get

$$[Q_{cc}(\tau_D^0), Q_{cc}^\perp] = \begin{pmatrix}
0 & \tau_D^{0,1} & -\tau_D^{0,3} \\
-\tau_D^{0,1} & \tau_D^{0,2} - \tau_D^{0,3} & \tau_D^{0,3} \\
(1 + \tau_D^{0,4}) & \tau_D^{0,3} & \tau_D^{0,3} \\
(1 + \tau_D^{0,4}) & 0 & \tau_D^{0,3} \\
0 & \tau_D^{0,3} & \tau_D^{0,3} \\
0 & 0 & \tau_D^{0,3} \\
. & . & . & . & . & I_4 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
\end{pmatrix}$$

and for $\tau_D^0 = (x, -1, y, -1)$ where $x \neq 0$ and $y \in \mathbb{R}$, we get the minimal system
\[
(\bar{A}(\bar{\tau}_B^0), B(\bar{\tau}_B^0), \bar{C}(\bar{\tau}_B^0), D(\bar{\tau}_B^0)) = \left( \begin{array}{ccc}
y - 1 & 0 & 0 \\
x & y - 1 & 0 \\
0 & 0 & 2 \\
0 & 0 & 0
\end{array} \right)
\]
(4.87)

where local identifiability is violated. This is because the rank of \([Q_{cc}(\bar{\tau}_B^0), Q_{cc}^\perp]\) decreases from 16 to 14 and thus the mapping \(F^\circ\) in (4.89) below cannot be locally injective; see theorem (4.10.2) below:

\[
\begin{bmatrix}
Q_{cc}(\bar{\tau}_B^0), Q_{cc}^\perp
\end{bmatrix} = \begin{bmatrix}
0 & 0 & x & 0 & \ldots & 0 & 0 & 1 & 1 \\
-x & 0 & 0 & x & \ldots & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 1 \\
0 & 0 & -x & 0 & \ldots & 0 & 1 & 1 & 0 \\
\ldots & \ldots & \ldots & I_4 & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & -1 \\
0 & -2 & 0 & 0 & \ldots & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & -2 & \ldots & 0 & 0 & 0 & -1 \\
\ldots & \ldots & \ldots & I_4 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots
\end{bmatrix}
\]

For fixed \(y\), the parameter vectors \(\bar{\tau}_B^0 = (x, -1, y, -1)\) and \(\bar{\tau}_B^0 = (-x, -1, y, -1)\) represent the same transfer function of McMillan degree two (if \(\Delta_Y(\bar{A}(\bar{\tau}_B^0), B(\bar{\tau}_B^0), \bar{C}(\bar{\tau}_B^0), D(\bar{\tau}_B^0))\) is well defined and minimal): The state transformation \(T = \text{diag}(-1, 1)\) maps \(\bar{A}(\bar{\tau}_B^0)\) and \(\bar{C}(\bar{\tau}_B^0)\) in (4.87) onto \(\bar{A}(\bar{\tau}_B^0)\) and \(\bar{C}(\bar{\tau}_B^0)\), respectively, and this shows that the systems \(\Delta_Y(\bar{A}(\bar{\tau}_B^0), B(\bar{\tau}_B^0), \bar{C}(\bar{\tau}_B^0), D(\bar{\tau}_B^0))\) and \(\Delta_Y(\bar{A}(\bar{\tau}_B^0), B(\bar{\tau}_B^0), \bar{C}(\bar{\tau}_B^0), D(\bar{\tau}_B^0))\) are observationally equivalent. Note that this is true for arbitrarily large \(|x|\), i.e. the \texttt{81sbdlc} parametrization is neither locally identifiable at every point in the parameter space nor is it globally identifiable.

However, in both of the cases treated here, we can observe that every transfer function in \(V_B^0\) has just a finite number of representatives within \(T_B^0\).

(v) \(V_B^0\) is clearly open in \(\pi(T_B^0)\). Note that the dimension of the parameter space \(T_B^0\) is strictly less than the dimension of the manifold \(M(n)\). Hence, \(V_B^0\) cannot be open in \(M(n)\).

(vi) We now discuss two cases where \(\pi(T_B^0) = \bar{V}_B^0\) and a case where \(\pi(T_B^0) \subset \bar{V}_B^0\) where the inclusion is strict:

(a) In the special case considered in (i) we had \(T_B^0 = \mathbb{R} \setminus \{-2\}\). The systems (4.85) correspond to the transfer function

\[
\bar{k}(z) = \pi(\bar{a}(\bar{\tau}_B^0), \bar{c}(\bar{\tau}_B^0)) = \frac{1 - (\bar{\tau}_B^0 + \bar{\tau}_B^{0+2})}{1 - \bar{\tau}_B^0 z}
\]

For the original model, we have

\[
\Delta_Y(\bar{a}(\bar{\tau}_B^0), B(\bar{\tau}_B^0), \bar{c}(\bar{\tau}_B^0), D(\bar{\tau}_B^0)) = (a(\bar{\tau}_B^0), k(\bar{\tau}_B^0), c(\bar{\tau}_B^0)) = \left( \bar{\tau}_B^0 + \frac{\bar{\tau}_B^0 + 2}{1 + (\bar{\tau}_B^0 + 1)^2}, -1 \right)
\]

corresponding to

\[
k(z) = \pi(a(\bar{\tau}_B^0), k(\bar{\tau}_B^0), c(\bar{\tau}_B^0)) = \frac{1 - \bar{\tau}_B^0 z}{1 - (\bar{\tau}_B^0 + \bar{\tau}_B^{0+2})}
\]
For \( \tau_D^0 \to \pm \infty \), we have \((a(\tau_D^0), k(\tau_D^0), c(\tau_D^0)) \to (\pm \infty, 0, -1)\) and the corresponding sequence of transfer functions also diverges in \( \mathcal{T}_D \):

\[
(K_j)_{j \in \mathbb{N}} = \left( 1, \frac{\tau_D^0 + 2}{1 + (\tau_D^0 + 1)^2}, K_1 a(\tau_D^0), K_1 a(\tau_D^0)^2, \ldots \right) \to (1, 0, 1, \pm \infty, \ldots)
\]

because the degree of the numerator polynomial of \( K_2 \) is seven, whereas the denominator polynomial is of degree six only. This implies that no additional transfer function can be obtained in \( \bar{V}_D^0 \). Hence, we have \( \pi(\bar{T}_D^0) = \bar{V}_D^0 \), containing lower degree transfer functions. Note that we have divergence in \( \mathcal{T}_D \) despite the fact that there is a pole-zero cancellation at \( z = 0 \) in the limit.

(b) In the special case considered in (iii), we had \( \mathcal{T}_D = \mathbb{R} \) and consequently the parameter space did not contain any boundary points: \( \pi(\bar{T}_D^0) = \pi(\mathcal{T}_D^0) = \bar{V}_D^0 \). As can be easily seen by calculations analogous to (a), for \( \tau_D^0 \to \pm \infty \) the corresponding sequence of transfer functions again diverges in \( \mathcal{T}_D \): No additional transfer function (neither of order one nor of order zero) can be obtained in \( \bar{V}_D^0 \), again implying that \( \pi(\mathcal{T}_D^0) = \bar{V}_D^0 \). Note that \( \bar{V}_D^0 = V_D^0 \) in this case.

(c) In order to show that \( \pi(\bar{T}_D^0) \neq \bar{V}_D^0 \) is also possible, we consider an example where \( n = 1, s = 2 \) and \( m = 0 \):

\[
(a, B, C, D) = \left( 1, \begin{pmatrix} 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right)
\]

Then we have

\[
Q_{cc} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\quad \text{and} \quad
Q_{cc}^{-1} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
\sqrt{2} & 0 & 0 & 0 \\
0 & \sqrt{2} & 0 & 0 \\
0 & 0 & \sqrt{2} & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

Using this \( Q_{cc}^{-1} \), the \( s1sDLC \) parametrization is given by

\[
\Delta^U_Y \begin{pmatrix}
vec(a(\tau_D^{0,1}, \tau_D^{0,2})) \\
vec(B(\tau_D^{0,1}, \tau_D^{0,2})) \\
vec(C(\tau_D^{0,1}, \tau_D^{0,2})) \\
vec(D(\tau_D^{0,1}, \tau_D^{0,2}))
\end{pmatrix}
= \Delta^U_Y \begin{pmatrix}
1 + \tau_D^{0,1} \\
1 + \sqrt{2} \tau_D^{0,2} \\
-1 + \sqrt{2} \tau_D^{0,2} \\
0
\end{pmatrix}
\quad (4.88)
\]

For \( Y^t_1 = (1, 0, 0, 1, 0, 0)' \) and \( \sigma^2 = 1 \) we get

\[
\tilde{X} = \tilde{X}(Y_1^{t,2}, C(\tau_D^{0,1}, \tau_D^{0,2}), a(\tau_D^{0,1}, \tau_D^{0,2}), \sigma^2) = -\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
1 + \sqrt{2} \tau_D^{0,2} & 0 & 0 \\
-1 + \sqrt{2} \tau_D^{0,2} & 0 & 0 \\
(1 + \sqrt{2} \tau_D^{0,2})(1 + \tau_D^{0,1}) & 1 + \sqrt{2} \tau_D^{0,2} & 0 \\
(-1 + \sqrt{2} \tau_D^{0,2})(1 + \tau_D^{0,1}) & -1 + \sqrt{2} \tau_D^{0,2} & 0
\end{pmatrix}
\]

which has full column rank for all \( \tau_D^0 \in \mathbb{T}_D \). Furthermore, simple calculations yield
\[
\bar{\tau} = \bar{\mathcal{K}}(\tau_{D1}^{0,1}, \tau_{D2}^{0,2}) = (\bar{X}, \bar{X})^{-1} \bar{X} \gamma_{\text{I}}^2 = \frac{1}{(\tau_{D1}^{0,2})^2 + 2} \left( 1 - \frac{1}{\sqrt{\tau_{D1}^{0,2}}} - 1 \right) \left( \frac{1}{\sqrt{\tau_{D2}^{0,2}}} - 1 \right) \left( 1 + \tau_{D1}^{0,1} \right)
\]
Hence, \((\bar{a}(\tau_{D1}^{0,1}, \tau_{D2}^{0,2}), \bar{\mathcal{K}}(\tau_{D1}^{0,1}, \tau_{D2}^{0,2}), \bar{C}(\tau_{D1}^{0,1}, \tau_{D2}^{0,2}))\) equals
\[
\left( 1 + \tau_{D1}^{0,1}, \frac{1}{(\tau_{D1}^{0,2})^2 + 2} \left( \frac{1}{\sqrt{\tau_{D1}^{0,2}}} - 1 \right) \left( \frac{1}{\sqrt{\tau_{D2}^{0,2}}} - 1 \right) \left( 1 + \frac{1}{\sqrt{\tau_{D2}^{0,2}}} - 1 \right) \left( 1 + \tau_{D1}^{0,1} \right) \right)
\]
For \(\tau_{D2}^{0,2} = \sqrt{2}\), we get a non minimal system, implying that \(\bar{T}_D = \mathbb{R}^2 \setminus \{(x, \sqrt{2}), x \in \mathbb{R}\}\). Clearly,
\[
\pi(\bar{T}_D^0) = \pi(\mathbb{R}^2) = V_D^0 \cup \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\} \ \text{where} \ \kappa(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\]
is the trivial transfer function of McMillan degree zero. However, in this case, \(\bar{T}_D^0\) contains an additional transfer function of degree one. To see this, consider a sequence \((\tau_{D1}^{0,1}, \tau_{D2}^{0,2}) = (0, \pm t), t \in \mathbb{N}\). Then we have
\[
(\bar{a}(\tau_{D1}^{0,1}, \tau_{D2}^{0,2}), \bar{\mathcal{K}}(\tau_{D1}^{0,1}, \tau_{D2}^{0,2}), \bar{C}(\tau_{D1}^{0,1}, \tau_{D2}^{0,2})) \rightarrow \left( 1, \left( \begin{array}{cc} 0 & 0 \\ \pm \infty \end{array} \right) \right)
\]
Note that the Markov parameters are given by
\[
(\bar{K})_{j \in \mathbb{N}} = \left( \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right), \frac{1}{(\tau_{D2}^{0,2})^2 + 2} \left( \frac{1}{\sqrt{\tau_{D2}^{0,2}}} - 1 \right) \left( \frac{1}{\sqrt{\tau_{D2}^{0,2}}} - 1 \right) \left( 1 + \tau_{D1}^{0,1} \right)
\]
such that
\[
(\bar{K})_{j \in \mathbb{N}} \rightarrow \left( \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right), \left( \frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t}} \right), \left( \frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t}} \right), \left( \frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t}} \right), \left( \frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t}} \right), \left( \frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t}} \right)
\]
Here we have convergence in \(T_{pl}\) to a degree one transfer function in \(\bar{T}_D^0\) which cannot be represented in \(\pi(\bar{T}_D^0)\).

To sum it up, it is possible that the same transfer functions are described in the closure of the parameter space \(T_D^0\) and in the closure of the corresponding transfer function space \(V_D^0\), but it is also possible that the closure of the transfer function space contains more transfer functions than can be represented in the closure of the parameter space.

Remark 4.10.4. In the example considered in (4.84) we saw that iterating the \(\text{sldDDLC}\) construction commencing from an arbitrary minimal initial \((a, k, c)\) leaves \(V_D^0\) unchanged: For any system on the (subset of the) straight line \((\bar{a} + \tau_{D1}^{0,1}, B, \bar{c}, D)\) corresponding to \(T_D^0\), the ortho-complement to its \(L^\infty_\mathbb{R}\) k-equivalence class coincides with the straight line corresponding to the "old" \(T_D^0\). Note that this is different from the situation for DDLC, where an iteration of the DDLC construction left \(V_D\) unchanged only for (Lyapunov balanced) minimal initial systems of the form \((a, k, \pm k)\); compare remark (4.9.2) in section (4.9.3).

Remark 4.10.5. Neglecting the scalar \(D\) in (4.84), it is evident that the (subset of the) straight line \((\bar{a} + \tau_{D1}^{0,1}, B, \bar{c})\) corresponding to \(T_D^0\) intersects "almost all" planes corresponding to the \(L^\infty_\mathbb{R}\) k-equivalence classes. Hence, the corresponding \(V_D^0\) only "leaves out" a thin set out of the \(n\) dimensional set of transfer functions where the optimization is performed: By only one iteration of \(\text{sldDDLC}\), we reach almost all transfer functions which are "candidates" for the optima of the criterion function. Again, this is different from the situation for DDLC, where \(V_D\) was never generic in \(M(n)\); compare remark (4.9.3) in section (4.9.3).

Remark 4.10.6. Note that in (4.84), \(V_D^0\) is independent of the particular initial system \((\bar{a}, B, \bar{c}, D)\) in the \(L^\infty_\mathbb{R}\) k-equivalence class where the \(\text{sldDDLC}\) construction is performed; This is again a difference to the situation for DDLC; compare remark (4.9.4) in section (4.9.3).
4.10.3 Topological and geometrical properties

The following theorem holds true for arbitrary $n$, $s$ and $m$:

**Theorem 4.10.1.** Let a minimal $(A, B, C, D, K)$ be given where $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{K}) = (A - KC, B - KD, -C, -D, K)$ is the corresponding minimal inverse system and let Assumption 3.6.1 be satisfied. The parametrisation by $\text{SLsDDLC}$ as given in (4.33) has the following properties:

(i) $T^0_B$ is an open and dense subset of $\mathbb{R}^n$.

(ii) There exists an open neighborhood $T^{0,\text{loc}}_B$ of $0 \in T^0_B$ such that $T^{0,\text{loc}}_B$ is identifiable and the mapping $\psi^{0,\text{loc}}_D : V^{0,\text{loc}}_D \rightarrow T^{0,\text{loc}}_B$ defined by $\psi^{0,\text{loc}}_D(\pi(\tau_D)) = \tilde{\tau}_D$ is a homeomorphism where $V^{0,\text{loc}}_D = \pi(T^{0,\text{loc}}_B)$.

(iii) $\pi(\tilde{T}_B)$ can (but need not necessarily) contain transfer functions of lower McMillan degree.

(iv) For "almost every" $(l, k) \in V^*_B$, the $(l, k)$-equivalence class in $T^0_B$ consists of a finite number of isolated points.

(v) $V^*_B$ is open (and trivially dense) in $\pi(T^0_B)$.

(vi) $\bar{\pi}(\tilde{T}_B) \subseteq V^*_B$ where equality can hold, but the inclusion can also be strict, and where $\tilde{T}_B$ is the set of all points in $\tilde{T}^0_B$ where $\tilde{\varphi}_D$ is well-defined.

**Remark 4.10.7.** Remarks analogous to those made in remark (4.9.5) below theorem (4.9.1) can also be made at this point. Note, however, the difference to DDLC in statement (ii): As the sets $T^{0,\text{loc}}_B$ and $V^{0,\text{loc}}_D$ depend on data, it is still an open question how the coordinate free consistency result in section (2.3.4) can be carried over to yield consistency in terms of parameter estimates.

**Remark 4.10.8.** The conjecture that equality holds in the single output case in statement (vi), whereas the inclusion is strict in the case $s > 1$, is still left for future research.

**Remark 4.10.9.** In analogy to remark (4.10.4), one could ask whether it is possible in general case that any transfer function within the same connected component as the initial transfer function can be reached by iterating the $\text{SLsDDLC}$ construction finitely times. This is still an open question, though there will be some further related discussion in chapter 5.

**Remark 4.10.10.** The conjecture that $V^*_B$ can also be non-generic within the set of all transfer functions corresponding to all observable pairs $(C', \tilde{A})$ and uniquely determined $(D, B, K)$ (by (3.27)) is still to be proved; in the simple case treated in remark (4.10.5), $V^*_B$ is generic within this set.

Before theorem 4.10.1 is proved, let us consider the mapping

$$F^o : GL(n) \times \mathbb{R}^{*n} \times \mathbb{R}^{*n} \times T^0_B \rightarrow S(n)$$

$$(T, \Theta_B, \Theta_D, \tau^0_B) \rightarrow \begin{pmatrix} \text{vec}(T \tilde{A}(\tau^0_B)T^{-1}) \\ \text{vec}(\Theta_B) \\ \text{vec}(\tilde{C}(\tau^0_B)T^{-1}) \\ \text{vec}(\Theta_D) \end{pmatrix} = \begin{pmatrix} T^{-1} \otimes T & 0 & 0 & 0 \\ 0 & I_s \otimes I_n & 0 & 0 \\ 0 & 0 & T^{-1} \otimes I_s & 0 \\ 0 & 0 & 0 & I_s \otimes I_n \end{pmatrix} \begin{pmatrix} \text{vec}(\tilde{A}(\tau^0_B)) \\ \text{vec}(\Theta_B) \\ \text{vec}(\tilde{C}(\tau^0_B)) \\ \text{vec}(\Theta_D) \end{pmatrix}$$

$T^{0,\text{vec}}$

Note that $\frac{\partial F^o}{\partial (T, \Theta_B, \Theta_D)}((T, \Theta_B, \Theta_D), \tau^0_B)$ has already been computed in (3.33) and describes the tangent space to the $L^0_k$ equivalence class $E^{0,\text{vec}}(\tilde{A}(\tau_B), B(\tau_B), \tilde{C}(\tau_B), D(\tau_B))$. Hence, we get

$$\frac{dF^o}{d((T, \Theta_B, \Theta_D), \tau^0_B)}((T, \Theta_B, \Theta_D), \tau^0_B) = X^*(\tau^0_B)$$

$$T^{0,\text{vec}} \cdot \begin{pmatrix} \tilde{A}(\tau^0_B)' \otimes I_n - I_n \otimes \tilde{A}(\tau^0_B) \\ -I_n \otimes \tilde{C}(\tau^0_B) \\ 0 \\ Q_{cc} \end{pmatrix} = \begin{pmatrix} Q_{cc}^+ \\ Q_{cc} \end{pmatrix}$$

$$\begin{pmatrix} I_n \otimes T^{-1} & 0 & 0 & 0 \\ 0 & I_s \otimes I_n & 0 & 0 \\ 0 & 0 & I_s \otimes I_s & 0 \\ 0 & 0 & 0 & I_{ns} \end{pmatrix}$$
and it is evident that
\[
\begin{align*}
\operatorname{rk} \left( \frac{dF^o}{d((T, \Theta_B, \Theta_D), \tau_D^0)} \right) (T, \Theta_B, \Theta_D) & = \operatorname{rk} \left( \frac{dF^o}{d((T, \Theta_B, \Theta_D), \tau_D^0)} \right) (I, 0, 0) \\
& = \operatorname{rk} \left( X^o(\tau_D^0) \right)
\end{align*}
\]
for all \( T \in \text{GL}(n) \). In analogy to theorem (4.9.2), we get

**Theorem 4.10.2.** Let \( T_D^0 \subset \mathbb{R}^{ns} \) denote an open subset of \( \mathbb{R}^{ns} \) and let \( \varphi_D : T_D^0 \to S_m(n) \) be the s1sDLDC parametrization given in (4.88). Then the s1sDLDC parametrization

(i) is locally identifiable at \( \tau_D^0 = \tau_D^{0,0} \) if and only if the mapping \( F^o \) in (4.89) is locally injective at
\[
\left((T, \Theta_B, \Theta_D), \tau_D^0 \right) = \left((I, 0, 0), \tau_D^{0,0} \right).
\]
(ii) If the rank of \( X^o(\tau_D^0) \) equals \( r \) for all \( \tau_D^0 \in U(\tau_D^{0,0}) \), where \( X^o(\tau_D^0) \) is given in (4.90) and \( U(\tau_D^{0,0}) \) is some open neighborhood of \( \tau_D^{0,0} \) in \( T_D^0 \), then the s1sDLDC parametrization is locally identifiable at the point \( \tau_D^0 = \tau_D^{0,0} \) if and only if \( r = n^2 + 2ns + s^2 \), or, equivalently, if and only if \( \det(X^o(\tau_D^{0,0})) \neq 0 \).

**Proof.** The proof proceeds along the same lines as the proof of theorem (4.9.2). Arguments for the equivalence classes \( \mathcal{E}(A, B, C, D) \) have to be replaced by arguments for the \( \mathcal{F}^o \) \( k \)-equivalence classes which are characterized by lemma (3.6.2).

Of course, the domain of definition of \( F^o \) is an open subset of \( \mathbb{R}^{n^2 + 2ns + s^2} \) and \( F^o \) is rational in \( T \) and linear in \( \Theta_B, \Theta_D \) and \( \tau_D^0 \).

We are now ready to prove theorem (4.10.1):

**Proof.** (i) We consider the mapping \( \Delta^o : T^o \to \mathbb{R} \) attaching \( \det(W^o_n(\tau_D^0)W^o_n(\tau_D^0)) \) to \( \tau_D^0 \) where \( W^o_n(\tau_D^0) = C(n)(\tau_D^0) \in \mathbb{R}^s \times n \) with
\[
O_n(\tau_D^0) = O_n(\Delta^U(\alpha(\tau_D^0), B(\tau_D^0), C(\tau_D^0), D(\tau_D^0)) = O_n(\varphi_D(\tau_D^0)))
\]
being the corresponding (finite) observability matrix; see definition (1.4.2). The matrix \( W^o_n(\tau_D^0) \) is given analogously. We can directly apply lemma (2.3.2) for \( \tau_D^0 = 0 \in \mathbb{R}^{ns} \) to see that \( \Delta^o \) is in fact well-defined (and analytic) on the open and dense subset \( T^o \) of \( \mathbb{R}^{ns} \). Clearly, \( \det(W^o_n(\tau_D^0)W^o_n(\tau_D^0)) \neq 0 \) if and only if \( \varphi_D(\tau_D^0) \) is minimal. Hence, \( T_D^0 = (\Delta^o)^{-1}(\mathbb{R} \times \{0\}) \) is open in \( T^o \) (and therefore also in \( \mathbb{R}^{ns} \)) because it is the inverse image of an open set and \( \Delta^o \) is trivially also continuous. Denseness of \( T_D^0 \) in \( T^o \) (and therefore also in \( \mathbb{R}^{ns} \)) follows from the fact that \( \Delta^o(\tau_D^0) = 0 \) can only hold true on a thin subset of each connected component of \( T^o \) because \( \Delta^o \) is analytic; the fact that \( \Delta^o \) does not vanish on a whole connected component of \( T^o \) is easily seen.

(ii) The rank condition for \( X^o(\tau_D^0) \) of theorem 4.10.2 (ii) is easily verified for \( \tau_D^{0,0} = 0 \) as \( X^o(0) = [Q_{cc} : Q_{cc}] \). Moreover, \( X^o(\tau_D^0) \) has constant rank \( n^2 + 2ns + s^2 \) in a neighborhood \( O \) of \( \tau_D^{0,0} = 0 \) because \( X^o(\tau_D^0) \) depends continuously on \( \tau_D^0 \) and the determinant is a continuous function of the entries in \( X^o(\tau_D^0) \). This shows local identifiability at \( \tau_D^0 = 0 \) by theorem 4.10.2, i.e., injectivity of the function \( \pi|_{T_D^0} = \pi \circ \varphi_D|_{T_D^0} = \pi \circ \Delta^o \circ F^0|_{(I, B, D), X T_D^0} \) where \( T_D^0 \subseteq O \) is chosen appropriately and \( \Delta^o \) is given in (3.27). Note that \( \pi \circ \varphi_D|_{T_D^0} \) is clearly continuous because \( \pi \) and \( \varphi_D|_{T_D^0} \) are continuous. Set \( \pi(T_D^0) = V_D^0 \). Then \( \pi|_{D^0} : T_D^0 \to V_D^0 \) is surjective by definition, it is injective by what was said above, and thus it is bijective and continuous. Note that \( \pi \circ \Delta^o \circ F^0|_{(I, B, D), X T_D^0} = \pi \circ \Delta^o \circ F^0|_{(GL(n), \mathbb{R}^{n \times n \times n}, \mathbb{R}^{n \times n \times n}) X T_D^0} \) is seen to be an open mapping:

\[
\begin{align*}
F^0|_{(GL(n), \mathbb{R}^{n \times n \times n}) X T_D^0} \text{ is a diffeomorphism and therefore open by the same argument as } F
\end{align*}
\]
was shown to be open in remark (4.9.11); clearly, \( T_D^0 \subseteq O \) has been chosen appropriately.
4.11 Orthogonal data driven local coordinates

Before we provide a detailed description of orthoDDLc, let us briefly review the basic philosophy behind the use of sllsDDLc and the extension to orthoDDLc:

As it is the case for the whole class of data driven local parametrizations, we try to avoid the use canonical forms in the first place. Secondly, we try to keep the number of parameters to a minimum, and making use of the separable least squares principle we can reduce the original optimization problem to a nonlinear estimation problem in only \( ns \) parameters. The parameter space is then again constructed "locally"
around the current system estimate, following the same idea as for conventional DDLC. This results in the \texttt{s1sDDLC} parametrization introduced in section (4.10).

For \texttt{orthoDDLC}, we will additionally restrict the optimization to (a subset of) the set $BSA(n)$, which is known to be a subset of the compact ball of radius $\sqrt{n+s}$ in $S(n)$; see lemma (3.5.1). This restriction is achieved by easily implementable modifications of an iterative estimation algorithm.

### 4.11.1 Introduction

The term \texttt{orthoDDLC} stands for Orthogonal Data Driven Local Coordinates and refers to a (class of) estimation algorithms which makes use of the \texttt{s1sDDLC} parametrization as given in section (4.10) above. For introducing \texttt{orthoDDLC}, the following two lemmas will be needed:

**Lemma 4.11.1 (Orthogonal projection onto the tangent space $Q_{\{\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D}\}}^{BSA}$).** Let $(\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D})$ be a $BSA$ realization, let $R \in \mathbb{R}^{(n+s) \times (n+s)}$ be the corresponding realization matrix and let $\Delta R \in \mathbb{R}^{(n+s) \times (n+s)}$ be arbitrary. Then the orthogonal projection (in Frobenius norm) of $R + \Delta R$ onto the affine space $Q_{\{\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D}\}}^{BSA}$ is given by

$$\hat{R} = R + (R' \cdot \Delta R)_{skew}$$

where $X_{skew}$ denotes the skew symmetric part of the matrix $X$, i.e. $X_{skew} = \frac{1}{2}(X - X')$.

**Proof.** From (3.25) it is obvious that $\hat{R}$ is an element of $Q_{\{\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D}\}}^{BSA}$. Moreover, $\Delta R = R (R' \cdot R)_{skew} + \{\Delta R - R (R' \cdot \Delta R)_{skew}\}$ is the first and the second component are orthogonal:

$$\text{tr} \left( (R' \cdot \Delta R)_{skew} \right) (\Delta R - R (R' \cdot \Delta R)_{skew}) = 0$$

$$\text{tr} [ - (R' \cdot \Delta R)_{skew} R' (R' \cdot R)_{skew} + (R' \cdot \Delta R)_{skew} R' R (R' \cdot \Delta R)_{skew} ] = 0$$

$$\text{tr} (R' \cdot (R' \cdot \Delta R)_{skew} (R' \cdot \Delta R)_{sym} = 0$$

The last but one equality follows from the fact that $R' R = I$ and the last equality follows from the fact that if $X = X_{skew} + X_{sym}$ with $X_{skew} = \frac{1}{2}(X - X')$ and $X_{sym} = \frac{1}{2}(X + X')$, then $\text{tr}(X_{skew} \cdot X_{sym}) = \frac{1}{4}\text{tr}(XX - (XX)' + XX' - XX') = 0$, which follows by cyclical permutation within the trace operator. Hence, the orthogonal projection of $R + \Delta R$ onto $Q_{\{\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D}\}}^{BSA}$ is given by (4.91).

The second result is the following:

**Lemma 4.11.2 (Best approximation in $O(n+s)$).** Let $\hat{R} \in \mathbb{R}^{(n+s) \times (n+s)}$ be a realization matrix. Then the unique matrix $R \in O(n+s)$ that solves the minimization problem

$$\min_{R \in O(n+s)} \| \hat{R} - R \|_F$$

is given by $R = U \hat{V}'$ where $\hat{R} = U \hat{\Sigma} V'$ is a singular value decomposition of $\hat{R}$.

**Proof.** As $\| \hat{R} - R \|_F^2 = \text{tr} (\hat{R}' \hat{R}) - 2 \text{tr} (\hat{R}' R) + \text{tr} (R' R) = \| \hat{R} \|_F^2 - 2 \text{tr} (\hat{R}' R) + (n+s)$, we have to maximize $\text{tr} \hat{R}^2 R$. Using a singular value decomposition $\hat{R} = U \hat{\Sigma} V'$, we can define an orthogonal $Z = V' R U'$, where $R$ is any orthogonal matrix. We have to show that the maximizing $R$ is given by $U V'$, i.e. $Z = I$. Clearly,

$$\text{tr} (R' \hat{R}) = \text{tr} (R' U \hat{\Sigma} V') = \text{tr} (V' R' U \hat{\Sigma}) = \text{tr} (Z \hat{\Sigma}) = \sum_{i=1}^{n+s} z_i \tilde{\sigma}_i \leq \sum_{i=1}^{n+s} \tilde{\sigma}_i$$

where the last inequality follows from the fact that all entries of an orthogonal matrix must have modulus less than or equal to one. The upper bound is attained if and only if $z_i = 1$ for all $i = 1, \ldots, n+s$. Hence, $Z = I$ (because $Z$ is orthogonal) and $R = UV'$ which is unique although $U$ and $V$ may not be uniquely determined.
4.11. ORTHOGONAL DATA DRIVEN LOCAL COORDINATES

Remark 4.11.1. The situation treated in lemma (4.11.2) above is a special case of the (balanced) orthogonal Procrustes problem of finding a matrix \( Q \in O(n) \) that solves the minimization problem

\[
\min_{Q \in O(n)} \| A - BQ \|_F
\]

for given matrices \( A \in \mathbb{R}^{n \times n} \) and \( B \in \mathbb{R}^{n \times n} \); see (Golub and VanLoan, 1996). The solution to this problem is given by \( Q = UV^t \) where \( B' A = U \Sigma V' \).

We are now ready to state the following

Definition 4.11.1 (Orthogonal data driven local coordinates (OrthoDDLc)). Let a minimal and strictly minimum phase \((A, B, C, D, K)\) be given where \((\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K}) = (A - KC, B - KD, -C, -D, K)\) is the corresponding minimal and stable inverse system and let Assumption 3.6.1 be satisfied. An algorithm that proceeds according to the following steps is called an OrthoDDLc algorithm, or, briefly, OrthoDDLc:

(i) Consider the Lyapunov equation (3.17) and calculate the output Gramian \( \bar{Q} > 0 \) corresponding to the system \((\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K})\). Perform a state transformation of the form \( T = Q^{1/2} \) where \( Q \) denotes any square root of \( \bar{Q} \). Then the resulting system is output normal (i.e., its output Gramian is the identity matrix) and will, with slight abuse of notation, also be denoted by \((\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K})\).

(ii) Compute the singular value decomposition which can be chosen to be of the following form

\[
\begin{pmatrix}
\bar{C} \\
\bar{A}
\end{pmatrix} = \begin{pmatrix}
\bar{C} \\
\bar{A}
\end{pmatrix} U_2 \begin{pmatrix}
I_n \\
0_{s \times n}
\end{pmatrix} V' \quad \text{and set} \quad R = \begin{pmatrix}
D & \bar{C} \\
B & \bar{A}
\end{pmatrix} = \begin{pmatrix}
U_2, & \begin{pmatrix}
\bar{C} \\
\bar{A}
\end{pmatrix}
\end{pmatrix}
\]

(iii) Consider the criterion function \( L_{D}^{\varphi}(Y_T^1; U_T^1, \tau_D^0) \) in (2.85), where \( \tau_D^0 \) denotes the parameter vector corresponding to the slssDDLc parametrization in (4.83). Calculate a parameter update \( \Delta \tau_D^0 \), e.g., the gradient

\[
\Delta \tau_D^0 = \nabla L_{D}^{\varphi}(Y_T^1; U_T^1, \tau_D^0)
\]

and the corresponding system \((\bar{A}(\Delta \tau_D^0), \bar{B}(\Delta \tau_D^0), \bar{C}(\Delta \tau_D^0), \bar{D}(\Delta \tau_D^0))\); see (4.83). In terms of the corresponding realization matrix, the update \( \Delta R \) is then given by

\[
\Delta R = R(\Delta \tau_D^0) - R = \begin{pmatrix}
0 & \Delta \bar{C} \\
0 & \Delta \bar{A}
\end{pmatrix}
\]

If \( \| \Delta R \|_F < \varepsilon \), compute \((A, B, C, D, K) = \Delta Y_V(\bar{A}, \bar{C}, \bar{D})\) as given in (3.27) and stop. Here, \( \varepsilon \) is some a priori chosen positive real number.

(iv) Project the update \( \Delta R \) onto the tangent space \( Q_{(A, B, C, D)}^{BSA} \) to obtain \( \hat{R} \) which is given by:

\[
\hat{R} = R + R (R' \cdot \Delta R)_{skew} \in Q_{(A, B, C, D)}^{BSA}
\]

Here, \( X_{skew} \) denotes the skew symmetric part of the matrix \( X \), i.e., \( \frac{1}{2}(X - X^t) \).

(v) Determine the closest (in Frobenius norm) orthogonal realization matrix to \( \hat{R} \) which is again denoted by \( \hat{R} \) and is given by

\[
\hat{R} = UV^t
\]

Here, \( \hat{R} = U \Sigma V^t \) is a singular value decomposition of \( \hat{R} \). Go to (iii).

Let us first show that the statements in the definition above are indeed correct:
(i) It is evident that the transformation \( T = Q^{1/2} \) yields an output normal system.

(ii) As \((\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K})\) is output normal, \(\bar{A}'\bar{A} + \bar{C}'\bar{C} = I\) holds, i.e. the columns of the matrix \((\bar{C}', \bar{A}')\) are orthogonal. Hence, the singular value decomposition can indeed be chosen to be of the form \((4.92)\). As the original system is assumed to be strictly minimum phase, \(\bar{A}\) is stable, and the square system \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) is BSA (and hence minimal) according to lemma \((3.4.3)\).

(iii) Using the \texttt{s1sDDLC} parametrization, any update \(\Delta R\) will be in \(Q_{(\bar{A}, \bar{B}, \bar{C}, \bar{D})}^{c\perp} \) and hence of the form \((4.93)\).

(iv) The fact that the orthogonal projection (with respect to the Frobenius norm) of \(\Delta R\) onto the tangent space \(Q_{(\bar{A}, \bar{B}, \bar{C}, \bar{D})}^{BSA} \) is given by formula \((4.94)\) has been shown in lemma \((4.11.1)\) above.

(v) Formula \((4.95)\) has been shown in lemma \((4.11.2)\).

We now give a few remarks:

\textbf{Remark 4.11.2.} Clearly, \((iii)\) contains a termination condition: If the norm of the update \(\Delta R\) is less than some a priori specified positive real number \(\varepsilon > 0\), the algorithm terminates by computing the original system \((A, B, C, D, K) = \Delta_U(\bar{A}, \bar{B}, \bar{C}, \bar{D})\) as given in \((3.27)\). Of course, it is possible that the resulting system \((A, B, C, D, K)\) is non minimal or even that \(\Delta_U(\bar{A}, \bar{B}, \bar{C}, \bar{D})\) is not well defined.

\textbf{Remark 4.11.3.} Using \texttt{orthoDDLC}, the original optimization problem in the parameters \((\tau, \sigma)\) is translated to an optimization problem in the parameters \(\tau^0\) only, where \(\tau^0\) denotes the parameter vector for the inverse system using \texttt{s1sDDLC} as given in section \((4.10)\) above.

\textbf{Remark 4.11.4.} As \(\hat{R}\) corresponds to a system in \(Q_{(\bar{A}, \bar{B}, \bar{C}, \bar{D})}^{BSA}\), \(\hat{R}\) is orthogonal "in first order". Hence, \((v)\) can be viewed as a "higher order correction" to obtain a new orthogonal \(R = UV'\).

\textbf{Remark 4.11.5.} By lemma \((3.4.3)\), the resulting system \((\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})\) corresponding to the new orthogonal realization matrix \(R\) is BSA if it is minimal. Note that minimality is guaranteed on an open and dense subset of \(O(n + s)\) by lemma \((3.5.1)\).

\subsection{4.11.2 An illustrative example}

Consider again the case where \(n = s = 1\) and \(m = 0\): no exogenous inputs are present and thus \(l(z)\) vanishes. We commence from a minimal system \((\alpha, k, c) = (1/2, 1 - \sqrt{2}, 1/2) \in S_m(1)\). Then, the corresponding minimal inverse system is given by \((\bar{a}, \bar{\kappa}, \bar{c}) = (1/\sqrt{2}, 1 - \sqrt{2}, -1/2)\). The steps of definition \((4.11.1)\) are particularly simple in this case:

(i) The Lyapunov equation \((3.17)\) is given by \(1/2 q + \frac{1}{k} = q\), yielding \(q = \frac{1}{k}\), and \(q^{1/2} = \pm 1/\sqrt{2}\). Transforming the inverse system by \(q^{1/2} = -1/\sqrt{2}\) yields a new \((\bar{a}, \bar{\kappa}, \bar{c}) = (1/\sqrt{2}, -\sqrt{2}(1 - \sqrt{2}), 1/\sqrt{2})\).

(ii) The singular value decomposition in \((4.92)\) becomes

\[
\begin{pmatrix}
\bar{c} \\
\bar{a}
\end{pmatrix} = \begin{pmatrix}
1/\sqrt{2} & -1/\sqrt{2} \\
1/\sqrt{2} & 1/\sqrt{2}
\end{pmatrix} \begin{pmatrix}
1 \\
0
\end{pmatrix} \quad \text{and we can set} \quad R = \begin{pmatrix}
\bar{D} & \bar{C} \\
\bar{B} & \bar{A}
\end{pmatrix} = \begin{pmatrix}
-1/\sqrt{2} & 1/\sqrt{2} \\
1/\sqrt{2} & 1/\sqrt{2}
\end{pmatrix}
\]

(iii) The \texttt{s1sDDLC} parametrization is given in analogy to \((4.84)\) by

\[
\Delta_U \begin{pmatrix}
\alpha(\tau_D^0) \\
\beta(\tau_D^0) \\
\sigma(\tau_D^0) \\
\delta(\tau_D^0)
\end{pmatrix} = \Delta_U \begin{pmatrix}
1/\sqrt{2} + \tau_D^0 \\
1/\sqrt{2} \\
1/\sqrt{2} \\
-1/\sqrt{2}
\end{pmatrix}
\]

and for a calculated parameter update \(\Delta \tau_D^0\) we get \(\Delta R = \begin{pmatrix}
0 & 0 \\
\Delta \tau_D^0
\end{pmatrix}\).

Note that the update is in the \(\bar{a}\)-direction only; see the green straight line in figure \((4.5)\).

(iv) The projection of the update \(\Delta R\) onto the tangent space \(Q_{(\bar{A}, \bar{B}, \bar{C}, \bar{D})}^{BSA}\) is given by \((4.94)\). This projection step corresponds to the black thin line connecting the green line and the red line in figure \((4.5)\).
(v) The final step involving a singular value decomposition to determine the closest orthogonal realization matrix to $\hat{R}$ is given by (4.95); see the second black thin line connecting the red line and the unit circle in figure (4.5).

Clearly, orthoDDLC makes use of the sIsDDLC parametrization. For the topological and geometrical properties of sIsDDLC, see theorem (4.10.1) in section (4.10) above.

### 4.12 State-space parametrizations in a simple case

Here, we review the simple case where $n = s = 1$ and $m = 0$: no exogenous inputs are present and thus $l(z)$ vanishes. Clearly, $S(1) = \mathbb{R}^3$ and $S_m(1) = \{(a, k, c) \in \mathbb{R}^3 | (a, k, c) \text{ is minimal} \}$. This example has been addressed many times in this chapter. In (4.75) below we present all state-space parametrizations discussed in this thesis for the simple case just mentioned:

<table>
<thead>
<tr>
<th>Full Par</th>
<th>$\begin{pmatrix} a(a, k, c) \ k(a, k, c) \ c(a, k, c) \end{pmatrix}$</th>
<th>$\begin{pmatrix} a \ k \ c \end{pmatrix}$</th>
<th>$S_m(1) = \mathbb{R}^3 \rhd {(x, 0, z) \in \mathbb{R}, x \in \mathbb{R} } \cup \ldots {(x, y, 0) \in \mathbb{R}, y \in \mathbb{R}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Echelon (overlapping)</td>
<td>$\begin{pmatrix} a(\tau_a^{1/2}, \tau_b^{1/2}) \ k(\tau_a^{1/2}, \tau_b^{1/2}) \ c(\tau_a^{1/2}, \tau_b^{1/2}) \end{pmatrix}$</td>
<td>$\begin{pmatrix} \tau_a^{1/2} \ \tau_b^{1/2} \ 1 \end{pmatrix}$</td>
<td>$T_a = \mathbb{R}^2 \rhd {(x, 0) \in \mathbb{R}, x \in \mathbb{R}}$</td>
</tr>
<tr>
<td>LyapBAL (Ober)</td>
<td>$\begin{pmatrix} a(\sigma, b_c) \ k(\sigma, b_c) \ c(\sigma, b_c) \end{pmatrix}$</td>
<td>$\begin{pmatrix} 2\sigma - s^2 \ 2(\sigma + \delta) \ \pm 2\sqrt{2\sigma \delta + \sigma^2} \end{pmatrix}$</td>
<td>$T_\delta = \mathbb{R}^+ \times \mathbb{R}^+$</td>
</tr>
<tr>
<td>StochBAL (Ober)</td>
<td>$\begin{pmatrix} a(\rho, b_c) \ k(\rho, b_c) \ c(\rho, b_c) \end{pmatrix}$</td>
<td>$\begin{pmatrix} 2\rho - (1 + \sigma)\rho^2 \ 2(\rho + \delta) \ 2\sqrt{2\rho \delta + \rho^2} \end{pmatrix}$</td>
<td>$T_\delta^{(1)} = (0, 1) \times \mathbb{R}^+$</td>
</tr>
<tr>
<td>MinBAL (McGinnie)</td>
<td>$\begin{pmatrix} a(\sigma, b_c) \ k(\sigma, b_c) \ c(\sigma, b_c) \end{pmatrix}$</td>
<td>$\begin{pmatrix} 2\sigma - k^3 \ 2\sigma + k \end{pmatrix}$</td>
<td>$T_\delta^{(2)} = \mathbb{R}^+ \times \mathbb{R}^+$</td>
</tr>
<tr>
<td>DDLC</td>
<td>$\begin{pmatrix} a(\tau_a^{1/2}, \tau_b^{1/2}) \ k(\tau_a^{1/2}, \tau_b^{1/2}) \ c(\tau_a^{1/2}, \tau_b^{1/2}) \end{pmatrix}$</td>
<td>$\begin{pmatrix} a \ k \ c \end{pmatrix} + \begin{pmatrix} 1 \ 0 \ 0 \end{pmatrix} \begin{pmatrix} \tau_a^{1/2} \ \tau_b^{1/2} \end{pmatrix}$</td>
<td>$T_D = \mathbb{R}^2 \rhd {(x, -\frac{a}{c} \sqrt{k^2 + c^2}, x \in \mathbb{R}) } \cup \ldots {(x, -\frac{a}{c} \sqrt{k^2 + c^2}, x \in \mathbb{R}) }$</td>
</tr>
<tr>
<td>sIsDDLC</td>
<td>$\begin{pmatrix} a(\tau_a^{1/2}) \ k(\tau_a^{1/2}) \ c(\tau_a^{1/2}) \end{pmatrix}$</td>
<td>$\Delta^U \begin{pmatrix} \bar{a} \ \bar{b} \ \bar{c} \ \bar{D} \end{pmatrix}$</td>
<td>$T_\bar{D} = \mathbb{R} \rhd {-2}$</td>
</tr>
</tbody>
</table>

In case of orthoDDLC, the sIsDDLC parametrization is used where the initial system $(\bar{a}, \bar{B}, \bar{c}, D)$ is balanced stable allpass.

### 4.13 Figures

Here, we present all figures which have been discussed in this chapter:
Figure 4.2: The case $n = s = 1$ and $m = 0$ where $S(1) = \mathbb{R}^3$

(a) The plot shows the set $\varphi_\alpha(T_\alpha)$ for $\alpha = (1)$, coinciding with the set $\varphi_{\delta}^{(1)}(T_{\delta}^{(1)})$. The green affine plane (excluding the intersection with the plane given by $k = 0$) corresponds to the echelon state-space parameter space $T_\alpha$ and at the same time to the parameter space $T_{\delta}^{(1)}$ of the overlapping state-space parametrization.

(b) The plot shows the sets $\varphi_\delta(T_\delta)$ for $\gamma = (1)$ and $\gamma = (-1)$, respectively. The two sets $T_{\delta}$ of free parameters for Ober’s Lyapunov balanced canonical form correspond to the two open half plane segments $k = c, k > 0$ for $\gamma = (1)$ and $k = -c, k > 0$ for $\gamma = (-1)$, respectively. Note that $\alpha$ is restricted in both cases by $-1 < \alpha < 1$.

(c) The plot shows the sets $\varphi_{\delta}^{(1)}(T_{\delta}^{(1)})$ for $\gamma = (1)$ and $\gamma = (-1)$, respectively. The two sets $T_{\delta}^{(1)}$ of free parameters for Ober’s stochastically balanced canonical form correspond to the red and green surface for $\gamma = (1)$ and $\gamma = (-1)$, respectively. Note that $\alpha$ is restricted in both cases by $-1 < \alpha < 1$ and all systems are strictly minimum phase, too.

(d) The plot shows the sets $\varphi_{\delta}^{(2)}(T_{\delta}^{(2)})$ for $\gamma = (1)$ and $\gamma = (-1)$, respectively. The two sets $T_{\delta}^{(2)}$ of free parameters for McGinnie’s minimum phase balanced canonical form correspond to the red and green surface for $\gamma = (1)$ and $\gamma = (-1)$, respectively. Note that $\alpha$ is restricted in both cases by $-1 < \alpha < 1$ and all systems are strictly minimum phase, too.
Figure 4.3: The DDLC construction in $S(1) = \mathbb{R}^3$ (a) and projection onto the $(k, c)$-plane (b). The affine plane corresponding to $T_D$ is shown for the minimal initial system $(a, k, c) = (0, 3, 1)$ (dark point). Note the exclusion of the two straight lines given by $k = 0$ and $c = 0$, respectively, and the straight line where $k$-equivalence classes (such as the thin hyperbola in the figure) touch $T_D$.

Figure 4.4: The skDDLC construction in the $(\bar{a}, B, \bar{c})$-space (a), neglecting the scalar $D$ and projection onto the $(B, \bar{c})$-plane (b). The straight line corresponding to $T_D^0$ is shown for the minimal system $(\bar{a}, B, \bar{c}) = (0, 1, 1)$ (dark point). Note the exclusion of the point $(\bar{a}(-2), B(-2), \bar{c}(-2)) = (-2, 1, 1)$ corresponding to $T_D^0 = -2$.
Figure 4.5: The orthoDDLC construction in the \((\bar{a}, B, \bar{c})\)-space ((a), neglecting the scalar \(D\)) and projection onto the \((B, \bar{c})\)-plane (b). The green straight line corresponding to \(T_{D}^{\bar{c}}\) is shown for the minimal system \((\bar{a}, B, \bar{c}, D) = (1/\sqrt{2}, 1/\sqrt{2}, 1/\sqrt{2}, -1/\sqrt{2})\) (black point). The red straight line corresponds to the tangent space \(\mathcal{G}^{BSA}_{(\bar{a}, B, \bar{c}, D)}\) as described in remark (3.5.2). The two black thin lines represent the projection step described in (iv) in definition (4.11.1) and the orthogonal best approximation described in (v) in (4.11.1), respectively. This yields the new BSA system given by the blue point.

Figure 4.6: (a) The case \(n = s = 1\) and \(m = 0\) where \(S(1) = \mathbb{R}^3\): The dark area indicates the projection of the image set \(F(GL(n) \times T_D)\) onto the \((k, c)\)-plane, where the mapping \(F\) is given in (4.79). Clearly, the \(k\)-axis and the \(c\)-axis are not included in this set. Note that \(F(GL(n) \times T_D)\) is not open in this case.
Chapter 5

Numerical considerations

In this chapter we will consider computationally relevant issues in connection with maximum likelihood estimation. In section (5.1), extended Kalman (or prediction error) filters are presented which are used for calculating the derivatives of the estimated innovations. Formulae for the evaluation of the asymptotic form $L(\tau, \sigma)$ of the likelihood function, its gradient and (approximate) Hessian are then derived in section (5.2), where the evaluation is not necessarily at the parameter values corresponding to the true transfer function. For reasons of completeness, the well known calculation of the analogous quantities in the finite sample case is presented in the subsequent section (5.3). In any case, the formulae are given in terms of the full state-space parametrization, such that the derivatives of the state-space matrices with respect to the parameters are needed; see section (5.4). Section (5.5) presents a short overview of some traditional optimization techniques, motivating the importance of the Fisher information matrix also in relation to computational aspects of maximum likelihood estimation. Numerical properties of parameter estimation algorithms in connection with various parametrizations are then treated in section (5.6). In particular, DDLC and orthoDDLC are compared in this respect. A few rather obvious questions for future research are given in section (5.7), and section (5.8) presents a number of plots.

5.1 Preliminaries on extended prediction error filters

The following discussion will be restricted to the state-space framework. Of course, an analogous treatment could also be given for ARMAX representations.

The state-space matrices $(A, B, C, D, K)$ are assumed to be parametrized using any of the parametrizations considered in chapter 4. That is, we consider the case

$$(A, B, C, D, K) = (A(\tau), B(\tau), C(\tau), D(\tau), K(\tau)), \quad \Sigma = \Sigma(\sigma)$$

where $\sigma \in \mathbb{R}^{(s+1)/2}$ again denotes the vector of off- and above-diagonal elements of $\Sigma$ and $\tau$ is a vector of free parameters. We will, however, most often omit the arguments $\tau$ in all state-space matrices: $(A, B, C, D, K)$ and $(A_0, B_0, C_0, D_0, K_0)$ will be used instead of $(A(\tau), B(\tau), C(\tau), D(\tau), K(\tau))$ and $(A(\tau_0), B(\tau_0), C(\tau_0), D(\tau_0), K(\tau_0))$, respectively. The same holds true for the covariance matrix, i.e. $\Sigma$ will be used instead of explicitly writing $\Sigma(\sigma)$. The symbols $A$, $B$ etc. denote the derivatives of the corresponding matrices with respect to one particular component $\tau_i$ of the (system!) parameter vector $\tau$. Sometimes also the notation $A^{[k]}$, $B^{[k]}$ etc. will be used in order to explicitly denote the derivative with respect to the $k.$ component $\tau_k$ of the parameter vector $\tau$. If $L = L(\tau, \sigma)$ is a sufficiently smooth scalar function of $(\tau, \sigma)$ or $\nu = \nu(\tau, \sigma)$ is a vector depending sufficiently smoothly on $(\tau, \sigma)$, then $L$ or $\nu$ are defined analogously. A superscript $(k)$ will be used as a shorthand notation for second derivatives with respect to the $k.$ and $l.$ entry in $\tau$.

Assume that the observed process $(y_t)$ can be be written as


\[ x_{t+1}^0 = A_0 x_t^0 + B_0 u_t + K_0 \varepsilon_t \]

\[ y_t = C_0 x_t^0 + D_0 u_t + \varepsilon_t \] (5.1)

where \((A_0, B_0, C_0, D_0, K_0)\) is the state-space system corresponding to the true parameter value \(\tau_0\); as usual \(\varepsilon_t\) is a white noise process with covariance matrix \(\mathbb{E} \varepsilon_t \varepsilon_t' = \Sigma_0\).

Our aim is to evaluate the likelihood and its asymptotic form as well as first and (approximate) second derivatives for arbitrary parameter values \((\tau, \sigma)\), i.e. for some \((A, B, C, D, K)\) and some \(\Sigma\). Hence, we consider a model of the form

\[ x_{t+1} = Ax_t + Bu_t + Ke_t \]

\[ y_t = Cx_t + Du_t + e_t \] (5.2)

where \((e_t)\) is again white noise with covariance matrix \(\mathbb{E} e_t e_t' = \Sigma\). The state-space model with inputs \((u_t', y_t')'\) and output \(e_t\) is given in (2.75) and is repeated here for ease of reference:

\[ x_{t+1} = (A - KC)x_t + (B - KD)u_t + Ky_t \]

\[ e_t = -Cx_t - Du_t + y_t \] (5.3)

Note that only \(x_t\) and \(e_t\) depend on the parameters \((\tau, \sigma)\), in contrast to \(y_t, x_t^0, \varepsilon_t\) and \(u_t\). Hence, differentiation of (5.3) with respect to one particular component \(\tau_i\) of the parameter vector \(\tau\) yields

\[ \dot{x}_{t+1} = (\tilde{A} - \tilde{K}C - \tilde{K}\tilde{C})x_t + (A - KC)\dot{x}_t + (B - KD - K\tilde{D})u_t + \tilde{K}y_t \]

\[ \dot{e}_t = -\tilde{C}x_t - C\dot{x}_t - \tilde{D}u_t \] (5.4)

This is the state-space system generating the derivatives \(e_t\) from \((u_t', y_t')'\). Note that by combining the systems (5.3) and (5.4), we get a system generating both \((e_t', \dot{e}_t')'\) from the observed quantities \((u_t', y_t')'\):

\[
\begin{pmatrix}
  x_{t+1} \\
  \dot{x}_{t+1} \\
  e_t \\
  \dot{e}_t
\end{pmatrix} =
\begin{pmatrix}
  \tilde{A} \\
  \tilde{A} - \tilde{K}C - \tilde{K}\tilde{C} \\
  -\tilde{C} \\
  0
\end{pmatrix}
\begin{pmatrix}
  x_t \\
  \dot{x}_t \\
  x_t \\
  \dot{x}_t
\end{pmatrix} +
\begin{pmatrix}
  \tilde{B} \\
  B - KD - K\tilde{D} \\
  0 \\
  -\tilde{D}
\end{pmatrix}
\begin{pmatrix}
  u_t \\
  u_t \\
  u_t \\
  u_t
\end{pmatrix} +
\begin{pmatrix}
  K \\
  Ky_t
\end{pmatrix}
\begin{pmatrix}
  y_t \\
  y_t
\end{pmatrix}
\] (5.5)

As usual, \(\tilde{A} = (A - KC)\) and \(\tilde{B} = (B - KD)\); see (2.74). This realization need not be minimal, however, we will use it in section (5.3) for the computations involving the finite sample likelihood. Note that the derivative \(\dot{e}_t\) does not depend on \(y_t\), but only on past \(y_{t-j}^u, j \geq 1\).

**Remark 5.1.1.** Note that for a parameter vector \(\tau\) consisting of \(d\) components, we can easily augment (5.5) to a system of "state dimension" \((d + 1)n\) which generates \(e_t\) and all derivatives \(e_t^{(1)}, \ldots, e_t^{(d)}\). It is shown in (Gupta and Mehr, 1974) that this system is in general non minimal and that the corresponding transfer function has order less than or equal to \((m + s + 1)n\). The authors also propose an algorithm for reducing the state dimension, but we will not discuss this here.

**Remark 5.1.2.** An alternative way to calculate the derivatives \(e_t^{(1)}, \ldots, e_t^{(d)}\) is to consider the problem in the frequency domain; see (Akaike, 1973). Here, one can then make use of numerically efficient techniques such as Fast Fourier Transforms. Again, we will not discuss this approach here.
For the case when we are concerned with the asymptotic form of the likelihood, we may link the systems (5.3) and (5.4) through \( y_t \) with the true data generating state-space system (5.1) in order to get a realization generating both \((e'_t, \tilde{e}'_t)\)' from the quantities \((u'_t, \varepsilon'_t)\)'. Doing this, (5.3) becomes

\[
x_{t+1} = (A - KC)x_t + (B - KD)u_t + KC_0x_0 + KD_0u_0 + \kappa_t
\]

and (5.4) becomes

\[
\dot{x}_{t+1} = (\dot{A} - \dot{K}C - \dot{K}\dot{C})x_t + (A - KC)\dot{x}_t + (\dot{B} - \dot{K}D - \dot{K}\dot{D})u_t + \dot{K}C_0x_0 + \dot{K}D_0u_0 + \dot{\kappa}_t
\]

and by combining the systems (5.6) and (5.7), we get a system generating both \((e'_t, \tilde{e}'_t)\)' from \((u'_t, \varepsilon'_t)\)'

\[
\begin{pmatrix}
  x_{t+1}^0 \\
  \dot{x}_{t+1}^0
\end{pmatrix}
= \begin{pmatrix}
  A_0 & 0 \\
  \dot{K}C_0 & A
\end{pmatrix}
\begin{pmatrix}
  x_t^0 \\
  \dot{x}_t^0
\end{pmatrix}
+ \begin{pmatrix}
  B_0 \\
  \dot{B} + KD_0
\end{pmatrix}
\begin{pmatrix}
  u_t \\
  \dot{u}_t
\end{pmatrix}
+ \begin{pmatrix}
  K_0 \\
  \dot{K}
\end{pmatrix}
\begin{pmatrix}
  \varepsilon_t \\
  \dot{\varepsilon}_t
\end{pmatrix}
\]

In the sequel, it will sometimes be necessary to explicitly denote the component of the parameter vector \(\tau\) with respect to which differentiation is performed. Hence, in (5.8) we may replace all "dots" in \(\dot{x}_t, \dot{e}_t, \dot{A}\) etc. by a superscript \((k)\) or \((l)\) to denote differentiation with respect to \(\tau_k\) and \(\tau_l\), respectively. Using an obvious notation for the case of differentiation with respect to \(\tau_k\), for instance, we will abbreviate (5.8) to

\[
\begin{pmatrix}
  x_{t+1,k}^0 \\
  \dot{x}_{t+1,k}^0
\end{pmatrix}
= \begin{pmatrix}
  A_k & 0 \\
  \dot{K}C_0 & A
\end{pmatrix}
\begin{pmatrix}
  x_{t,k}^0 \\
  \dot{x}_{t,k}^0
\end{pmatrix}
+ \begin{pmatrix}
  B_0 \\
  \dot{B} + KD_0
\end{pmatrix}
\begin{pmatrix}
  u_t \\
  \dot{u}_t
\end{pmatrix}
+ \begin{pmatrix}
  K_0 \\
  \dot{K}
\end{pmatrix}
\begin{pmatrix}
  \varepsilon_t \\
  \dot{\varepsilon}_t
\end{pmatrix}
\]

where \(A_k \in \mathbb{R}^{3n \times 3n}\), \(\dot{B}_k \in \mathbb{R}^{3n \times m}\), \(\dot{C}_k \in \mathbb{R}^{2s \times 3n}\), \(\dot{D}_k \in \mathbb{R}^{2s \times m}\), \(\dot{K}_k \in \mathbb{R}^{3n \times s}\) and \(\tilde{x}_{t,k} = (x'_t, z'_t, \varepsilon'_t)^\prime\). The corresponding transfer function will be denoted by \((\tilde{I}_k, \tilde{K}_k)\).

The matrices \(\hat{A} \in \mathbb{R}^{2n \times 2n}\), and \(\bar{K} \in \mathbb{R}^{2n \times s}\) are submatrices of \(A_k\) and \(K_k\):

\[
\hat{A} = \begin{pmatrix}
  A_0 & 0 \\
  \bar{K}C_0 & \bar{A}
\end{pmatrix}
\quad \bar{K} = \begin{pmatrix}
  K_0 \\
  \bar{K}
\end{pmatrix}
\]

Note that \(\hat{A}\) and \(\bar{K}\) do not contain derivatives of system matrices with respect to any parameters!

In analogy to (1.12) the covariance between \((e'_t, \tilde{e}'_t)^\prime\) and \((e'_t, \tilde{e}'_t)^{\prime \prime}\)' is given by

\[
\mathbb{E}\left\{ \begin{pmatrix}
  e_t \\
  \tilde{e}_t
\end{pmatrix}
\begin{pmatrix}
  e_t^{\prime \prime} \\
  \tilde{e}_t^{\prime \prime}
\end{pmatrix}
\right\}
= \mathbb{E}\left\{ \begin{pmatrix}
  e_t \\
  \tilde{e}_t
\end{pmatrix}
\begin{pmatrix}
  e_t \\
  \tilde{e}_t
\end{pmatrix}
\right\}'
= \hat{C}_k \hat{P}^{(k)}(k) \bar{C}_l' + \begin{pmatrix}
  I_s \\
  0
\end{pmatrix}
\Sigma_0 \begin{pmatrix}
  I_s \\
  0
\end{pmatrix}
\]

where

\[
\hat{P}^{(k)} = \hat{A}_k \hat{P}^{(k)}(k) \hat{A}_l' + \bar{K}_k \Sigma_0 \bar{K}_l'
\]
Note the lower block triangular structure of the matrices \( \hat{\mathbf{A}}_k \) and \( \hat{\mathbf{A}}_t \); the upper left block matrix \( \hat{\mathbf{A}} \) appears in both \( \hat{\mathbf{A}}_k \) and \( \hat{\mathbf{A}}_t \). We will now partition \( \mathbf{P}^{(kl)} \) accordingly to get

\[
\mathbf{P}^{(kl)} = \begin{pmatrix}
\mathbf{P}_{11}^{(kl)} & \mathbf{P}_{12}^{(kl)} \\
\mathbf{P}_{21}^{(kl)} & \mathbf{P}_{22}^{(kl)}
\end{pmatrix}
\]  

(5.13)

Making use of this decomposition, (5.12) can be decomposed into a set of four nested Sylvester equations:

\[
\begin{pmatrix}
\mathbf{P}_{11}^{(kl)} & \mathbf{P}_{12}^{(kl)} \\
\mathbf{P}_{21}^{(kl)} & \mathbf{P}_{22}^{(kl)}
\end{pmatrix} = \begin{pmatrix}
(K^{(k)} \mathbf{C}_0, A^{(k)} - K^{(k)} \mathbf{C} - K^{(k)} \mathbf{C}^{-1}) & 0 \\
0 & (K^{(l)} \mathbf{C}_0, A^{(l)} - K^{(l)} \mathbf{C} - K^{(l)} \mathbf{C}^{-1})
\end{pmatrix}
\begin{pmatrix}
\mathbf{P}_{11}^{(kl)} & \mathbf{P}_{12}^{(kl)} \\
\mathbf{P}_{21}^{(kl)} & \mathbf{P}_{22}^{(kl)}
\end{pmatrix} + 
\begin{pmatrix}
\hat{\mathbf{K}} \Sigma_0 \hat{\mathbf{K}}' \\
K^{(k)} \Sigma_0 K^{(l)}
\end{pmatrix}
\begin{pmatrix}
\mathbf{P}_{11}^{(kl)} & \mathbf{P}_{12}^{(kl)} \\
\mathbf{P}_{21}^{(kl)} & \mathbf{P}_{22}^{(kl)}
\end{pmatrix} + 
\begin{pmatrix}
\mathbf{P}_{11}^{(kl)} & \mathbf{P}_{12}^{(kl)} \\
\mathbf{P}_{21}^{(kl)} & \mathbf{P}_{22}^{(kl)}
\end{pmatrix}
\]  

Multiplying out each block, i.e. using (A.47) in the appendix (A.5), and putting

\[
\Delta_k = (K^{(k)} \mathbf{C}_0, A^{(k)} - K^{(k)} \mathbf{C} - K^{(k)} \mathbf{C}^{-1}) \in \mathbb{R}^{n \times 2n}
\]

\[
\Delta_l = (K^{(l)} \mathbf{C}_0, A^{(l)} - K^{(l)} \mathbf{C} - K^{(l)} \mathbf{C}^{-1}) \in \mathbb{R}^{n \times 2n}
\]

we get

\[
\mathbf{P}_{11}^{(kl)} = \hat{\mathbf{A}} \mathbf{P}_{11}^{(kl)} \mathbf{A}' + \hat{\mathbf{K}} \Sigma_0 \hat{\mathbf{K}}' \implies \mathbf{P}_{11}^{(kl)} = \hat{\mathbf{P}}
\]  

(5.16)

Hence, \( \hat{\mathbf{P}} \) is independent of any differentiation. For the (1, 2) and the (2, 1) block we get

\[
\begin{align*}
\mathbf{P}_{12}^{(kl)} &= \hat{\mathbf{A}} \mathbf{P}_{12}^{(kl)} \mathbf{A}' + \hat{\mathbf{A}} \Delta_k \hat{\mathbf{A}}' + \hat{\mathbf{K}} \Sigma_0 K^{(l)} \\
\mathbf{P}_{21}^{(kl)} &= \hat{\mathbf{A}} \mathbf{P}_{21}^{(kl)} \mathbf{A}' + \Delta_k \hat{\mathbf{A}}' + K^{(k)} \Sigma_0 \hat{\mathbf{K}}'
\end{align*}
\]  

(5.17)

Note that \( \hat{\mathbf{A}} \) and \( \mathbf{A} \) are stable, and therefore the solutions \( \mathbf{P}_{12}^{(kl)} \) and \( \mathbf{P}_{21}^{(kl)} \) can be written as

\[
\begin{align*}
\mathbf{P}_{12}^{(kl)} &= \sum_{j=1}^{\infty} \hat{\mathbf{A}}^{-j} \left( \hat{\mathbf{A}} \Delta_k + \hat{\mathbf{K}} \Sigma_0 K^{(l)} ight) \hat{\mathbf{A}}^{-1} \\
\mathbf{P}_{21}^{(kl)} &= \sum_{j=1}^{\infty} \hat{\mathbf{A}}^{-j} \left( \Delta_k \hat{\mathbf{A}} + K^{(k)} \Sigma_0 \hat{\mathbf{K}} ight) \hat{\mathbf{A}}^{-1}
\end{align*}
\]  

(5.18) (5.19)

Finally, for the (2, 2) block, the following equation is obtained:

\[
\begin{align*}
\mathbf{P}_{22}^{(kl)} &= \mathbf{A} \mathbf{P}_{22}^{(kl)} \mathbf{A}' + \Delta_k \hat{\mathbf{P}} \Delta_k' + \Delta_k \mathbf{P}_{12}^{(kl)} \mathbf{A}' + \mathbf{A} \mathbf{P}_{21}^{(kl)} \Delta_k' + K^{(k)} \Sigma_0 K^{(l)} \\
&= \mathbf{A} \mathbf{P}_{22}^{(kl)} \mathbf{A}' + \Delta_k \mathbf{P}_{12}^{(kl)} \mathbf{A}' + \mathbf{A} \mathbf{P}_{21}^{(kl)} \Delta_k' + K^{(k)} \Sigma_0 K^{(l)}
\end{align*}
\]  

(5.20)
5.2 The asymptotic likelihood $L(\tau, \sigma)$

The asymptotic likelihood (2.87) is repeated for ease of reference, introducing the trace operator:

$$ L(\theta) = \log \det \Sigma + tr \Sigma^{-1} E e_t e_t' = k^{-1}(z) (k_0(z)\varepsilon_t + (l(z) - l_0(z))u_t) $$

(5.21)

where $\theta$ refers to $(l, k, \sigma(\Sigma))$ in this case. In subsection (5.2.1) we will discuss how to evaluate this function for various parametrizations, and in subsections (5.2.2) and (5.2.3) we treat the evaluation of the gradient and (an approximation to) the Hessian.

5.2.1 Evaluating $L(\tau, \sigma)$ for a particular parametrization

We will consider the asymptotic likelihood (5.21) to be a function of the parameters $\tau$ in $(A, B, C, D, K)$ and $\sigma$ in $\Sigma$, ensuring that $\pi(A, B, C, D, K) = (l, k) \in M_{\text{sym}}(n)$ and $\Sigma = \Sigma' > 0$:

$$ L(\tau, \sigma) = L(A, B, C, D, K, \Sigma) = L((A, B, C, D, K, \Sigma), (A_0, B_0, C_0, D_0, K_0, \Sigma_0)) $$

(5.22)

where the last expression explicitly clarifies the dependence of the asymptotic likelihood on the true parameters $\tau_0$ and $\sigma_0$.

In order to evaluate $E e_t e_t'$, we only need to consider the first output of system (5.8). The covariance matrix of $e_t$ is given in the $(1, 1)$ block of (5.11), i.e. by

$$ \text{Cov}(e_t, e_t) = \begin{pmatrix} (C_0, -C) & 0 \end{pmatrix} \tilde{P} \begin{pmatrix} (C_0, -C)' \end{pmatrix} + \Sigma_0 $$

$$ = \begin{pmatrix} C_0 & -C \end{pmatrix} \tilde{P} \begin{pmatrix} C_0' \ -C' \end{pmatrix} + \Sigma_0 $$

where $\tilde{P}$ can be any $\tilde{P}^{(kl)}$ because the upper left block will always be $\tilde{P}$. However, we need the (non-centered) second moments of $e_t$. From (5.8) we see that the transfer function from $u_t$ to $e_t$, $l_{ue}(z)$, say, is of the following form

$$ l_{ue}(z) = \begin{pmatrix} C_0 & -C \end{pmatrix} (z^{-1}I - \hat{A})^{-1} \begin{pmatrix} B_0 \\
B + KD_0 \end{pmatrix} + (D_0 - D) = \begin{bmatrix} A_0 & 0 \\
KC_0 & \hat{A} \\
C_0 & -C \end{bmatrix} \begin{bmatrix} B_0 \\
B + KD_0 \\
D_0 - D \end{bmatrix} $$

(5.23)

In analogy to (2.61) in section (2.3.1), the value of the asymptotic likelihood is thus given by

$$ L(\tau, \sigma) = \log \det \Sigma + tr \Sigma^{-1} \left\{ \begin{pmatrix} C_0 & -C \end{pmatrix} \tilde{P} \begin{pmatrix} C_0' \\
-C' \end{pmatrix} + \Sigma_0 + \int_0^\pi l_{ue}(e^{i\omega})dF_u(\omega)l_{ue}(e^{i\omega})^* \right\} $$

(5.24)

where $\tilde{P} \in \mathbb{R}^{2n \times 2n}$ is given in (5.16) and the transfer function $l_{ue}(z)$ is given in (5.23).

5.2.2 Evaluating $\nabla L(\tau, \sigma)$ for a particular parametrization

Now we can turn to the problem of calculating the gradient of the asymptotic likelihood function. Note that an analogous derivation of the gradient can be found in (Scherrer, 2002) for the case without exogenous inputs where additionally $(A, K, C)$ is a lower order approximant of the true system $(A_0, K_0, C_0)$. From (5.21) we immediately obtain

$$ \dot{L}(\tau, \sigma) = 2tr \Sigma^{-1} E e_t e_t' $$

(5.25)
In order to evaluate $\mathbb{E} \hat{e}_t e'_t$, we first need to consider the covariance matrix of $\hat{e}_t$ and $e_t$ which is given in the $(2,1)$ block of (5.11), i.e. by

$$
Cov(\hat{e}_t, e_t) = \begin{pmatrix} (0, -\hat{C}) & -\hat{C} \end{pmatrix} \hat{P} \begin{pmatrix} (C_0, -C') \\ 0 \end{pmatrix}
$$

$$
= \begin{pmatrix} (0, -\hat{C}) \hat{P} - C \hat{P}_{21} - (0, -\hat{C}) \hat{P}_{12} - C \hat{P}_{22} \end{pmatrix} \begin{pmatrix} (C_0, -C') \\ 0 \end{pmatrix}
$$

$$
= (0, -\hat{C}) \hat{P} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right) - C \hat{P}_{21} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right)
$$

If we are interested in the derivative with respect to $\tau_k$, then $\hat{P} = \hat{P}^{(kl)}$ has to be chosen, where $l$ can be taken to be the index of any other component of $\tau$.

Again, we actually need the (non-centered) second moments $\mathbb{E} \hat{e}_t e'_t$. From (5.8) we see that the transfer function from $u_t$ to $\hat{e}_t$, $l_{ue}(z)$, say, is of the following form

$$
l_{ue}(z) = \begin{pmatrix} 0 & -\hat{C} \end{pmatrix} (z^{-1} I - \hat{A}_k)^{-1} \hat{B}_k - \hat{D}
$$

$$
= \begin{bmatrix}
\hat{A}_0 & 0 \\
\hat{K} \hat{C}_0 & \hat{A}
\end{bmatrix}
\begin{bmatrix}
\hat{B}_0 \\
\hat{B} + \hat{K} \hat{D}_0
\end{bmatrix} + \int_{-\pi}^{\pi} l_{ue}(e^{i\omega}) dF_u(\omega) l_{ue}(e^{i\omega})^*
$$

(5.26)

In total, we get

$$
\hat{L}(\tau, \sigma) = 2 \cdot tr \Sigma^{-1} \left\{ (0, -\hat{C}) \hat{P} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right) - C \hat{P}_{21} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right) + \int_{-\pi}^{\pi} l_{ue}(e^{i\omega}) dF_u(\omega) l_{ue}(e^{i\omega})^* \right\}
$$

(5.27)

Now, $\hat{P}$ is independent of the differentiation whereas the solution $\hat{P}_{21} = \hat{P}_{21}^{(kl)}$, which is given in (5.17), clearly depends on the component with respect to which we differentiate. However, using the representation (5.19) of $\hat{P}_{21}^{(kl)}$, we can simplify the second term in (5.27):

$$
-2 \cdot tr \Sigma^{-1} C \hat{P}_{21} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right) = -2 \cdot tr \Sigma^{-1} C \sum_{j=1}^{\infty} \hat{A}^{j-1} \left( \Delta_k \hat{P} \hat{A}' + K(k) \Sigma_0 \hat{K}' \right)^{j-1} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right)
$$

$$
= -2 \cdot tr \left( \Delta_k \hat{P} \hat{A}' + K(k) \Sigma_0 \hat{K}' \right) \sum_{j=1}^{\infty} \hat{A}^{j-1} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right) \Sigma^{-1} C \hat{A}^{j-1}(5.28)
$$

Note that $X \in \mathbb{R}^{2n \times n}$ does not depend on the parameter with respect to which we differentiate. This yields

$$
\hat{L}(\tau, \sigma) = 2 \cdot tr \Sigma^{-1} \left\{ (0, -\hat{C}) \hat{P} \left( \begin{pmatrix} C_0' \\ -C' \end{pmatrix} \right) + \int_{-\pi}^{\pi} l_{ue}(e^{i\omega}) dF_u(\omega) l_{ue}(e^{i\omega})^* \right\} - 2 \cdot tr \left\{ \left( \Delta_k \hat{P} \hat{A}' + K(k) \Sigma_0 \hat{K}' \right) X \right\}
$$

(5.29)

where $\hat{P} \in \mathbb{R}^{2n \times 2n}$ is given in (5.16), $X \in \mathbb{R}^{2n \times n}$ is the solution of the Sylvester equation in (5.28), $\Delta_k \in \mathbb{R}^{n \times 2n}$ is given in (5.14) and $l_{ue}(z)$ and $l_{ue}(z)$ are given in (5.26) and (5.23), respectively.
5.2. THE ASYMPTOTIC LIKELIHOOD $L(\tau, \sigma)$

**Remark 5.2.1.** We have implemented formula (5.29) for the computation of the gradient of the asymptotic likelihood function in MATLAB (as well as formula (5.24) for the computation of the value of the asymptotic likelihood itself). Of course, such a computation also requires formulae for the derivatives of the state-space matrices with respect to the system parameters used. For the state-space parametrizations treated in this thesis, this will be discussed in section (5.4) below. In plot (a) in figure (5.4), a gradient step for the asymptotic likelihood $L((a_0, k_0, \sigma_0^2), (a_2, k_0, \sigma_0^2))$ is shown for various parametrizations in the case when $n = s = 1$ and $m = 0$. We want to make two remarks here:

(i) The implementation of formulae (5.24) and (5.29) makes it possible to run gradient-type algorithms (see remark (5.2.2) below) for different parametrizations to find the minimum of the asymptotic likelihood function. Note that one might think this to be a nice test setting for comparing different parametrizations, as the global minimizer of the asymptotic likelihood is known under mild conditions; see (2.88) and the discussion below (2.88). However, for a comparison of gradient-type algorithms, this is (in a certain sense) not the case as will be briefly motivated in (ii).

(ii) Let us consider a simple "steepest descent" algorithm where the parameter update is chosen to be some (positive) scalar multiple of the negative gradient. Such an update direction in the parameter space corresponds to a certain update "direction" e.g. in the space $M_{\text{sup}}(n)$ of transfer functions. Plot (c) in figure (5.4) illustrates the fact that the update "directions" in $M_{\text{sup}}(n)$ can be different for different parametrizations, making it principally difficult to assess the influence of the particular parametrization on, e.g. the convergence speed of the steepest descent algorithm. In other words, it is not possible to distinguish in a finite precision setting (as it occurs for the actual implementation of the algorithm on a computer) whether poor performance is due to poor finite precision properties of the parametrization used (numerical ill-conditioning) or due to the fact that the search algorithm already takes a longer path (that might even lead to a different local minimum) in the perfect precision setting. For a further discussion of related issues, we refer to (Hanzon, 1989) and (Peeters, 1994); see also remark (5.2.3) below.

**Remark 5.2.2.** We designed a graphical user interface (GUI) in MATLAB for estimation of state-space systems using the parametrizations discussed in this thesis together with some of the algorithms discussed in section (5.5) below. In plot (d) of figure (5.2), the evolution of the value of the asymptotic likelihood function is shown for the full state-space parametrization when a simple gradient search as in (5.61) in section (5.5) below (together with a simple line search) is used. Clearly, formulae (5.24) and (5.29) are used here.

### 5.2.3 Evaluating an approximate Hessian of $L(\tau, \sigma)$ for a particular parametrization

Turning to the Hessian of the asymptotic likelihood, we immediately get from (5.25)

$$L^{(kl)}(\tau, \sigma) = 2tr\Sigma^{-1}E_{t}^{(k)}e_{t}' + 2tr\Sigma^{-1}E_{t}^{(kl)}e_{t}'$$

(5.30)

First, we want to stress the fact that at the true parameter values, i.e. for $\tau = \tau_0$ and $\Sigma = \Sigma_0$, the second term in (5.30) vanishes. For simplicity, we will therefore neglect this term at any $\tau$ and $\Sigma$ in any case, being aware of the fact that the resulting quantity will just be an approximation to the true Hessian. In a sense, this approach corresponds to the computation of the Gauss-Newton approximation of the Hessian, yet not for the finite sample likelihood $L_T(\tau, \sigma)$, but for its asymptotic form $L(\tau, \sigma)$; see section (5.5) for a further discussion of this statement.

Now, the covariance of $e_{t}^{(k)}$ and $e_{t}^{(l)}$ is given in the (2,2) block of (5.11), i.e. by

$$\text{Cov}(e_{t}^{(k)}, e_{t}^{(l)}) = \begin{pmatrix} (0, -C^{(k)}) & -C \end{pmatrix} \begin{pmatrix} \hat{P}_{12}^{(kl)} & \hat{P}_{12}^{(kl)} \\ \hat{P}_{21}^{(kl)} & \hat{P}_{22}^{(kl)} \end{pmatrix} \begin{pmatrix} (0, -C^{(l)})' \\ -C' \end{pmatrix}$$

---

1 Clearly, this is not the case if the finite sample likelihood is considered instead.
In order to obtain the (non-centered) second moments \( \mathbb{E} [e_t^{(k)} e_t^{(l)'}, \mathbb{E} [u_t^{(k)} u_t^{(l)'}] \), let us denote the transfer function from \( u_t \) to \( e_t^{(k)} \) by \( i_T^{(k)}(z) \) and the transfer function from \( u_t \) to \( e_t^{(l)} \) by \( i_T^{(l)}(z) \); these transfer functions are given in (5.26), where all "dots" have to be replaced by differentiation with respect to \( \tau_k \) and \( \tau_l \), respectively. Hence, we finally get

\[
L^{(k)}(\tau, \sigma) = 2 \cdot \text{tr} \Sigma^{-1} \left\{ \begin{array}{c}(0, C^{(k)}) \\ \frac{\hat{P}_{11}^{(k)}}{\hat{P}_{21}^{(k)}} \\ \frac{\hat{P}_{12}^{(k)}}{\hat{P}_{22}^{(k)}} \end{array} \right\} + \int_{-\infty}^{\infty} i_T^{(k)}(e)^i dF_{\sigma}(\sigma) \int_{-\infty}^{\infty} i_T^{(l)}(e)^i dF_{\sigma}(\sigma) \]  

(5.31)

where \( \hat{P} \in \mathbb{R}^{n \times 2n} \) is given in (5.16), \( \hat{P}_{12}^{(k)} \in \mathbb{R}^{n \times n} \) and \( \hat{P}_{21}^{(k)} \in \mathbb{R}^{n \times n} \) are given in (5.17), \( \hat{P}_{22}^{(k)} \in \mathbb{R}^{n \times n} \) is given in (5.20) and the transfer functions \( i_T^{(k)}(z) \) and \( i_T^{(l)}(z) \) are given in (5.26), where all "dots" have to be replaced by differentiation with respect to \( \tau_k \) and \( \tau_l \), respectively.

Note that the 'A-matrices' in all Sylvester equations are the same for differentiation with respect to any component of the parameter vector. This in turn means that for different partial differentiations each of the Sylvester equations in (5.16), (5.17) and (5.20) correspond to a system of linear equations with different "right hand sides" but the same "coefficient matrix".²

Remark 5.2.3. We have implemented formula (5.31) for the computation of the Gauss-Newton approximation of the Hessian of the asymptotic likelihood function in MATLAB. In plot (b) of figure (5.4), a Gauss-Newton step for the asymptotic likelihood \( L(\alpha, k, c, \sigma^2) \) is shown for various parametrizations in the case when \( n = s = 1 \) and \( m = 0 \); for the formulae for the computation of the derivatives of the state-space matrices with respect to the system parameters see section (5.4) below. We want to make a few remarks again:

(i) The implementation of formulae (5.31) and (5.29) makes it possible to run Gauss-Newton type algorithms (see remark (5.2.4) below) for various parametrizations to find the minimum of the asymptotic likelihood function. Again, this is a nice test setting as one knows the global minimizer in this case, which is clearly not the case if we considered the finite sample likelihood instead.

(ii) In analogy to (ii) in remark (5.2.1) we can now calculate a parameter update obtained by a Gauss-Newton step of the form \( G^{-1}(\nabla L) \) and consider the corresponding update "direction" in the space \( \mathbb{M}_{\text{amp}}(n) \) of transfer functions. In contrast to remark (5.2.1), these update "directions" are now the same (in first order) if we use different parametrizations, provided, of course, that locally the same set of transfer functions is described; see plot (d) of figure (5.4). The reason for this is the fact that the Gauss-Newton matrix \( G \) of the asymptotic likelihood can be interpreted as a Riemannian metric tensor for \( \mathbb{M}_{\text{amp}}(n) \), turning the corresponding Gauss-Newton algorithm into a Riemannian gradient algorithm which is known to be (in first order) parameterization independent. In other words, the problem of choosing a parameterization (that relates well with the chosen Riemannian metric) and the problem of minimizing the criterion function can be decoupled in this case. For a further discussion, we again refer to (Hanzon, 1989) and (Peeters, 1994).

(iii) Note that a "steepest descent" algorithm may be considered to be preferable because each gradient step optimizes the decrease of the criterion function over all steps of fixed (small) length in all possible directions. An appealing reason for the choice of a Riemannian gradient is the fact that the length of such steps should be measured in the transfer function space, and not in (an open subset of) the Euclidean space of one or another parametrization. Riemannian gradient algorithms will thus act on the manifold of transfer functions in a coordinate independent way, but, of course, there is still the question as to what Riemannian metric one should choose. This choice should reflect the user's notion of which transfer functions are considered to be "close" or "far apart". For a discussion of a number of different Riemannian metrics for \( \mathbb{M}_k(n) \) or \( \mathbb{M}_{\text{amp}}(n) \), see (Peeters, 1994), for instance.

Remark 5.2.4. As mentioned in remark (5.2.2) above, we designed a graphical user interface (GUI) in MATLAB for estimation of state-space systems using the parametrizations discussed in this thesis together with some of the algorithms discussed in section (5.5) below. In plot (c) of figure (5.2), the evolution of the

²Considering a Sylvester equation of the form \( P = A \otimes P^T + Q \), it is immediate to rewrite it as \( (I - A \otimes A) \text{vec}(P) = \text{vec}(Q) \) which is an ordinary system of linear equations.
value of the asymptotic likelihood function is shown for McGinnie’s minimum phase balanced canonical form when a Gauss-Newton algorithm as in (5.65) in section (5.5) below (together with a simple line search) is used. Clearly, formulae (5.24), (5.29) and (5.31) are used here.

5.2.4 Evaluations at the true parameter values

For \((\tau, \sigma) = (\tau_0, \sigma_0)\), i.e. for \((A, B, C, D, K, \Sigma) = (A_0, B_0, C_0, D_0, K_0, \Sigma_0)\), the evaluations simplify considerably. Note that in this case, \(x_t^0 = x_t\) and \(e_t = e_t\).

Plugging in \((A, B, C, D, K, \Sigma)\) for the true data generating system, (5.6) becomes

\[
x_{t+1} = A x_t + B u_t + K e_t
\]

\[
e_t = e_t
\]

(5.32)

showing that \(l_{ue}(z) = 0\). The system (5.7) simplifies to

\[
\begin{align*}
\dot{x}_{t+1} &= (\dot{A} - K \dot{C}) x_t + (A - K C) \dot{x}_t + (B - K \dot{D}) u_t + \dot{K} e_t \\
\dot{e}_t &= -\dot{C} x_t - C \dot{x}_t - \dot{D} u_t
\end{align*}
\]

(5.33)

and by combining the systems (5.32) and (5.33) we get the analogue to (5.8):

\[
\begin{pmatrix}
    x_{t+1} \\
    \dot{x}_{t+1}
\end{pmatrix} = \begin{pmatrix}
    A & 0 \\
    \dot{A} - K \dot{C} & A
\end{pmatrix} \begin{pmatrix}
    x_t \\
    \dot{x}_t
\end{pmatrix} + \begin{pmatrix}
    B \\
    \dot{B} - K \dot{D}
\end{pmatrix} u_t + \begin{pmatrix}
    K \\
    \dot{K}
\end{pmatrix} e_t
\]

(5.34)

showing that the transfer function \(l_{ue}(z)\) becomes

\[
l_{ue}(z) = (\dot{C} - C) \left( z^{-1} I - \begin{pmatrix} A & 0 \\ \dot{A} - K \dot{C} & A \end{pmatrix} \right)^{-1} \begin{pmatrix} B \\ \dot{B} - K \dot{D} \end{pmatrix} - \dot{D}
\]

(5.35)

By \(\bar{P}^{(kl)}\) we denote the solution to the Sylvester equation of the form

\[
\bar{P}^{(kl)} = \left( A^{(k)} \ A \right) \begin{pmatrix}
    A & 0 \\
    A^{(l)} - K C^{(k)} & A
\end{pmatrix} \bar{P}^{(kl)} \left( A^{(l)} \ A \right) + \begin{pmatrix} K \\
    K^{(k)} \Sigma ( K' K^{(l)'} )
\end{pmatrix}
\]

(5.36)

By partitioning \(\bar{P}^{(kl)}\) in an obvious way, we can again use (A.47) in the appendix (A.5) to get the following nested Sylvester equations:

\[
\bar{P}^{(kl)}_{11} = A \bar{P}^{(kl)}_{11} A' + K \Sigma K' \implies \bar{P}^{(kl)}_{11} = P
\]

(5.37)

For the (1,2) and the (2,1) block we get

\[
\begin{align*}
\bar{P}^{(kl)}_{12} &= A \bar{P}^{(kl)}_{12} A + A P(A^{(l)} - K C^{(l)})' + K \Sigma K^{(l)'} P A' + K^{(k)} \Sigma K'
\end{align*}
\]

(5.38)
Finally, for the \((2,2)\) block the following equation is obtained:

\[
\tilde{P}_{22}^{(k)} = \tilde{A}P_{22}^{(k)} \tilde{A} + (A^{(k)} - KC^{(k)})P(A^{(l)} - KC^{(l)})' + (A^{(k)} - KC^{(k)})P_{12}^{(k)} \tilde{A} + \tilde{A}P_{21}^{(k)} (A^{(l)} - KC^{(l)})' + K^{(k)} \Sigma K^{(l)}
\]  

(5.39)

**The likelihood at the true parameter values**

Using (5.34) and the fact that \(l_{ue}(z) = 0\), the likelihood function in (5.24) simplifies to

\[
L(\tau, \sigma) = \log \det \Sigma + tr \Sigma^{-1} \{0 + \Sigma\} = \log \det \Sigma + s
\]  

(5.40)

which is in accordance with what has been claimed in (2.88) in section (2.3.4).

**The gradient at the true parameter values**

The gradient of the asymptotic likelihood function in (5.29) simplifies to

\[
\dot{L}(\tau, \sigma) = 2 \cdot tr \Sigma^{-1} \left\{ -\dot{C} - C \right\} \tilde{P} \left( \begin{array}{c} 0 \\ 0 \end{array} \right) + 0 = 0 + 0 = 0
\]  

(5.41)

which can again be seen from (5.34) and the fact that \(l_{ue}(z) = 0\). This is in accordance with the fact that the likelihood is minimized at the true parameter values, implying that the gradient becomes zero there.

**The Hessian at the true parameter values**

The approximation to the Hessian in (5.31) now becomes the true Hessian because the neglected term in (5.30) vanishes at the true parameter values. Instead of (5.31), we therefore get the well known formula

\[
L^{(k)}(\tau, \sigma) = 2 \cdot tr \Sigma^{-1} \left\{ C^{(k)} C \right\} \left( \begin{array}{c} P \\ P^{(k)}_{21} \end{array} \right) \left( \begin{array}{c} C^{(l)}' \\ C' \end{array} \right) \right) + \int_{-\infty}^{\infty} l_{ue}^{(k)}(\omega) dF_\omega(\omega) l_{ue}^{(l)}(\omega)'
\]  

(5.42)

where \(P \in \mathbb{R}^{nxn}\) is given in (5.37), \(P_{12}^{(k)} \in \mathbb{R}^{nxn}\) and \(P_{21}^{(k)} \in \mathbb{R}^{nxn}\) are given in (5.38), \(P_{22}^{(k)} \in \mathbb{R}^{nxn}\) is given in (5.39) and the transfer functions \(l_{ue}^{(k)}(z)\) and \(l_{ue}^{(l)}(z)\) are given in (5.35), where all "dots" have to be replaced by differentiation with respect to \(\tau_k\) and \(\tau_l\), respectively.

**Remark 5.2.5.** Clearly, formula (5.42) yields the \((k,l)\) element of the Fisher information matrix. In table (5.6) in section (5.6), we present this matrix for various parametrizations for the case \(n = s = 1\) and \(m = 0\). Note that the computation is carried out symbolically in Mathematica with the help of so called Fadee sequences (and using the decomposition of the "larger" Sylvester equation into four "smaller" ones as outlined above); see (Peeters and Hanzon, 1999) for a detailed description of this approach. The Fisher information matrix is also used in figures (5.3) and (5.1) in section (5.8).

### 5.3 The finite sample likelihood \(L_T(\tau, \sigma)\)

In section (2.3.2) we discussed how the exact finite sample likelihood function (2.67) can be evaluated using its prediction error form and Kalman filtering techniques. Assuming that the inputs are perfectly predictable, it was illustrated in theorem (2.3.1) how to initialize the Kalman filter to evaluate this function. However, a disadvantage of this approach is that both quantities \(\Sigma_{\eta^{-1}}(\theta)\) and \(\epsilon_t(\theta)\) in the prediction error decomposition form of the likelihood (2.70) will be quite complicated functions of system parameters \(\tau\) and covariance parameters \(\sigma\) due to the nature of the recursive Kalman filtering equations. This makes the computation of the gradient \(\nabla L_T\) very cumbersome.

In the sequel we will therefore restrict ourselves to an approximation of the finite sample likelihood and its derivatives. For ease of presentation, we will repeat the simplifying assumptions:
• The input is assumed to be perfectly predictable, i.e., \( \hat{u}_{ts} = u_t \) for \( t > s \geq 0, u_1 \neq 0 \); see remark (2.1.10) in section (2.1).

• The Kalman filter is initialized with \( \hat{x}_{1|0} = 0 \) and \( P_1 = P_{1|0} = 0 \), implying (2.71) and (2.75), i.e., the quantities \( e_i(\theta) \) are given by filtering the observed inputs and outputs by the formal inverse of the system \((A, B, C, D, K)\). Note that these quantities now only depend on system parameters \( \tau \) for \((l, k)\). We obtain the likelihood function \( L_T(Y_l^T; U_l^T, (\tau, \sigma)) \) in (2.76).

• There are no cross restrictions between the unknown and unrestricted covariance parameters \( \sigma \) and the system parameters \( \tau \), hence we could also "concentrate out" the parameters in \( \Sigma \) to obtain the concentrated likelihood function \( L_T^c(Y_l^T; U_l^T, \tau) \) in (2.77). For any given \( \tau \), the corresponding \( \Sigma \) minimizing (2.76) is given by \( \Sigma = \frac{1}{T} \sum_{t=1}^{T} e_t(\tau)e_t(\tau)' \).

• Assume that the vector of free system parameters is of the form \( \tau = (\tau^o, \tau^u) \) where \( \tau^u \) consists of all entries in \( \bar{D}, \bar{B} \) and \( \bar{K} \). Assume, furthermore, that we are given an estimate \( \hat{\Sigma} \) of the innovations covariance matrix. If there are no cross restrictions between the parameters \( \tau^o \) (which only appear in \( \bar{A} \) and \( \bar{C} \) corresponding to the inverse system) and \( \tau^u \), then we could also "concentrate out" the parameters in \( \tau^u \) to obtain the double concentrated likelihood function \( L_T^{cc}(Y_l^T; U_l^T, \tau^o) \) in (2.85).

### 5.3.1 Evaluating \( L_T(\tau, \sigma), L_T^{cc}(\tau) \) and \( L_T^{cc}(\tau^o) \) for a particular parametrization

Under the assumptions discussed above, the quantities \( e_i(\tau) \) only depend on system parameters for \((l, k)\) and are thus given by (5.5), i.e., by

\[
\begin{align*}
x_{t+1} &= \bar{A}x_t + \bar{B}u_t + \bar{K}y_t \\
e_t &= \bar{C}x_t + \bar{D}u_t + y_t
\end{align*}
\]

(5.43)

where the dependence of \( e_t \) on \( \tau \) is not made explicit and \( (\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{K}) = (A-KC, B-KD, -C, -D, K) \);

see (2.74). This allows us to evaluate \( L_T(Y_l^T; U_l^T, (\tau, \sigma)) \) in (2.76) as well as \( L_T^{cc}(Y_l^T; U_l^T, \tau) \) in (2.77) and \( L_T^{cc}(Y_l^T; U_l^T, \tau^o) \) in (2.85).

### 5.3.2 Evaluating \( \nabla L_T(\tau, \sigma), \nabla L_T^{cc}(\tau) \) and \( \nabla L_T^{cc}(\tau^o) \) for a particular parametrization

The partial derivatives of \( L_T(Y_l^T; U_l^T, (\tau, \sigma)) \) in (2.76), \( L_T^{cc}(Y_l^T; U_l^T, \tau) \) in (2.77) and \( L_T^{cc}(Y_l^T; U_l^T, \tau^o) \) in (2.85) are obtained as follows:

\[
\begin{align*}
\dot{L}_T(Y_l^T; U_l^T, (\tau, \sigma)) &= 2tr \left\{ \frac{1}{T} \sum_{t=1}^{T} \dot{e}_t e_t' \Sigma(\sigma)^{-1} \right\} \\
\dot{L}_T^{cc}(Y_l^T; U_l^T, \tau) &= 2tr \left\{ \left( \sum_{t=1}^{T} e_t e_t' \right)^{-1} \sum_{t=1}^{T} \dot{e}_t e_t' \right\} \\
\dot{L}_T^{cc}(Y_l^T; U_l^T, \tau^o) &= 2tr \left\{ \sum_{t=1}^{T} \dot{e}_t e_t' \right\}
\end{align*}
\]

(5.44) \hspace{2cm} (5.45) \hspace{2cm} (5.46)

where \( \dot{e}_t = \Sigma^{-1/2} e_t \) in (5.46) and the dependence of \( e_t \) on \( \tau \) in (5.44) and (5.45) and \( \dot{e}_t \) on \( \tau^o \) in (5.46)

is not made explicit. Note that (5.44) follows from the fact that there are no cross restrictions between the unknown covariance parameters \( \sigma \) and the system parameters \( \tau \). Furthermore, note that the partial derivatives of \( L_T(Y_l^T; U_l^T,(\tau, \sigma)) \) with respect to the entries in \( \sigma \) are directly obtained by considering the on- and above diagonal elements of (2.78) in section (2.3.2).
For (5.44) and (5.45), the quantities $e_t$ and $\hat{e}_t$ are given by (5.5), i.e. by

$$
\begin{pmatrix}
  x_{t+1} \\
  \hat{x}_{t+1} \\
  e_t \\
  \hat{e}_t
\end{pmatrix}
= 
\begin{pmatrix}
  \hat{A} & 0 & 0 & 0 \\
  \hat{A} & \hat{A} & 0 & 0 \\
  C & 0 & \hat{C} & 0 \\
  \hat{C} & \hat{C} & \hat{C} & \hat{C}
\end{pmatrix}
\begin{pmatrix}
  x_t \\
  \hat{x}_t \\
  x_t \\
  \hat{x}_t
\end{pmatrix}
+ 
\begin{pmatrix}
  \hat{B} \\
  \hat{B} \\
  \hat{D} \\
  \hat{D}
\end{pmatrix}
\begin{pmatrix}
  u_t \\
  \hat{u}_t \\
  u_t \\
  \hat{u}_t
\end{pmatrix}
+ 
\begin{pmatrix}
  \hat{K} \\
  \hat{K} \\
  \hat{I_s} \\
  0
\end{pmatrix}
\begin{pmatrix}
  y_t \\
  \hat{y}_t \\
  y_t \\
  \hat{y}_t
\end{pmatrix}
$$

(5.47)

where $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{K})$ is again given by (2.74). Note that $\hat{A}$, $\hat{B}$, $\hat{C}$, $\hat{D}$ and $\hat{K}$ are given as in (5.5) because $\tau$ is assumed to be a vector of free parameters for the state-space matrices $(A, B, C, D, K)$ and not for its inverse system $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{K})$ in this case.

For (5.46), we first remark that the weighted innovations $(\hat{e}_1^T, \ldots, \hat{e}_T^T)' = \hat{E}_1^T$ are given in (2.81), where vec$(\hat{D})$, vec$(\hat{B})$ and vec$(\hat{K})$ have to be replaced by (2.83):

$$
\hat{E}_1^T = y_T^T - \hat{X} (\hat{X}' \hat{X})^{-1} \hat{X}' \hat{y}_T^T = \left( I_{sT} - \hat{X} (\hat{X}' \hat{X})^{-1} \hat{X}' \right) \hat{y}_T^T
$$

(5.48)

The partial derivatives $\hat{e}_t$ with respect to the entries in $\tau^0$ are then obtained by differentiating (5.48):

$$
\hat{E}_1^T = - (\hat{X} (\hat{X}' \hat{X})^{-1} \hat{X}' + \hat{X} (-\hat{X}' \hat{X})^{-1} (\hat{X}' \hat{X} + \hat{X}' \hat{X}) (\hat{X}' \hat{X})^{-1} \hat{X}' + (\hat{X}' \hat{X})^{-1} \hat{X}'') \hat{y}_T^T
$$

(5.49)

where $\hat{X}' = (\hat{X}' \hat{X})^{-1} \hat{X}'$. Finally, from (2.80) and (2.81), $\hat{X}$ is easily seen to be

$$
\hat{X} = (I_T \otimes \Sigma^{-\frac{1}{T}}) \cdot 
\begin{pmatrix}
  0_{s \times ms} & 0_{s \times (m+s)n} \\
  0_{s \times ms} & - (u_1, y_1)' \otimes \hat{C} \\
  0_{s \times ms} & - (u_1, y_1)' \otimes (\hat{C} \hat{A} + \hat{C} \hat{A}) - (u_2, y_2)' \otimes \hat{C} \\
  \vdots & \vdots \\
  0_{s \times ms} & - (u_1, y_1)' \otimes (\hat{C} \hat{A}^T - 2 + \hat{C} \sum_{j=0}^{T-3} \hat{A}^j \hat{A}^T - 3 - j) - (u_{T-1}, y_{T-1})' \otimes \hat{C}
\end{pmatrix}
$$

$$
= (I_T \otimes \Sigma^{-\frac{1}{T}}) \cdot 
\begin{pmatrix}
  0_{sT \times ms} & 0_{sT \times (m+s)n} \\
  0_{sT \times ms} & - (u_1, y_1)' \otimes \hat{K}_1 + (u_1, y_1)' \otimes \hat{K}_2 + \cdots + (u_1, y_1)' \otimes \hat{K}_{T-1}
\end{pmatrix}
$$

where $(\hat{K}_i)$ denotes the impulse response of the derivative system of $(\hat{A}, I_n, \hat{C}, 0_{s \times n})$ which is obtained completely analogous to (5.4) and hence given by

$$
\left( 
\begin{pmatrix}
  \hat{A} \\
  \hat{A}
\end{pmatrix}
, 
\begin{pmatrix}
  I_n \\
  0_{n \times n}
\end{pmatrix}
, 
\begin{pmatrix}
  \hat{C} \\
  \hat{C}
\end{pmatrix}
, 
0_{s \times n}
\right)
$$

(5.50)

The impulse response of (5.50) can be calculated efficiently by filtering the corresponding impulse sequences. Note that in this case $\tau^0$ is assumed to be a vector of free parameters for $\hat{A}$ and $\hat{C}$ and not for the $\hat{A}$ and the $\hat{C}$ matrix of the original system. In case of the SLSDDLC parametrization in (4.83), for instance, vec$(\hat{A})$ and vec$(\hat{C})$ are given by the corresponding rows of $Q_{cc}^+$ in (4.83).
5.3.3 Evaluating an approximate Hessian of $L_T(\tau, \sigma)$, $L_T^\sigma(\tau)$ and $L_T^{\sigma\sigma}(\tau^o)$ for a particular parametrization

The Hessians of $L_T(Y_1^T; U_1^T, (\tau, \sigma))$ in (2.76), $L_T^\sigma(Y_1^T; U_1^T, \tau)$ in (2.77) and $L_T^{\sigma\sigma}(Y_1^T; U_1^T, \tau^o)$ in (2.85) are obtained from (5.44), (5.45) and (5.46) as follows:

$$
L_{\tau}^{(k)}(Y_1^T; U_1^T, (\tau, \sigma)) = 2tr\left\{ \frac{1}{T} \sum_{t=1}^{T} e_t(\tau) e_t^T(\tau) \Sigma(\sigma)^{-1} + e_t^{(k)} e_t^{(k)\top} \Sigma(\sigma)^{-1} \right\}
$$

$$
L_{\tau}^{\sigma(\tau)}(Y_1^T; U_1^T, \tau) = 2tr\left\{ \left( \sum_{t=1}^{T} e_t^{(k)} e_t^T(\tau) \right) \left( \sum_{t=1}^{T} e_t e_t^T(\tau) \right)^{-1} - 2 \left( \sum_{t=1}^{T} e_t e_t^T(\tau) \right)^{-1} \left( \sum_{t=1}^{T} e_t e_t^T(\tau) \right)^{-1} \left( \sum_{t=1}^{T} e_t^{(k)} e_t^T(\tau) \right)^{-1} \right\}
$$

$$
L_{\tau}^{\sigma(\sigma)}(Y_1^T; U_1^T, \tau) = 2tr\left\{ \left( \sum_{t=1}^{T} e_t^{(k)} e_t^T(\tau) \right) \left( \sum_{t=1}^{T} e_t^{(k)} e_t^T(\tau) \right)^{-1} - 2 \left( \sum_{t=1}^{T} e_t^{(k)} e_t^T(\tau) \right)^{-1} \left( \sum_{t=1}^{T} e_t^{(k)} e_t^T(\tau) \right)^{-1} \left( \sum_{t=1}^{T} e_t e_t^T(\tau) \right)^{-1} \right\}
$$

Again, we have only considered first order derivatives of the estimated innovations for the approximation of the Hessians in (5.51), (5.52) and (5.53), yielding the Gauss-Newton approximation of the Hessians. For (5.51) and (5.52), the estimated innovations and their derivatives are given in (5.47), whereas for (5.53), the corresponding quantities are obtained by (5.48) and (5.49).

5.4 Derivatives of state-space matrices

In sections (5.2) and (5.3) we have frequently used the symbols $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{K})$ or $\hat{A}$ and $\hat{C}$ to denote derivatives of state-space matrices with respect to system parameters. In this section we will explicitly calculate these quantities for the state-space parameterizations considered in chapter 4.

Note that in chapter 4 the mappings attaching the state-space matrices $(A, B, C, D, K)$ to the parameter vectors $\tau \in T$ have been denoted by $\phi$. Different subscripts (and superscripts) have been used for the parameter vectors $\tau$, the parameter spaces $T$ and the corresponding mappings $\phi$ in order to distinguish different parametrizations. We will subsequently make use of this unified notation, always considering the vectorized version $\phi(\tau)$, i.e. $\phi(\tau) = (vec(A(\tau)), vec(B(\tau)), vec(C(\tau)), vec(D(\tau)))$. All formulae presented below have been implemented in MATLAB. This makes it possible to compute gradients and approximate Hessians of the likelihood type criterion functions considered above for the state-space parametrizations considered in this thesis.

Derivatives for the full state-space parametrization

Note that the parameter vector is given by $\tau = (vec(A), vec(B), vec(C), vec(D))$ in case of the full state-space parametrization. The mapping $\phi$ attaching the state-space matrices to the parameter vector is the identity mapping, and $\frac{d\phi}{d\tau} = I_{n^2+2ms+m(n+s)}$.

Derivatives for echelon state-space forms

Here, the mapping $\phi_\alpha$ from the parameter vectors $\tau_\alpha$ to the state-space matrices simply inserts the components of $\tau_\alpha$ in the correct positions of $(A, B, C, D, K)$. Hence, $\frac{d\phi_\alpha}{d\tau_\alpha}(\tau_\alpha) = (0, \ldots, 0, 1, 0, \ldots, 0)'$ where $\tau_\alpha^i$ denotes the i. component of $\tau_\alpha$ and the vector $(0, \ldots, 0, 1, 0, \ldots, 0)'$ contains only zeros except for a one at the position where $\tau_\alpha^i$ appears in $(vec(A), vec(B), vec(C), vec(D))$. 
Derivatives for the overlapping state-space parametrization

The mapping $\varphi^{(1)}_\alpha$ from the parameter vectors $\tau^{(1)}_\alpha$ to the state-space matrices again simply inserts the components of $\tau^{(1)}_\alpha$ in the correct positions of $(A, B, C, D, K)$. Hence, $\frac{\partial \varphi^{(1)}_\alpha}{\partial \tau^{(1)}_\alpha} = (0, \ldots, 0, 1, 0, \ldots, 0)'$, where $\tau^{(1)}_\alpha(i)$ denotes the $i$. component of $\tau^{(1)}_\alpha$ and the vector $(0, \ldots, 0, 1, 0, \ldots, 0)'$ contains only zeros except for a one at the position where $\tau^{(1)}_\alpha(i)$ appears in $(vec(A)', vec(B)', vec(C)', vec(D)')'$. 

Derivatives for Lyapunov balanced canonical forms (Ober)

It is seen from theorem (4.6.1) that the mapping $\varphi_\delta$ from the parameter vectors $\tau_\delta \in T_\delta$ to the state-space matrices is composed of two mappings: First, the parameters are mapped to the continuous time matrices $(A_c, B_c, C_c, D_c)$, and then the homeomorphism $\rho$ given in (A.44) is applied. We will first consider the derivatives of $\rho$:

$$
\frac{\partial A}{\partial \tau^i_\delta} = \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1} + (I + A_c)(I - A_c)^{-1} \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1} = (I + A) \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1}
$$

$$
\frac{\partial B}{\partial \tau^i_\delta} = \sqrt{2}(I - A_c)^{-1} \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1} B_c + \sqrt{2}(I - A_c)^{-1} \frac{\partial B_c}{\partial \tau^i_\delta} = (I - A_c)^{-1} \frac{\partial A_c}{\partial \tau^i_\delta} B + \sqrt{2}(I - A_c)^{-1} \frac{\partial B_c}{\partial \tau^i_\delta}
$$

$$
\frac{\partial C}{\partial \tau^i_\delta} = \sqrt{2} \frac{\partial C_c}{\partial \tau^i_\delta} (I - A_c)^{-1} + \sqrt{2} C_c (I - A_c)^{-1} \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1} = \sqrt{2} \frac{\partial C_c}{\partial \tau^i_\delta} (I - A_c)^{-1} + C \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1}
$$

$$
\frac{\partial D}{\partial \tau^i_\delta} = \frac{\partial D_c}{\partial \tau^i_\delta} + \frac{\partial C_c}{\partial \tau^i_\delta} (I - A_c)^{-1} B_c + C_c (I - A_c)^{-1} \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1} B_c = \frac{\partial D_c}{\partial \tau^i_\delta} + \frac{\partial C_c}{\partial \tau^i_\delta} \frac{1}{\sqrt{2}} B + \frac{1}{\sqrt{2}} C \frac{\partial A_c}{\partial \tau^i_\delta} B + \frac{1}{\sqrt{2}} C \frac{\partial B_c}{\partial \tau^i_\delta}
$$

Vectorization yields

$$
\begin{pmatrix}
vec_\frac{\partial A}{\partial \tau^i_\delta} \\
vec_\frac{\partial B}{\partial \tau^i_\delta} \\
vec_\frac{\partial C}{\partial \tau^i_\delta} \\
vec_\frac{\partial D}{\partial \tau^i_\delta}
\end{pmatrix} =
\begin{pmatrix}
vec(I + A) \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1} \\
vec(I - A_c)^{-1} \frac{\partial A_c}{\partial \tau^i_\delta} B + \sqrt{2} \frac{\partial C_c}{\partial \tau^i_\delta} (I - A_c)^{-1} + \sqrt{2} C_c (I - A_c)^{-1} \frac{\partial A_c}{\partial \tau^i_\delta} (I - A_c)^{-1} B_c + C_c (I - A_c)^{-1} \frac{\partial B_c}{\partial \tau^i_\delta}
\end{pmatrix}
$$

which can be written as

$$
\begin{pmatrix}
(I - A_c)^{-1} \otimes (I + A) \\
B^T \otimes (I - A_c)^{-1} \\
(I - A_c)^{-1} \otimes C \\
B^T \otimes \frac{1}{\sqrt{2}} C
\end{pmatrix}
\begin{pmatrix}
0_{n^2 \times n(m+s)} \\
\sqrt{2} I_{(m+s)} \otimes (I - A_c)^{-1} \\
0_{n(m+s) \times ns} \\
\frac{1}{\sqrt{2}} I_{(m+s)} \otimes C
\end{pmatrix}
\begin{pmatrix} \frac{\partial A}{\partial \tau^i_\delta} \\
\frac{\partial B}{\partial \tau^i_\delta} \\
\frac{\partial C}{\partial \tau^i_\delta} \\
\frac{\partial D}{\partial \tau^i_\delta}
\end{pmatrix}
$$

What remains is the problem of calculating the derivatives of $(A_c, B_c, C_c, D_c)$ with respect to the parameters in $\tau_\delta$. This will only be done for the generic piece, i.e. for the matrices given in (4.54). For
the case of the Lyapunov balanced forms considered here, this was done in (Chou, 1994) and (Chou and Maciejowski, 1997). However, we will present the formulae once again, because the cited formulae contain some minor (and almost unavoidable) errors which are hopefully not present in this exposition (although we do not dare to make guarantees at this point). We want to remind the reader of the form of the parameter vector given in (4.53): below we use the notation \( \tau_c = (s_1, \ldots, s_n) = (\sigma_1 - \sigma_2, \ldots, \sigma_{n-1} - \sigma_n, \sigma_n) \) for the first part of the parameter vector. Hence,

\[
\frac{\partial A_c}{\partial s_i} = \frac{\partial A_c}{\partial \sigma_i} - \frac{\partial A_c}{\partial \sigma_{i+1}} \quad \text{for } i = 1, \ldots, (n-1)
\]  \( (5.55) \)

Dotted rows and columns are filled according to the pattern recognizable from the other entries, and empty spaces contain only zeros:

\[
\frac{\partial A_c}{\partial \sigma_i} = \begin{pmatrix}
\frac{\sigma_i^2 + \sigma_j^2}{(\sigma_i^2 - \sigma_j^2)} \hat{B}_1 \hat{B}_1' - \frac{2\sigma_i \sigma_j}{(\sigma_i^2 - \sigma_j^2)} C_i' C_j' \\
\cdots \\
\frac{\sigma_i^2 + \sigma_j^2}{(\sigma_i^2 - \sigma_j^2)} \hat{B}_i \hat{B}_i' - \frac{2\sigma_i \sigma_j}{(\sigma_i^2 - \sigma_j^2)} C_i' C_j' \\
\cdots \\
\frac{\sigma_i^2 + \sigma_j^2}{(\sigma_i^2 - \sigma_j^2)} \hat{B}_n \hat{B}_n' - \frac{2\sigma_i \sigma_j}{(\sigma_i^2 - \sigma_j^2)} C_i' C_j' \\
\end{pmatrix}
\]

\[
\frac{\partial A_c}{\partial B_{ij}} = \begin{pmatrix}
\frac{1}{\sigma_i^2 - \sigma_j^2} (\sigma_1 \hat{B}_{ij} - \frac{1}{B_1} \sigma_i \hat{B}_{ij} C_i' C_j') \\
\cdots \\
\frac{1}{\sigma_i^2 - \sigma_j^2} (\sigma_n \hat{B}_{nj} - \frac{1}{B_1} \sigma_n \hat{B}_{nj} C_i' C_j') \\
\end{pmatrix}
\]

\[
\frac{\partial A_c}{\partial \phi_{i,j}} = \begin{pmatrix}
- \frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} C_i' \left( \frac{\partial C_i}{\partial \phi_{i,j}} \right) \\
\cdots \\
- \frac{\sigma_n}{\sigma_n^2 - \sigma_j^2} C_n' \left( \frac{\partial C_n}{\partial \phi_{i,j}} \right) \\
\end{pmatrix}
\]

\[
\frac{\partial A_c}{\partial d_{i,j}} = 0_{n \times n}
\]

where in \( \frac{\partial A}{\partial B_{ij}} \) all \( \hat{B}_{c,ij} \) have been replaced by the shorter \( \hat{B}_{ij} \) because of the space limitations. Note that all derivatives of \( A_c \) contain nonzero elements only in the \( i \)th row and the \( i \)th column. Additionally,

\[
\frac{\partial \hat{B}_c}{\partial \hat{B}_{c,ij}} = \begin{pmatrix} 1 \end{pmatrix} \quad \text{at position } (i,j)
\]

\[
\frac{\partial \hat{B}_c}{\partial \tau_\delta} = 0_{n \times (m+s)} \quad \text{otherwise}
\]

\[
\frac{\partial C_c}{\partial \hat{B}_{c,ij}} = \begin{bmatrix} 0_{\times (i-1)} & C_i \cdot \hat{B}_{c,ij} \cdot \hat{B}_1' \hat{B}_1' \end{bmatrix} 0_{\times (n-i)}
\]

\[
\frac{\partial C_c}{\partial \phi_{i,j}} = \begin{bmatrix} 0_{\times (i-1)} & \sqrt{\hat{B}_1' \hat{B}_1' \left( \frac{\partial \hat{B}_1'}{\partial \phi_{i,j}} \right)} \end{bmatrix} 0_{\times (n-i)}
\]

\[
\frac{\partial C_c}{\partial \tau_\delta} = 0_{n \times n} \quad \text{otherwise}
\]
\[
\frac{\partial \tilde{D}_c}{\partial \tau^i_\delta} = \begin{pmatrix}
1
\end{pmatrix} \text{ at position } (i, j)
\]
\[
\frac{\partial \tilde{D}_c}{\partial \phi^i_\delta} = \begin{pmatrix}
- \left( \frac{\partial \phi^i_\delta}{\partial \tau^i_\delta} + C_c(I - A_c)^{-1} \frac{\partial A_c}{\partial \phi^i_\delta} \right) \left( I - A_c \right)^{-1} K_c - C_c(I - A_c)^{-1} \frac{\partial K_c}{\partial \phi^i_\delta} \end{pmatrix} \text{ for } \tau^i_\delta \neq d_{c,ij} (5.56)
\]

Note that the derivatives in (5.56) will only be needed once. To see this, first note that on the one hand the second part \(E_c \text{ of } \tilde{D}_c\) contains all parameters except for the \(d_{c,ij}\), but on the other hand it is constructed in such a way that the direct feedthrough \(k(0)\) in discrete time is the identity matrix \(I_c\) for all parameter values. Hence, the corresponding discrete time derivative will be zero. Second, the derivative of \(\tilde{D}_c\) does not enter the calculation of the derivatives of the discrete time matrices \(A, B\) and \(C\) as can be seen from (5.54). However, when considering the derivative of the discrete time \(D\) matrix corresponding to \(l(0)\), the derivatives in (5.56) have to be calculated.

The derivative \(\frac{\partial U_i}{\partial \phi^i_\delta}\) is given by
\[
\frac{\partial U_i}{\partial \phi^i_\delta} = \begin{pmatrix}
K_j \cdot U(\gamma_j)
\end{pmatrix}_{0(s-j-1) \times 1} (5.57)
\]

Here, \(\gamma_j = [\phi_{i,1}, \ldots, \phi_{i,j-1}, \phi_{i,j} + \frac{\tau^i_\delta}{2}]\), \(K_j = 1\) for \(j = (s - 1)\) and \(K_j = \prod_{k=j+1}^{s-1} \cos(\phi_{i,k})\) otherwise. Moreover, \(U(\gamma_j)\) denotes the \((j + 1)\) dimensional vector corresponding to the polar coordinates in \(\gamma_j\) via (4.52).

**Derivatives for stochastically balanced canonical forms (Ober)**

It is seen from theorem (4.7.1) that the mapping \(\phi^{(1)}_\delta\) from the parameter vectors \(\tau^{(1)}_\delta \in T^{(1)}_\delta\) to the state-space matrices is again composed of two mappings. We can use the derivatives in (5.54) and then calculate the derivatives of \((A_c, \tilde{B}_c, C_c, \tilde{D}_c)\) with respect to the parameters in \(\tau^{(1)}_\delta\). This is done in the same way as discussed for Ober's Lyapunov balanced canonical form above and for McGinnie's minimum phase balanced canonical form below. We only implemented the latter two balanced forms in MATLAB and will therefore not present the formulae for Ober's stochastically balanced canonical form here. However, their derivation is just a matter of "careful" differentiation.

**Derivatives for minimum phase balanced canonical forms (McGinnie)**

Again, it is clear from theorem (4.8.1) that the mapping \(\phi^{(2)}_\delta\) from the parameter vectors \(\tau^{(2)}_\delta \in T^{(2)}_\delta\) to the state-space matrices is composed of two mappings. Using the derivatives in (5.54), we need to calculate the derivatives of \((A_c, \tilde{B}_c, C_c, \tilde{D}_c)\) with respect to the parameters in \(\tau^{(2)}_\delta\). Again, this will only be done for the generic piece which is not explicitly stated in section (4.8), but should be clear from (4.72) in theorem (4.8.1).

As for Ober's Lyapunov balanced canonical form, the following derivatives of \(A_c\) contain non-zero elements only in the \(i\) row and the \(i\) column:

\[
\frac{\partial A_c}{\partial \phi^i_\delta} = [\tilde{A}_{ki}]
\]
\[
\tilde{A}_{ki} = \begin{cases}
\frac{\sigma_0}{\sigma_i^2 - \sigma_i^2} A_c, i + \frac{\sigma_0}{\sigma_i^2 - \sigma_i^2} K_c \frac{\sigma_0^2}{\sqrt{1 + \sigma_0^2}} \frac{1}{\sqrt{1 + \sigma_0^2}} & \text{ for } k = i, i \neq i \\
\frac{\sigma_0}{\sigma_k^2 - \sigma_i^2} A_c, k + \frac{\sigma_0}{\sigma_k^2 - \sigma_i^2} K_c \frac{\sigma_0^2}{\sqrt{1 + \sigma_0^2}} \frac{1}{\sqrt{1 + \sigma_0^2}} & \text{ for } k \neq i, i = i
\end{cases}
\]

\[
\tilde{A}_{ki} = \frac{\sigma_k}{\sigma_i^2 - \sigma_i^2} A_c, k + \frac{\sigma_k}{\sigma_i^2 - \sigma_i^2} K_c \frac{\sigma_k^2}{\sqrt{1 + \sigma_k^2}} \frac{1}{\sqrt{1 + \sigma_k^2}} & \text{ for } k \neq i, i = i
\]

\[
\tilde{A}_{ki} = \frac{A_c, i, i}{\sigma_i} & \text{ for } k = l = i
\]
\[
\frac{\partial A_c}{\partial K_{ij}} = \begin{bmatrix} \dot{A}_{kl} \end{bmatrix}
\]
\[
\dot{A}_{kl} = \frac{\sigma_l (1 + \sigma_l^2) K_{ij}}{\sigma_l^2 - \sigma_i^2} - \frac{\sigma_l \sqrt{1 + \sigma_l^2} \sqrt{1 + \sigma_i^2}}{(\sigma_l^2 - \sigma_i^2) K_{ij} K_i K_j} K_i K_j U_i U_i \quad \text{for } k = i, l \neq i
\]
\[
\dot{A}_{kl} = \frac{\sigma_k (1 + \sigma_k^2) K_{ij}}{\sigma_k^2 - \sigma_i^2} K_{ij} K_k K_j K_k U_i U_i \quad \text{for } k \neq i, l = i
\]
\[
\dot{A}_{kl} = -\frac{K_{ij}}{\sigma_i} \quad \text{for } k = l = i
\]
\[
\frac{\partial A_c}{\partial \phi_{i,j}} = \begin{bmatrix} \dot{A}_{kl} \end{bmatrix}
\]
\[
\dot{A}_{kl} = -\frac{1}{\sigma_i^2 - \sigma_l^2} \sigma_i \sqrt{1 + \sigma_i^2} \sqrt{1 + \sigma_l^2} \sqrt{K_i K_j K_i K_l} \left( \frac{\partial U_i}{\partial \phi_{i,j}} \right) \quad \text{for } k = i, l \neq i
\]
\[
\dot{A}_{kl} = \frac{1}{\sigma_i^2 - \sigma_l^2} \sigma_i \sqrt{1 + \sigma_i^2} \sqrt{1 + \sigma_l^2} \sqrt{K_i K_j K_k K_l} K_i K_j U_k \left( \frac{\partial U_i}{\partial \phi_{i,j}} \right) \quad \text{for } k \neq i, l = i
\]
\[
\dot{A}_{kl} = 0 \quad \text{for } k = l = i
\]
\[
\frac{\partial A_c}{\partial \theta_{i,j}} = 0_{n \times n}
\]
\[
\frac{\partial A_c}{\partial \epsilon_{i,j}} = 0_{n \times n}
\]

The derivatives of \( \dot{B}_c \) are calculated as follows:
\[
\frac{\partial \dot{B}_c}{\partial K_{i,j}} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{at position } (i, j)
\]
\[
\frac{\partial \dot{B}_c}{\partial U_i} = \begin{pmatrix} \sigma_i \\ \vdots \\ \sigma_i \end{pmatrix} \quad \text{at position } (i, j + m)
\]
\[
\frac{\partial \dot{B}_c}{\partial \phi_{i,j}} = 0_{n \times (m + \sigma)} \quad \text{for other components of } \dot{A}_c^{(2)}
\]

Moreover, we have for \( \dot{C}_c \):
\[
\frac{\partial \dot{C}_c}{\partial \theta_{i}} = \begin{bmatrix} 0 \times (n + i) & K_i' + \frac{\sigma_i}{\sqrt{1 + \sigma_i^2}} U_i \cdot \sqrt{K_i K_i'} & 0 \times (n - i) \end{bmatrix}
\]
\[
\frac{\partial \dot{C}_c}{\partial K_{i,j}} = \begin{bmatrix} 0 \times (n + i) & 0 \\ \sigma_i \\ \vdots \\ 0 \\ \vdots \\ \sigma_i \\ \vdots \\ 0 \end{bmatrix} + \frac{\sigma_i}{\sqrt{1 + \sigma_i^2}} U_i \cdot \frac{K_i K_j K_i'}{\sqrt{K_j K_i'}} 0 \times (n - i) \quad \text{row } j
\]
\[
\frac{\partial \dot{C}_c}{\partial \phi_{i,j}} = \begin{bmatrix} 0 \times (n + i) & \sqrt{1 + \sigma_i^2} (\frac{\partial U_i}{\partial \phi_{i,j}}) K_i K_i' \end{bmatrix} 0 \times (n - i)
\]
\[
\frac{\partial \dot{C}_c}{\partial \theta_{i,j}} = 0_{n \times n}
\]
\[
\frac{\partial \dot{C}_c}{\partial \epsilon_{i,j}} = 0_{n \times n}
\]

Note that \( \dot{D}_c = [D_c, E_c] \):

\[
\frac{\partial \dot{D}_c}{\partial \theta_{i,j}} = \begin{bmatrix} D_i \times (n + i) & E_i \times (n - i) \end{bmatrix}
\]
\[
\frac{\partial \dot{D}_c}{\partial \phi_{i,j}} = \begin{bmatrix} D_i \times (n + i) & E_i \times (n - i) \end{bmatrix}
\]
\[
\frac{\partial \dot{D}_c}{\partial \epsilon_{i,j}} = \begin{bmatrix} D_i \times (n + i) & E_i \times (n - i) \end{bmatrix}
\]

5.4. **Derivatives of State-Space Matrices**

163
\[
\frac{\partial E_c}{\partial \tau_{\delta}^{(2),i}} = -E_c \left( \frac{\partial \hat{C}_c}{\partial \tau_{\delta}^{(2),i}} + \hat{C}_c (I - A_c)^{-1} \frac{\partial A_c}{\partial \tau_{\delta}^{(2),i}} (I - A_c)^{-1} K_c + \hat{C}_c (I - A_c)^{-1} \frac{\partial K_c}{\partial \tau_{\delta}^{(2),i}} \right) E_c \text{ for other components of } \tau_{\delta}^{(2)}
\]

Of course, the derivatives of \( D_c \) are trivially obtained by

\[
\frac{\partial D_c}{\partial \tau_{\delta}^{(2),i}} = \begin{pmatrix} 1 \end{pmatrix} \text{ at position } (i,j) \\
\frac{\partial D_c}{\partial \tau_{\delta}^{(2),i}} = 0_{m \times m} \text{ for other components of } \tau_{\delta}^{(2)}
\]

Having calculated the derivatives of \( \hat{C}_c \) and \( E_c \), we can get the derivatives of \( C_c \) by

\[
\frac{\partial C_c}{\partial \sigma_i} = \frac{\partial E_c}{\partial \sigma_i} \hat{C}_c + E_c \frac{\partial \hat{C}_c}{\partial \sigma_i} \\
\frac{\partial C_c}{\partial K_{ij}} = \frac{\partial E_c}{\partial K_{ij}} \hat{C}_c + E_c \frac{\partial \hat{C}_c}{\partial K_{ij}} \\
\frac{\partial C_c}{\partial \phi_{ij}} = \frac{\partial E_c}{\partial \phi_{ij}} \hat{C}_c + E_c \frac{\partial \hat{C}_c}{\partial \phi_{ij}} \\
\frac{\partial C_c}{\partial b_{ij}} = 0_{n \times n} \\
\frac{\partial C_c}{\partial d_{ij}} = 0_{n \times n}
\]

The derivative \( \frac{\partial U_c}{\partial \tau_{\delta}^{(2),i}} \) is again given in (5.57). Note that now both the \( A_c \) and \( C_c \) matrix depend on the minimum phase singular values such that we have to use (5.55) for both matrices!

### Derivatives for DDLC

For any fixed minimal \( (A, B, C, D, K) \), the mapping \( \varphi_D \) from the parameter vectors \( \tau_D \in T_D \) to the state-space matrices is affine as can be seen from (4.74) in definition (4.9.1). We immediately obtain

\[
\frac{d \varphi_D}{d \tau_D} = Q^L.
\]

### Derivatives for slsDDLC

It is seen from (4.83) in definition (4.10.1) that the mapping \( \varphi_D^\sigma \) from the parameter vectors \( \tau_D^\sigma \in T_D^\sigma \) to the state-space matrices \( (A, B, C, D, K) \) is in general non linear because the mapping \( \Delta^U \) is non linear; see (3.27) in lemma (3.6.1). However, we have seen in sections (3.3.2) and (3.3.3) that the gradient and a Gauss-Newton approximation of the Hessian of \( L^\sigma_c (Y^T_c, U^T_c, \tau_D^\sigma) \) can be computed when only \( \hat{A} \) and \( \hat{C} \) are known. Note that these quantities are directly obtained from the corresponding rows of \( Q^L_c \) in (4.83):

\[
\frac{d \text{vec}(A), \text{vec}(B), \text{vec}(C), \text{vec}(D)^T}{d \tau_D^\sigma} = Q^L_c.
\]

### 5.5 A few notes on optimization algorithms

After the choice of an appropriate model class and parametrization, maximum likelihood estimation leads to the problem of minimizing a function of a number of parameters over some open subset of an Euclidean space. Apart from special cases, there are no analytic formulae for the solution to this optimization problem available, such that we have to employ some kind of numerical optimization procedure.
One can, of course, find an enormous number of different optimization techniques in the literature. We could think of rather recent approaches such as genetic algorithms or, to be more specific, evolutionary strategies, or particle swarm algorithms. These methods can be used even if only very little knowledge about qualitative properties of the criterion function (e.g. differentiability) is available.

For all parametrizations considered in this thesis, the parameter spaces $T$ have been defined in chapter 4 in such a way that likelihood type criterion functions are smooth when considered as functions from $T$ into the real numbers. Note that at boundary points of $T$ (which do not belong to $T$), the likelihood function even might not be continuous; see, e.g. (Deistler and Poetscher, 1984). This smoothness allows us to consider rather traditional iterative optimization algorithms such as gradient methods, sometimes called "steepest descent methods" for obvious reasons, Newton’s method, quasi-Newton methods, conjugate gradient (CG) methods, trust region methods, the method of Gauss-Newton or its extension, the Levenberg-Marquardt algorithm, just to mention a few. As is well known and will become clear from the discussion below, these search procedures typically converge to a local minimum of the criterion function, which is an apparent disadvantage if several local minima different from the global minimum exist.

We will now briefly sketch four of these traditional methods. To start with, let $\mathcal{D}$ be an open subset of $\mathbb{R}^d$ and let the mapping

$$
\Phi : \mathcal{D} \to \mathbb{R}
$$

$$
x \mapsto \Phi(x)
$$

be twice continuously differentiable. Assume that $x_0$ is a global minimizer of $\Phi$ and that the other local minima of $\Phi$ are isolated minima. The optimization problem is then to find the global minimum $x_0$ of $\Phi$ over the domain $\mathcal{D}$. A necessary condition for any minimum $\bar{x}$ is that the gradient vanishes at $\bar{x}$. To simplify the notation a bit, let us denote the gradient $\nabla \Phi(x) \in \mathbb{R}^d$ by $f(x)$. Then the problem is to find the solution to the nonlinear system of equations given by $f(x) = 0$ that corresponds to the global minimum of $\Phi$, where

$$
f : \mathcal{D} \to \mathbb{R}^d
$$

$$
x \mapsto f(x)
$$

Clearly, the equation $f(x) = 0$ can be rewritten as a fixed point problem of the form

$$
x = \mu M^{-1}f(x) = x
$$

$$
\rightarrow \Phi(x)
$$

(5.58)

where the scalar step size parameter $\mu \in (0, 1)$ and the nonsingular matrix $M \in GL(d)$ can be freely chosen. If there exists a compact and convex subset $\overline{\mathcal{D}} \subset \mathcal{D}$ such that $\varphi(\overline{\mathcal{D}}) \subset \overline{\mathcal{D}}$, $x_0 \in \overline{\mathcal{D}}$ and $\varphi|_{\overline{\mathcal{D}}}$ is a contraction mapping, i.e. $||\varphi(x_1) - \varphi(x_2)|| \leq L \cdot ||x_1 - x_2||$ for an arbitrary vector norm $||\cdot||$, some Lipschitz constant $L < 1$ and all $x_1, x_2 \in \overline{\mathcal{D}}$, then by Banach’s fixed point theorem (see, e.g. (Heuser, 1992)), the iteration

$$
x_{k+1} = \varphi(x_k)
$$

(5.59)

converges to the unique fixed point $x_0 \in \overline{\mathcal{D}}$. Note that the application of the mean value theorem yields

$$
||\varphi(x_1) - \varphi(x_2)|| = ||\int_0^1 d\varphi(x_1 + \lambda(x_2 - x_1))d\lambda|| (x_1 - x_2) ||
$$

$$
\leq \int_0^1 ||d\varphi(x_1 + \lambda(x_2 - x_1))||d\lambda \cdot ||x_1 - x_2||
$$

$$
\leq \max_{x \in \overline{\mathcal{D}}} ||d\varphi(x)|| \cdot ||x_1 - x_2||
$$

For the following methods to be applicable, it suffices to require the criterion function to be twice (or even just once) continuously differentiable.
where $d\varphi(x) \in \mathbb{R}^{d \times d}$ denotes the derivative of $\varphi$. Therefore, one can choose $L = \max_{x \in \mathbb{R}^d} ||d\varphi(x)||$ if this quantity is less than one. It is clear now that $||x_{k+1} - x_0|| = ||\varphi(x_k) - \varphi(x_0)|| \leq L||x_k - x_0|| \leq \cdots \leq L^k||x_1 - x_0|| \to 0$ for $k \to \infty$, where the convergence behaviour is linear, i.e. $\frac{||x_{k+1} - x_0||}{||x_k - x_0||} \leq L < 1$.

The main observation is that convergence becomes faster if $L$ becomes smaller, and that therefore the choice of $\mu$ and $M$ should aim at making $L$ as small as possible. Note that

$$d\varphi(x) = I - \mu M^{-1} df(x)$$

(5.60)

if $M$ is considered to be constant. Clearly, $df(x) \in \mathbb{R}^{d \times d}$ denotes the matrix of second derivatives, i.e. the Hessian, of $\Phi(x)$. In fact, we can now distinguish some well known algorithms by their choice of $\mu$ and $M$:

(a) Choosing $\mu = 1$ and $M = I$ yields a simple gradient algorithm of the form

$$x_{k+1} = x_k - f(x_k)$$

(5.61)

Recall that $f(x)$ is the gradient of the function $\Phi(x)$ that we want to minimize. Another possibility, yet admittedly a bit academic, is that we want to use a gradient algorithm but can freely choose $\mu$ when starting the search algorithm, leaving it fixed afterwards. If one has some a priori approximations of the quantities $\lambda_{\text{max}}(df(x_0))$ and $\lambda_{\text{min}}(df(x_0))$ (without necessarily having knowledge of $df(x_0)$ itself), then an optimal choice for $\mu$ is $\mu_{\text{opt}} = \frac{\lambda_{\text{max}}(df(x_0)) + \lambda_{\text{min}}(df(x_0))}{2}$; see, e.g. chapter 12.2 in (Gevers and Li, 1993). In such a situation it is clear that

$$L \geq ||d\varphi(x_0)|| \geq \frac{\lambda_{\text{max}} \left( I - \frac{2}{\lambda_{\text{min}}(df(x_0)) + \lambda_{\text{max}}(df(x_0))} \cdot df(x_0) \right) \right|}{\lambda_{\text{min}}(df(x_0)) + \lambda_{\text{max}}(df(x_0))} = 1 - \frac{2 \cdot \lambda_{\text{min}}(df(x_0))}{\lambda_{\text{min}}(df(x_0)) + \lambda_{\text{max}}(df(x_0))} = \frac{\lambda_{\text{max}}(df(x_0)) - 1}{\lambda_{\text{min}}(df(x_0)) + 1} = \frac{\kappa(df(x_0)) - 1}{\kappa(df(x_0)) + 1}$$

(5.62)

because the spectral radius $\lambda_{\text{max}}(X)$ is a lower bound on any matrix norm of $X$ and $df(x_0)$ is clearly symmetric and positive definite by assumption (because $x_0$ is an isolated minimum), implying that the condition number $\kappa(df(x_0))$ equals the quotient of its largest over its smallest eigenvalue.

(b) Choosing $\mu = 1$ and $M = df(x)$ yields Newton’s algorithm of the form

$$x_{k+1} = x_k - [df(x_k)]^{-1} f(x_k)$$

(5.63)

Note that $M$ is a function of $x$ now. For this choice of $\mu$ and $M$, we see from (5.58) that $df(x) \cdot (\varphi(x) - x) + f(x) = 0$. Differentiating once more and noting that $d^2 f(x)$ actually represents a (bilinear) mapping from $\mathbb{R}^d \times \mathbb{R}^d$ into $\mathbb{R}^d$, we get $d^2 f(x)(\varphi(x) - x, \cdot) + df(x)(d \varphi(x) - I) + df(x) = d^2 f(x)(\varphi(x) - x, \cdot) + df(x) \cdot d \varphi(x) = 0$ and at $x = x_0$, we obtain $df(x_0) \cdot d \varphi(x_0) = 0$, implying that $d \varphi(x_0) = 0$ as $df(x) > 0$ by assumption. The Lipschitz constant $L$ therefore becomes smaller and smaller in a neighborhood of $x_0$ and we get a quadratic convergence behaviour. This is seen by again using the mean value theorem and the fact that $d \varphi(x_0) = 0$:
\[ ||x_{k+1} - x_0|| = ||\varphi(x_k) - \varphi(x_0)|| \]
\[ = || \left( \int_0^1 d\varphi(x_0 + \lambda(x_k - x_0))d\lambda \right)(x_k - x_0)|| \]
\[ \leq \int_0^1 ||d\varphi(x_0 + \lambda(x_k - x_0)) - d\varphi(x_0)||d\lambda \cdot ||x_k - x_0|| \]
\[ \leq \int_0^1 L'||\lambda(x_k - x_0)||d\lambda \cdot ||x_k - x_0|| \]
\[ \leq L' \int_0^1 \lambda d\lambda \cdot ||x_k - x_0||^2 \]
\[ = \frac{L'}{2} \cdot ||x_k - x_0||^2 \]  

where \( L' \) is a Lipschitz constant of \( d\varphi(x) \) in \( \bar{D} \) (which need not be less than one). Despite this appealing property of quadratic convergence, Newton’s method also faces some difficulties:

1. The matrix \( M = df(x) \) has to be updated in each iteration. Note that \( df(x) \) is the matrix of second derivatives of \( \Phi(x) \) and that it may be very costly to compute this matrix in practice.

2. The Newton step in (5.63) becomes very large for large values of \( ||df(x)|^{-1}|| \), possibly causing the iteration to leave the set \( D \) and even yield \( ||f(x_{k+1})|| \geq ||f(x_k)|| \). This is the case if \( df(x) \) is almost singular. A possible approach to cope with this situation is to adjust the step size \( \mu \) by the additional step size parameter \( \mu \in (0,1) \), e.g. by trying \( \mu \in \{1/2,1/4,1/8,\ldots\} \). This yields the so called damped Newton method. Moreover, note that the Newton update is not defined at all if \( df(x) \) is singular.

3. From a rather global point of view (i.e. considering \( D \) and not \( \bar{D} \)), we remark that a non singular \( df(x_k) \) need not necessarily be positive definite. This can result in update directions that are not descent. Note that this is something qualitatively different from an increase in the criterion value due to too large step sizes as discussed in (2) above.

4. Even if the Newton step does not appear to be very large, we could still have \( ||f(x_{k+1})|| \geq ||f(x_k)|| \) if the Lipschitz constant \( L' \) of \( d\varphi \) is very large. However, this can again be “cured” by a damped Newton algorithm.

We remark that (5.63) can also be interpreted as the unique (in case of a non singular \( df(x_k) \)) solution to the linear least squares problem of minimizing \( ||f(x_k) + df(x_k)(x - x_k)||^2 \). From this point of view, each iteration of Newton’s algorithm first differentiates the criterion \( \Phi(x) \) to get \( f(x) \) and then solves the linear least squares problem obtained from a linearization of \( f(x) \).

(c) Clearly, one could also choose \( \mu = 1 \) and \( M = G(x) \) where \( G(x) \) denotes some approximation of \( df(x) \) based on first order derivatives \( f(x) \) only. We will briefly illustrate this idea for criterion functions of nonlinear least squares type, i.e. where

\[ \Phi(x) = ||E(x)||^2 = E'(x)E(x) \quad \text{and} \quad E : D \rightarrow \mathbb{R}^N, x \mapsto E(x) \]

is a function attaching the vector \( E(x) \in \mathbb{R}^N \) to \( x \in D \). As an example, consider the criterion function \( L_2^2(Y_1^T : U_1^T, \tau^o) \) for given observations \( Y_1^T \) and \( U_1^T \) in (2.85). Clearly, \( L_2^2(\tau^o) = ||E_1^T(\tau^o)||^2_2 \) where \( E_1^T(\tau^o) \in \mathbb{R}^T \); see (2.81) and (5.48). The gradient \( f(x) \) of \( \Phi(x) \) is then given (up to a constant factor of two) by \( f(x) = J'(x)E(x) \) where \( J(x) = dE(x) \in \mathbb{R}^{N \times d} \) denotes the Jacobian matrix of \( E(x) \).

Choosing \( \mu = 1 \) and \( M = G(x) = J'(x)J(x) \) yields the method of Gauss-Newton of the form

\[ x_{k+1} = x_k - G(x_k)^{-1}f(x_k) = x_k - [J'(x_k)J(x_k)]^{-1}J'(x_k)E(x_k) \]  

(5.65)
Here, we view $G(x)$ as an approximation to the Hessian $df(x)$ of the original criterion function $\Phi(x)$. Note that $G(x_0) = df(x_0)$ if the criterion value at $x_0$ is zero (or, trivially, if $E(x)$ is linear in $x$). If this is the case, the local convergence property near $x_0$ is the same as for Newton’s algorithm.

Note that formula (5.65) can also be interpreted as the solution to the problem where $E(x)$ is first directly linearized to get $E(x) = E(x_k) + J(x_k)(x - x_k) + O(||x - x_k||^2)$ and then the corresponding linear least squares problem of minimizing $||E(x_k) + J(x_k)(x - x_k)||^2_2$ is considered. Note the different order of the two steps as compared to Newton’s method.

We make the following remarks:

1. The matrix $G(x)$ still has to be updated in each iteration. However, it can be computed by using first order derivatives only.

2. The Gauss-Newton step can also become too large for large $||G(x_k)^{-1}||$, again necessitating the use of a step size controlling parameter $\mu \in (0,1)$. Note that the interpretation of the Gauss-Newton step (5.65) as the solution to the linear least squares problem of minimizing $||E(x_k) + J(x_k)(x - x_k)||^2_2$ enables us to come up with a well defined step even in the case where $G(x_k)$ is singular (i.e. $J(x_k)$ does not have full column rank): Among all solutions to the linear least squares problem, we could take the unique minimum norm solution.

3. From a rather global point of view (i.e. considering $D$ and not $D'$), we remark that any non singular $G(x_k)$ is positive definite by construction. This guarantees that the update directions will always be descent.

Instead of (artificially) introducing a step size controlling parameter and instead of considering the minimum norm solution to the problem of minimizing $||E(x_k) + J(x_k)(x - x_k)||^2_2$ in case of a singular $G(x_k) = J'(x_k)J(x_k)$, one might proceed as in (d):

(d) Choosing $M = G(x) + \mu DD'$ yields the Levenberg-Marquardt algorithm of the form

$$x_{k+1} = x_k - [G(x_k) + \mu_k D_k D_k^t]^{-1} f(x_k) = x_k - [J'(x_k) J(x_k) + \mu_k D_k D_k^t]^{-1} J'(x_k) E(x_k)$$

(5.66)

Here, $\mu_k \geq 0$ is called the Levenberg-Marquardt parameter and $D_k$ denotes a non-singular scaling matrix, so that $||D_k(x - x_k)||_2$ represents some norm of the step $x - x_k$.

Note that formula (5.66) can be interpreted as the solution to the linear least squares problem obtained from balancing the criterion $||E(x_k) + J(x_k)(x - x_k)||^2_2$ and the second objective of keeping the step size small, i.e. (5.66) is the solution to the linear least squares problem of minimizing $||E(x_k) + J(x_k)(x - x_k)||^2_2 + \mu_k ||D_k(x - x_k)||^2_2$. If $\mu_k > 0$, the solution is always unique. We may choose $D_k = I$, while adapting $\mu_k$ during the search procedure. A quite sophisticated way of doing this is to accept Gauss-Newton steps if they are inside a trust region (the Levenberg-Marquardt parameter is then set to zero) or, if this is not the case, to determine $\mu_k$ such that the resulting step lies (approximately) on the boundary of the trust region. This trust region is itself defined as the region where the actually achieved improvement of the criterion function is in appropriate accordance with the improvement of the linearized criterion $||E(x_k) + J(x_k)(x - x_k)||^2_2$. Clearly, this trust region is adapted during the search procedure.

Finally, it is worth noting that in order to obtain $x_{k+1}$ in (5.63), (5.65) or (5.66), one formally has to invert $M = M(x)$, i.e. the Hessian of $\Phi(x)$ or some approximation of it. In actual computations, of course, this inversion is avoided by solving the system of linear equations of the form

$$M(x_k) \cdot \Delta x_{k+1} = f(x_k) \quad \text{where} \quad \Delta x_{k+1} = x_{k+1} - x_k$$

(5.67)

e.g. by LU factorization.
5.6 Numerical considerations for various parametrizations

In section (5.5) above, we have briefly discussed some important traditional optimization methods. This was also done in order to be now able to motivate the importance of the Fisher information matrix for numerical properties of maximum likelihood parameter estimation algorithms.

Note that for the asymptotic likelihood function (2.87), the matrix of second derivatives at the true parameter values has been computed in (5.42) in section (5.2.4); (5.42) yields the \((k,l)\) element of the well known Fisher information matrix.

For the finite sample likelihood function, e.g. \(L_T(Y^T_1; U^T_1, (\tau, \sigma))\) in (2.76), it is well known that the second derivatives at the true parameters \((\tau_0, \sigma_0)\) converge to the Fisher information matrix if the number \(T\) of observations becomes larger and larger; note that we only consider the derivatives with respect to system parameters \(\tau\) in the sequel.

In accordance with the discussion in section (5.5) above, we shall denote the (approximation to the) matrix of second derivatives by \(M(\tau_k)\), where \(M(\tau_k)\) either denotes the true matrix of second derivatives in case of Newton’s algorithm (see (5.63)) or the approximations given in (5.65) or (5.66).

Concerning the role of \(M(\tau_k)\), we can now make the following remarks:

(i) **Accuracy of the parameter estimate:** It is well known in asymptotic statistical analysis that the inverse of the Fisher information matrix yields the (asymptotic) Cramer-Rao lower bound for asymptotically unbiased estimators. In this sense, it clearly determines the possible accuracy of the parameter estimates. From the numerical point of view, a large condition number of its finite sample approximations \(M(\tau_k)\) leads to ill-conditioned linear least squares problems in the iteration steps. This pops up when solving (5.67): Note that the relative error of the solution \(\Delta x_{k+1}\) to this system of linear equations (measured in Euclidean norm) is proportional to the condition number \(\kappa(M(\tau_k))\). In actual computations, \(M(\tau_k)\) and \(f(\tau_k)\) are inevitably subject to roundoff errors, and a high condition number may yield to wrong search directions and thereby corrupt the performance of the algorithm.

(ii) **Speed of convergence:** For the gradient algorithm discussed in (a) in section (5.5) above it is seen that the Lipschitz constant of \(\phi(\tau) = \tau - \mu_{opt}(\tau)\) is bounded from below by \(\frac{\kappa(M(\tau_0))^{-1}}{\kappa(M(\tau_0)) + 1}\), where \(M(\tau_0)\) here denotes the exact matrix of second derivatives; see (5.62). Hence, the convergence speed of the corresponding gradient algorithm may well decrease if \(\kappa(M(\tau_0))\) becomes large.

(iii) **Change of correct components of the parameter vector:** A trivial observation is that if \(M(\tau_k)\) is not diagonal and if one component of \(\tau_k\) has already reached its optimal value, then in the next iteration step this component can be moved away from its optimal value after solving (5.67).

With these observations in mind, we will now briefly investigate the properties of the Fisher information matrix for various different parametrizations. Recall that we have computed the Fisher information matrix symbolically (for very simple cases); for the case \(n = s = 1\) and \(m = 0\), see table (5.1), for instance. In figure (5.1) we also consider the case \(n = s = 1\) and \(m = 0\) and plot the condition number of the Fisher information matrices. There are a few remarks to be made for figure (5.1):

(i) First, observe that no parametrization is uniformly better over \(M_{amp}(n)\) in terms of the condition number of the Fisher information matrix. Note that we consider the Fisher information matrix at the true parameter values \((\tau_0, \sigma_0)\), implying that \(\tau_0\) always corresponds to \(\tau_0^0 = 0\) for the DDLC parametrization. As an initial system for the DDLC parametrization, we always considered Ober’s Lyapunov balanced representative. As will be discussed below, the condition number of the Fisher information matrix for DDLC strongly depends on the choice of the initial representative within the \(k\)-equivalence class \(\mathcal{E}(a, k, \epsilon)\); see observation (5.61.1) below for a more precise statement. Of course, one might now argue that the Lyapunov balanced initial representative is a bad choice and that DDLC could probably be made superior to all the other parametrizations considered by a clever (data-driven) choice of other representatives. However, it turns out (see the discussion of plots (a) and (b) of figure (5.3) below) that this is not the case here. In fact, the choice of an initial system in Ober’s Lyapunov balanced canonical form will guarantee a condition number as low as possible in
many cases. Note that none of the parametrizations treated in this thesis are explicitly constructed such that the conditioning of the matrix $M(\tau_k)$ is optimal; see remark (5.6.1) below.

(ii) From figure (5.1) it becomes apparent that the conditioning of the Fisher information matrix can become very bad for Ober's Lyapunov balanced canonical forms. This might at first sight be a bit surprising as there are a number of indications in the literature that Lyapunov balanced canonical forms have certain optimality or at least desirable properties from a numerical point of view. We just want to comment on some of these properties here:

- Finite word length (FWL) effects on the input signal of a linear system clearly yield to what is called a quantization error in the field of digital signal processing. These errors are usually modeled as an additive disturbance on the input signal with known variance\(^4\) and cause a disturbance of the output signal. The quotient of the variance of this output disturbance over the variance of a single quantization operation is known as the roundoff noise gain of the realization. In (Mullis and Roberts, 1976), the authors treat the problem of finding state-space realizations for single-input, single-output systems that have an optimal roundoff noise gain and additionally satisfy some scaling constraint on the state sequence (to prevent overflows). They do this in a fixed point arithmetic setting, considering two different situations: First, they assume that equally many bits are available for each component of the state vector. This yields to optimal realizations characterized by $Q = (\frac{1}{N} \sum_{i=1}^{N} \sigma_i)^2 P$, where $P$ (see (3.16)) and $Q$ (see (3.17)) are the controllability Gramian and the observability Gramian of the realization, respectively, and $\sigma_i$ denotes the $i$. Hankel singular value of the corresponding transfer function. Note that these realizations will in general not be Lyapunov balanced and that the proportionality factor is determined by the underlying transfer function only. In a second situation, they assume that the number of available bits can be distributed freely among the components of the state vector, and in this case the optimal realizations are the ones where both Gramians are simultaneously diagonal (but not necessarily equal). Hence, Lyapunov balanced state-space realizations belong to the set of optimal systems in this case. Note that in this setting, the roundoff noise on the output is due to the quantization of the signals and the errors occurring in the (fixed point) arithmetic operations needed to compute the output sequence. Clearly, all entries in the state-space matrices contribute to these FWL errors (and not only the free parameters one would need to describe the realizations).

- In (Thiele, 1986), the following sensitivity measure is minimized over the set of observationally equivalent minimal state-space systems (for single-input, single-output transfer functions):

$$L_{12}(A,b,c,d) = \left| \frac{\partial l(z)}{\partial A} \right|_{1}^{2} + \left| \frac{\partial l(z)}{\partial b} \right|_{2}^{2} + \left| \frac{\partial l(z)}{\partial c} \right|_{2}^{2} + 1 \quad (5.68)$$

where $l(z)$ denotes the transfer function under consideration. Note that (5.68) is in fact a mixture of two norms. Clearly,

$$\left| \frac{\partial l(z)}{\partial A} \right|_{1} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\partial l(e^{i\omega})}{\partial A} \right| d\omega \quad \text{and} \quad \left| \frac{\partial l(z)}{\partial b} \right|_{2} = \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\partial l(e^{i\omega})}{\partial b} \right|^{2} d\omega \right)^{1/2}$$

where the third term is given analogously and the Frobenius norm of a vector apparently coincides with the usual Euclidean norm. The $L_{12}$-optimal realizations are shown to be the ones where $P = Q$ holds, i.e. the controllability Gramian and the observability Gramian are equal, but not necessarily diagonal. Note that the class of Lyapunov balanced state-space realizations is a subset of these optimal realizations. From (5.68) it is clear that again sensitivities with respect to all matrix entries are considered here.

- In (Helmke and Moore, 1994), the authors have replaced (5.68) by a "pure" $L_2$-norm. In fact, they introduce two modified Gramian matrices called $L_2$-sensitivity Gramians \(^5\) $\bar{P}$ and $\bar{Q}$, say. Then the $L_2$-optimal realizations can be shown to be the ones where $\bar{P} = \bar{Q}$. These realizations are generally not Lyapunov balanced and have to be computed, e.g. by gradient

\(^4\)Of course, this variance depends on the number of bits used for the encoding of the input sequence.

\(^5\)The authors of (Helmke and Moore, 1994) have also introduced Euclidean norm Gramians $\bar{P} = AA' + BB'F$ and $\bar{Q} = A'A + C'C$. Then, the realizations of minimum Frobenius norm (which the authors call Euclidean norm) are characterized by $P = Q$; see section (3.3) in this thesis.
flow techniques. For a discretized version of this technique, see, e.g. (Gevers and Li, 1993). However, according to the authors of the latter work, numerical simulations indicate that realizations which are $L_1$-optimal (as e.g. Lyapunov balanced realizations) are also "nearly $L_2$-optimal". Recall once more that also $L_2$-sensitivity is computed with respect to all entries in the state-space matrices.

- In (Gevers and Li, 1993), the $L_2$ sensitivity criterion is also combined with the roundoff noise gain as introduced in (Mullis and Roberts, 1976) above to yield what the authors call the total noise gain. However, the optimal realizations with respect to this criterion function again have to be computed by an iterative scheme, and there is no immediate connection to Lyapunov balanced realizations.

- An argument which is more closely related to parameter estimation can be found in (Chou and Maciejowski, 1997) and (Maciejowski, 1985), the latter paper being the place where the notion of using balanced realizations as canonical forms was actually first introduced. The authors argue that the use of Lyapunov balanced parametrizations results in estimation problems as well conditioned as possible for a given transfer function. The reasoning is that controllability and observability are equally important in any Lyapunov balanced state-space realization and that if one measures the distance of the Gramians $P$ and $Q$ in (3.16) and (3.17) to the closest singular matrix, Lyapunov balanced parametrizations clearly maximize $\min(\sigma_{\min}(P), \sigma_{\min}(Q))$ where $\sigma_{\min}(X)$ stands for the smallest singular value of $X$. In this sense, Lyapunov balanced canonical forms give realizations which are *farthest away from non minimality*. However, note that the distance of the Gramians $P$ and $Q$ to the closest singular matrix is not measured in the parameter space $T_k$ here. Note, moreover, that the mapping from the parameters to the values of some criterion function may also be very distorting even if the parameters do not correspond to a system which is close to being non minimal. This is in fact the case for Ober's Lyapunov balanced canonical form in figure (5.1).

**Remark 5.6.1.** Of course, it is always possible to construct a local parametrization such that the (approximation to the) matrix of second derivatives of the criterion function at the current estimate is the identity matrix and thus optimally conditioned. In (McKelvey, 2002), this is done for the case where only the transfer function $l(z)$ is considered and the criterion function is given by $V(\tau) = \sum_{i=1}^{2n} ||L_i - L_i(\tau)||^2$, where $L_i$ denotes the measured $i$-th impulse response coefficient of the true underlying system and $L_i(\tau)$ is the $i$-th impulse response coefficient of the parametrized state-space model; $n$ denotes the McMillan degree of the transfer function corresponding to $(L_i(\tau))_{i=1}^N$. This parametrization is called "OrthImp". It is again affine and easy to implement; see figure (5.4). In principle, a similar approach is also possible in our maximum likelihood setting. However, such a transformation of the parameter space might be costly to obtain and the transformation itself can lead to ill-conditioned operations, thus just shifting the problem into another place. Note, moreover, that a reparametrization can in general only *locally* (i.e. in a neighborhood of the current estimate) yield a Fisher information matrix that is close to the identity; see also remark (5.6.2) below.

**Remark 5.6.2.** In chapter 12 of (Gevers and Li, 1993), the authors apply a similar idea for the estimation of ARX models in the single-input, single-output case. They use a so called $\gamma$-operator parametrization which represents the underlying transfer functions using the basis functions $\gamma(z) = \frac{a(z-\Delta)}{z^\Delta + 1}$ instead of $z$, where $\Delta > 0$ denotes some positive constant. The authors claim that this parametrization yields a Fisher information matrix with zero odd off-diagonal elements, leading to a (drastic) improvement of the condition number *irrespective* of the (spectral properties of) the data process and the unknown system. This is then illustrated by means of an example using a FIR model. However, in the real ARX case, this property is not valid any more. The reason for this is as follows: Consider an ARX model of the form $y_t = -a_1 y_{t-1} - \cdots - a_p y_{t-p} + b_0 u_t + b_1 u_{t-1} + \cdots + b_q u_{t-q} + \varepsilon_t$ which the authors write as $y_t = \theta'_r \psi_r(t) + \varepsilon_t$, where $\theta'_r = (a_1, \ldots, a_p, b_0, \ldots, b_q)$ and $\psi_r(t) = (y_{t-1}, \ldots, y_{t-p}, u_0, \ldots, u_{t-q})$. Then, the use of the $\gamma$-operator parametrization transforms the output $y_t$ and the regressor $\psi_r(t)$ into the filtered quantities denoted by $y_t^\gamma$ and $\psi_r(t)$ in (Gevers and Li, 1993). However, the new output $y_t^\gamma$ and the new regressor $\psi_r(t)$ now also depend on $y_t$ such that the model $y_t^\gamma = \theta'_r \psi_r(t) + \varepsilon_t$ is no longer a regression model in the sense that $\varepsilon_t$ is not orthogonal to the regressor $\psi_r(t)$ any more. This implies that the Fisher information matrix is *not* given by the expression in theorem 12.2 in (Gevers and Li, 1993) as is claimed there.

As already mentioned, the condition number of the Fisher information matrix for DDLC can strongly
depend on the choice of the initial representative within the \( k \)-equivalence class \( \mathcal{E}(a, k, c) \). This is illustrated by the plots (a) and (b) of figure (5.3), where again the case \( n = s = 1 \) and \( m = 0 \) is treated for two different transfer functions. We make the following remarks:

(i) Note that the condition numbers for DDLC seem to explode in both plot (a) and (b) for \( t \to \infty \) (corresponding to initial systems with large \( |k| \) and small \( |c| \)) and \( t \to 0 \) (corresponding to initial systems with large \( |c| \) and small \( |k| \)). This conjecture will be shown analytically in the discussion of observation (5.6.1) below. Note that it directly follows from considering the behaviour of the eigenvalues of the Fisher information matrices as given in plots (c) and (d), respectively. Whereas one eigenvalue becomes constant for \( t \to \infty \) and \( t \to 0 \), the other one seems to diverge to infinity.

(ii) Observe that applying the DDLC construction at Ober’s Lyapunov balanced initial system either yields a minimal – see plot (a) – or a locally maximal condition number; see plot (b). Again, this follows directly from considering the behaviour of the eigenvalues of the Fisher information matrices as given in plots (c) and (d), respectively. One eigenvalue reaches its minimum at the Lyapunov balanced realization (corresponding to \( t = t_{bal} \)) and increases monotonically in \( |t - t_{bal}| \), i.e. "at both sides" of the Lyapunov balanced realization. The other eigenvalue remains almost constant\(^6\) for varying \( t \), and therefore it depends on the size of this eigenvalue at the Lyapunov balanced system whether the quotient of the larger over the smaller eigenvalue (the eigenvalues and singular values coincide for symmetric positive (semi)definite matrices) is minimal – see plot (a) – or locally maximal; see plot (b).

(iii) Clearly, it can be the case that the DDLC parametrization yields larger condition numbers than the other parametrizations for any choice of initial system; see plot (a). This is also in accordance with figure (5.1) because at \((a, k, c) = (0.1, 1, 1)\) DDLC is worst even if we apply the DDLC construction at the Lyapunov balanced realization.

(iv) For plot (d), on the other hand, optimal conditioning can be achieved by a suitable choice of initial systems, and all other parametrizations will then clearly be inferior in that respect. Note that the condition number for Ober’s Lyapunov balanced canonical form is rather large in plot (d); compare figure (5.1) for \((a, k, c) = (0.9, 0.4, 1)\).

An intuitively appealing, yet not entirely correct, explanation of (i) and (ii) above is the following: The sensitivity of the asymptotic likelihood function with respect to changes in the DDLC parameter spaces in the \( a \)-direction remains constant along the equivalence class \( \mathcal{E}(a, k, c) \). This is because the distance between different hyperbolae (representing different transfer functions in \( \mathbb{M}(1) \)) stays constant in this direction. On the other hand, the sensitivity in the orthogonal direction of the DDLC parameter spaces is minimal for the DDLC parametrization applied at Ober’s Lyapunov balanced realization and monotonically increases "at both sides" of this realization as the hyperbolae approach each other. Note that this "sensitivity" has to be attributed to second order derivatives of the asymptotic likelihood function – the first order derivatives are zero as we consider the asymptotic likelihood function at the true parameter values.

We now continue discussing the dependence of the (condition number of the) Fisher information matrix on the choice of the initial system for DDLC in a slightly more general way. Let us recall that in statement (iii) of theorem (4.3.2) it was shown that the set \( \mathbb{M}(n) \) has the structure of a real analytic manifold of dimension \( 2ns + m(n + s) \). It is easy to see that \( \mathbb{M}_{\text{amp}}(n) \) is an open subset of \( \mathbb{M}(n) \), and therefore a submanifold. \( \mathbb{M}_{\text{amp}}(n) \) can be endowed with a differentiable Riemannian metric tensor – see the appendix (A.6) – and the Fisher information matrix (evaluated at the true parameter values) is one example for such a Riemannian metric tensor for \( \mathbb{M}_{\text{amp}}(n) \); see, e.g., (Hanzon, 1989) and the references there.

**Observation 5.6.1 (Choice of initial systems in \( \mathcal{E}(A, B, C, D) \) for DDLC).** Let \( \{ \varphi_T^T, T \in GL(n) \} \) be the following family of DDLC parametrizations

\(^6\)Note that the colours of the two curves are a bit misleading in plot (d) of figure (5.3); the parabola-type curve should be entirely blue and the rather straight line should be entirely white.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FullPar</strong></td>
<td>[ \begin{pmatrix} \frac{c^2 k (a - c)}{c k} &amp; \frac{c^2 k (-a + c)}{c k} \ \frac{-1 + a^2 - a c}{c k} &amp; \frac{-1 + a^2 - a c}{c k} \end{pmatrix} ]</td>
</tr>
<tr>
<td><strong>Echelon (OVL)</strong></td>
<td>[ \begin{pmatrix} \frac{(\tau_a^1 - \tau_a^2)}{\tau_a^2} &amp; \frac{\tau_a^1}{\tau_a^2} \ \frac{(\tau_a^2 - \tau_a^1)}{(1 + \tau_a^2 - \tau_a^1)} &amp; \frac{(\tau_a^2 - \tau_a^1)}{(1 + \tau_a^2 - \tau_a^1)} \end{pmatrix} ]</td>
</tr>
<tr>
<td><strong>LyapBAL (Ober)</strong></td>
<td>Expressions in Mathematica available, but too lengthy in order to be presented here...</td>
</tr>
<tr>
<td><strong>StochBAL (Ober)</strong></td>
<td>Not implemented in Mathematica...</td>
</tr>
<tr>
<td><strong>MinBAL (McGinnie)</strong></td>
<td>Expressions in Mathematica available, but too lengthy in order to be presented here...</td>
</tr>
<tr>
<td><strong>DDLC (a, k, c) = (0.5, 1, 1)</strong></td>
<td>( (1, 1) = \frac{-40 + 32 (\tau_a^2)^2 - 8 (\tau_a^1)^2 + 4 (\tau_a^1)^2}{(1 + \tau_a^2)^2} )</td>
</tr>
<tr>
<td></td>
<td>( (1, 2) = \frac{4 \sqrt{2} (\tau_a^1)^2 (1 + \tau_a^2)^2 + 4 \tau_a^2}{(1 + \tau_a^2)^2} )</td>
</tr>
<tr>
<td></td>
<td>( (2, 1) = \frac{-4 \sqrt{2} (\tau_a^1)^2 (1 + \tau_a^2)^2 + 4 \tau_a^2}{(1 + \tau_a^2)^2} )</td>
</tr>
<tr>
<td></td>
<td>( (2, 2) = \frac{-4 \sqrt{2} (\tau_a^1)^2 (1 + \tau_a^2)^2 + 4 \tau_a^2}{(1 + \tau_a^2)^2} )</td>
</tr>
<tr>
<td><strong>ARMA</strong></td>
<td>[ \begin{pmatrix} \frac{1}{(1 + a)^2} \ \frac{-1 + a}{(1 + a)^2} \end{pmatrix} ]</td>
</tr>
</tbody>
</table>

Table 5.1: Fisher information matrices at the true parameter values for various different parametrizations considered in this thesis. In case of the ARMA parametrization, the parameters are directly taken to be the coefficients in the numerator and denominator polynomial of \( k(z) = \frac{1 + b z}{1 + a z} \), respectively.
where the columns of $Q_T^\perp$ are chosen to be orthogonal for all $T \in \text{GL}(n)$. Moreover, let $\mathcal{R}(\tau_D^T) > 0$ represent a Riemannian metric tensor for $\mathbb{M}(n)$ where $\tau_D^T \in T_D^{\text{loc},T}$ and $T_D^{\text{loc},T}$ denotes an appropriately chosen open neighborhood of $0 \in T_D^T$; see statement (ii) in theorem (4.9.1). Using the notation $\mathcal{R}_T = \mathcal{R}(\tau_D^T)|_{\tau_D^T = 0}$, the following can hold true for $||T|| \rightarrow \infty$ or $||T|| \rightarrow 0$:

(i) $\lambda_{\text{max}}(\mathcal{R}_T) \rightarrow \infty$ and $\lambda_{\text{min}}(\mathcal{R}_T) \rightarrow M$, $M > 0$

(ii) $\lambda_{\text{max}}(\mathcal{R}_T) \rightarrow \infty$ and $\lambda_{\text{min}}(\mathcal{R}_T) \rightarrow 0$

(iii) $\kappa(\mathcal{R}_T) \rightarrow \infty$

where $\lambda_{\text{max}}(X)$ and $\lambda_{\text{min}}(X)$ denote the eigenvalues of the square matrix $X$ of maximum and minimum modulus, respectively; $\kappa(X)$ denotes the condition number (corresponding to the spectral norm) and is given by the quotient of the largest over the smallest singular value of $X$.

Remark 5.6.3. For the case $n = s = 1$ and $m = 0$ it will be shown below that statements (i) and (iii) actually hold true for all initial systems $(a, k, c) \in S_m(1)$. As far as statement (iii) is concerned, we conjecture that it holds for any choice of $n, m$ and $s$, but we do not have a proof for this.

We will now show each of these statements:

(i) To see this, let us reconsider the example in section (4.9.3) for the case $n = s = 1$, $m = 0$ and $(a, k, c) = (0, 3, 1)$; see also figure (4.3). Note that

$$Q_T = \begin{pmatrix} 0 & t \\
3t & -t \end{pmatrix}$$

and thus we can choose $Q_T^\perp = \begin{pmatrix} 1 & 0 \\
0 & \frac{1}{\sqrt{9t^2 + 1}} \end{pmatrix}$.

Let us now consider the two systems in $Q_{(0,3,1)}^\perp$ corresponding to steps of length $\varepsilon$ in the coordinate directions of the parameter space $T_D = T_D^T$. From statement (ii) in theorem (4.9.1) it is clear that the corresponding two transfer functions must then have unique representatives in the parameter space $T_D^{\text{loc},T}$ for each $T \in \text{GL}(n)$ provided that $\varepsilon > 0$ is chosen small enough. In order to find these two parameter values in $T_D^{\text{loc},T}$, we start by first applying a state transformation to the two systems:

\begin{align}
(0, 3, 1) + \varepsilon(1, 0, 0) & \mapsto (0, 3t, \frac{1}{t}) + \varepsilon(1, 0, 0) \in Q_{(0,3,t^{-1})}^\perp \\
(0, 3, 1) + \varepsilon(0, \frac{1}{\sqrt{10}}, \frac{3}{\sqrt{10}}) & \mapsto (0, 3t, \frac{1}{t}) + \varepsilon(0, \frac{1}{\sqrt{10}}, \frac{3}{\sqrt{10}}) \not\in Q_{(0,3,t^{-1})}^\perp \text{ if } t \neq \pm1
\end{align}

where the inclusions are easily verified by lemma (3.2.3). Note that the first system in $Q_{(0,3,1)}^\perp$ is mapped onto a system in $Q_{(0,3,t^{-1})}^\perp$, i.e. can be represented in the parameter space $T_D^{\text{loc},T}$. Moreover, note that the corresponding parameter vector also has length $\varepsilon$. In other words: A step of length $\varepsilon$ in the first coordinate direction of the parameter space $T_D = T_D^T$ yields the same transfer function as a step of length $\varepsilon$ (in the first coordinate direction) of the parameter space $T_D^T$. 

\begin{align}
(0, 3, 1) + \varepsilon(1, 0, 0) & \mapsto (0, 3t, \frac{1}{t}) + \varepsilon(1, 0, 0) \in Q_{(0,3,t^{-1})}^\perp \\
(0, 3, 1) + \varepsilon(0, \frac{1}{\sqrt{10}}, \frac{3}{\sqrt{10}}) & \mapsto (0, 3t, \frac{1}{t}) + \varepsilon(0, \frac{1}{\sqrt{10}}, \frac{3}{\sqrt{10}}) \not\in Q_{(0,3,t^{-1})}^\perp \text{ if } t \neq \pm1
\end{align}
In order to compute the representative of the second system (5.71) in $Q^\perp_{(0,3t,0)}$, we use lemma
lemma (3.2.3) again:

$$(0,3t,\frac{1}{t}) + (a_s,k_s,c_s) \in Q^\perp_{(0,3t,0)} \iff a_s \cdot 0 + 3tk_s = 0 \cdot a_s + \frac{1}{t}c_s \quad (5.72)$$

Without restriction of generality we consider $\varepsilon = 1$. Note that $(a_s,k_s,c_s)$ has to have the form

$$(a_s,k_s,c_s) = (0,3x,\frac{1}{x}) + (0,\frac{x}{\sqrt{10}},\frac{3}{x\sqrt{10}}) - (0,3t,\frac{1}{t}) = (0,3x-3t+\frac{x}{\sqrt{10}},\frac{1}{x} - \frac{1}{t} + \frac{3}{x\sqrt{10}})$$

for some $x \in \mathbb{R} \setminus \{0\}$. Therefore, for given $t$, we have to find the solutions $x$ to the equation (5.72):

$$3t(3x-3t+\frac{x}{\sqrt{10}}) = \frac{1}{t}(\frac{1}{x} - \frac{1}{t} + \frac{3}{x\sqrt{10}}) \quad (5.73)$$

This equation is quadratic in $x$, and the solutions are given by

$$x_{1,2} = \frac{0.00502579}{t^3} \left( -10 + 90t^4 = \sqrt{100+5954.73t^4+8100t^8} \right)$$

The second solution corresponds to

$$(a_s,k_s,c_s) = \left( 0, -\frac{1}{t} - \frac{3}{2}t^4 + \frac{1}{t^2} \sqrt{100+5954.73t^4+8100t^8}, -\left( \frac{1}{t} + \frac{387.737t^3}{-10+90t^4+\sqrt{100+5954.73t^4+8100t^8}} \right) \right)$$

Note that the term below the square root of $k_s$ can be written as $(90t^2 + \frac{5954.73}{180}t^2) - C$ where $C$ is some constant which is independent of $t$. For $t$ large enough, this square root is approximately $(90t^2 + \frac{5954.73}{180}t)$, such that the terms of degree four vanish in the numerator and we have

$$(a_s,k_s,c_s) \rightarrow (0,0,0) \quad \text{for } t \rightarrow \infty \quad (5.74)$$

Hence, a step of length $\varepsilon = 1$ in the second coordinate direction of the parameter space $T_D = T_D'$ yields the same transfer function as a step of smaller and smaller length (converging to zero) in the parameter space $T_D'$ if $t \rightarrow \infty$.

In total, this implies that any Riemannian metric tensor $\mathcal{R}_T = \mathcal{R}(\tau_D T)|_{x_D = 0}$ must satisfy statement (i) in the observation (5.6.1) above.

(ii) To see this, we have to consider another example where $n = 1$, $m = 0$ and $s = 2$:

$$A = 0.5 \quad K = (1,0) \quad C = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Note that again $T = t \in \mathbb{R} \setminus \{0\}$. We get

$$Q_T = \begin{pmatrix} 0 & t \\ t & 0 \end{pmatrix} \quad \text{and thus we can choose } Q_T^\perp = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{1+t^2}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{\sqrt{1+t^2}} \end{pmatrix}$$

Let us again consider two systems in $Q^\perp_{(0.5,(1,0),(1.0)|T}}$ corresponding to steps of length $\varepsilon$ in two
coordinate directions of the parameter space $T_D = T_D'$. Applying a state transformation to these
systems, we get
\[(0.5, (1, 0), (1, 0)^t) + \varepsilon(0, (0, 0), (0, 1)^t) \Rightarrow (0.5, (t, 0), \left(\frac{1}{t}, 0\right)^t) + \frac{1}{t}\varepsilon(0, (0, 0), (0, 1)^t) \in Q^{+}_{(0.5, t(1, 0), (1, 0)^t)} \]

\[(0.5, (1, 0), (1, 0)^t) + \varepsilon(0, (0, 1), (0, 0)^t) \Rightarrow (0.5, (t, 0), \left(\frac{1}{t}, 0\right)^t) + t \cdot \varepsilon(0, (0, 1), (0, 0)^t) \in Q^{+}_{(0.5, t(1, 0), (1, 0)^t)} \]

In the first case, a step of length \(\varepsilon\) in the third coordinate direction of the parameter space \(T_D = T_D^0\) yields the same transfer function as a step of smaller and smaller length \(\frac{1}{t} \cdot \varepsilon\) (in the third coordinate direction) of the parameter space \(T_D^0\) if \(t \to \infty\).

In the second case, a step of length \(\varepsilon\) in the second coordinate direction of the parameter space \(T_D = T_D^0\) yields the same transfer function as a step of larger and larger length \(t \cdot \varepsilon\) (in the second coordinate direction) of the parameter space \(T_D^0\) if \(t \to \infty\).

Both statements imply that any Riemannian metric tensor \(\mathcal{R} = \mathcal{R}(T_D^0)\) must satisfy statement (ii) in the observation (5.6.1) above.

(iii) The fact that statement (iii) holds for at least one minimal \((a, k, c) \in S_m(1)\) is an immediate consequence of (i) where the initial system for the DDLC parametrization was taken to be \((a, k, c) = (0, 3, 1)\). As far as the statement in remark (5.6.3) is concerned we note that it is only a matter of tedious calculations to see that (i) (and thus (iii)) holds true if any minimal \((a, k, c)\) is chosen as initial system for the DDLC parametrization.

Of course, analogous statements can be shown for \(t \to 0\).

**Remark 5.6.4.** In a certain sense, property (ii) in observation (5.6.1) is "worse" than (i): Assume we consider a neighborhood \(T_D^{loc, \varepsilon}\) of \(T_D^0\) of fixed size. For instance, we could consider balls of fixed radius \(\varepsilon\) with center 0 in the parameter spaces \(T_D^0\). It is then clear that the distance (measured in some Riemannian metric) from the transfer functions \(\pi(0), 0 \in T_D^0\) to the boundary of the corresponding open neighborhoods \(V^{loc, \varepsilon}_{D}\) in \(\mathbb{M}(n)\) can decrease to zero if \(\|T\| \to 0\) or \(\|T\| \to \infty\). This means that if one considers a step of bounded length in each of the parameter spaces \(T_D^0\), then the corresponding transfer functions become closer and closer in \(\mathbb{M}(n)\). In fact, it seems hard to determine beforehand whether an estimation algorithm can be prevented from entering the regions in \(S(n)\) where these problems might occur. Note that problems of this type can be avoided by the use of orthoDDLC; see remark (5.6.6) below.

The following lemma is the analogue to observation (5.6.1) for orthoDDLC:

**Lemma 5.6.2 (Choice of initial systems in \(E^{BSA}(\bar{A}, \bar{B}, \bar{C}, D)\) for orthoDDLC).** Let \(\{\varphi_{D,Q}^D, Q \in O(n)\}\) be the following family of \(s\)ldsDDLC parametrizations

\[\varphi_{D,Q}^D : T_D^{loc, \varepsilon} \to S_m(n)\]

\[\tau_{D,Q}^D \Rightarrow \Delta^V \left( \begin{array}{c} \text{vec}(Q \bar{A}Q') \\ \text{vec}(Q'B) \\ \text{vec}(CQ') \\ \text{vec}(D) \end{array} \right) + \left( \begin{array}{cccc} (Q \bar{A}Q') & 0_{n \times n^2} & 0_{n^2 \times n} & 0_{n^2 \times n^2} \\ 0_{n \times n^2} & I_n & 0_{n \times n} & 0_{n \times n} \\ 0_{n^2 \times n} & 0_{n \times n} & I_n & 0_{n \times n} \\ 0_{n^2 \times n} & 0_{n \times n} & 0_{n \times n} & I_n \end{array} \right) \tau_{D,Q}^D \]

where the columns of \(Q_{\varphi_{D,Q}^D}\) are chosen to be orthogonal for all \(Q \in O(n)\) and \((\bar{A}, \bar{B}, \bar{C}, D) \in BSA(n)\). Let \(T_D^{loc, \varepsilon}\) denote an appropriately chosen open neighborhood of \(\tau_{D,Q}^D = 0 \in T_D^{loc, \varepsilon}\); see statement (ii) in theorem (4.10.1) and let the matrix \(\mathcal{R} = \mathcal{R}_I > 0\) represent any symmetric and positive definite bilinear form in \(T_D^{loc, \varepsilon}\). Interpret (5.75) as a coordinate transformation from \(\tau_{D,I}^D \in T_D^{loc, \varepsilon}\) to \(\tau_{D,Q}^D \in T_D^{loc, \varepsilon}\) and let the matrix \(\mathcal{R}_Q\) represent the symmetric and positive definite bilinear form in \(T_D^{loc, \varepsilon}\). Then, \(\sigma(\mathcal{R}_Q) = \sigma(\mathcal{R})\) where \(\sigma(X)\) denotes the set of eigenvalues of the (square) matrix \(X\).

**Remark 5.6.5.** Note that we can, e.g., choose \(\mathcal{R}_Q\) to be the Fisher information matrix, i.e. the Hessian of the asymptotic likelihood function, evaluated at the true parameters \(\tau_{D,Q}^D = 0\); see remark (5.2.3) above.
Proof. From statement (ii) and (iii) of lemma (3.4.7) we see that a step of length \( \varepsilon \) in each coordinate direction of the parameter space \( T^{o,loc}_{D,I} \) is uniquely represented by a step of the same length in the parameter space \( T^{o}_{D,Q} \). Moreover, it follows immediately (in the same manner as statement (iii) of lemma (3.4.7) was proved) that any two steps in orthogonal directions of \( T^{o,loc}_{D,I} \) are uniquely represented by steps of the same length in orthogonal directions of \( T^{o}_{D,Q} \); note that the columns of \( Q_{\perp}^{a} \) are assumed to be orthogonal for all \( Q \in O(n) \) such that orthogonal directions in the parameter spaces yield orthogonal directions in \( \mathbb{R}^{n+2ns+2^s} \). This implies that the coordinate transformation from \( \tau_{D,I}^{o} \in T^{o}_{D,I} \) to \( \tau_{D,Q}^{a} \in T^{a}_{D,Q} \) must preserve the Frobenius norm and orthogonality. It is easy to see that the coordinate transformation is in fact linear (not even affine) and thus it must be represented by an orthogonal \( ns \times ns \) matrix, i.e., \( \tau_{D,Q}^{a} = G \cdot \tau_{D,I}^{o} \), \( G \in O(ns) \). This implies that \( R_{Q} = G R G' \), and such a transformation clearly leaves the eigenvalues unchanged.

It is clear from this lemma that the choice of the initial representative within the equivalence class \( E^{BSA}(\tilde{A}, B, \tilde{C}, D) \) is irrelevant for the condition number of the Fisher information matrix. Clearly, this is in contrast to the situation for DDLC; see statement (iii) in observation (5.6.1) above.

**Remark 5.6.6.** Assume we consider a neighborhood \( T^{o,loc}_{D,Q} \) of \( 0 \in T^{a}_{D,Q} \) of fixed size. For instance, we could consider balls of fixed radius \( \varepsilon \) with center 0 in the parameter spaces \( T^{a}_{D,Q} \). It then follows from (i), (ii) and (iii) in lemma (3.4.7) and lemma (3.6.2) that the corresponding open neighborhoods \( V^{o,loc}_{D,Q} \in \mathbb{M}(n) \) (!) are the same for all initial representatives in the equivalence class \( E^{BSA}(\tilde{A}, B, \tilde{C}, D) \). In particular, this means that the distance (measured in some metric) from the transfer functions \( \pi(0) \in T^{a}_{D,Q} \) to the boundary of the corresponding open neighborhoods \( V^{o,loc}_{D,Q} \) in \( \mathbb{M}(n) \) cannot decrease to zero, which is clearly in contrast to the situation for DDLC; see remark (5.6.4) above.

To sum it up, **orthoDDLC** has a number of favourable properties as compared to DDLC. Apart from the fact that less parameters are explicitly estimated, one need not bother about the choice of initial state-space realizations in the equivalence class. Moreover, the implementation of **orthoDDLC** is very simple: Step (iv) in theorem (4.11.1) is virtually costless, and step (v) consists of a singular value decomposition of a matrix of size \((n+s)\times(n+s)\) which trivially does not grow with the amount of data.

### 5.7 Future research

Of course, a list of possible future research topics is not too hard to obtain, in particular concerning **s1sDDLC** and **orthoDDLC**. The following points are just a few questions that turn up in a rather immediate way:

1. Of course, the (generalized) least squares step to determine \( \tilde{D}, \tilde{B} \) and \( \tilde{K} \) from \( \tilde{A} \) and \( \tilde{C} \) — see (2.83) — may be ill-conditioned for both **s1sDDLC** and **orthoDDLC**. Such problems will in particular occur if the pair \( (\tilde{C}, \tilde{A}) \) is close to being unobservable. Note however, that for **orthoDDLC**, this source of problems should not be too severe if the step lengths are not extremely large because the algorithm keeps the initial pairs \( (\tilde{C}, \tilde{A}) \) output normal. Of course, one can monitor the condition number of \( (X'X) \) in (2.83) during the search procedure.

2. The question whether for any given minimal initial system \((A, B, C, D, K)\), each transfer function within (one connected component of) \( \mathbb{M}(n) \) can be represented in the DDLC parameter space after a finite number of iterations of the DDLC construction starting at \((A, B, C, D, K)\) is still subject to research; see also remark (4.9.7). In the light of the remark (5.6.4), however, the answer here might also be negative.

3. An analogous question to the one in (ii) could be posed for the **orthoDDLC** approach: Is there any guarantee that after a finite number of iterations of the **orthoDDLC** method one can represent every transfer function within (one connected component of) \( \mathbb{M}(n) \) (by allowing the data to vary)?

4. The question whether the coordinate free consistency result in section (2.3.4) can be carried over to yield consistency in terms of the parameter estimates is still open; see remark (4.10.7).
5.8 Figures

Here, we present all figures which have been discussed in this chapter:

Condition number of FIMs at true parameter values

Figure 5.1: (a) The case $n = s = 1$ and $m = 0$ where $S(1) = \mathbb{R}^3$: Condition number of the Fisher information matrix for different parametrizations: Note that $c = 1$ is fixed, such that each point in the $(a,k)$-plane of the plot corresponds to a different transfer function and a unique representative in terms of the canonical forms. In case of DDLC, Ober’s Lyapunov balanced representative was chosen as the initial system. In case of the ARMA parametrization, the parameters are directly taken to be the coefficients in the numerator and denominator polynomial of $k(z) = \frac{1}{1 + \alpha z^{-1}}$, respectively. Note that the condition number corresponding to the slsDDLC parametrization (and therefore, also for orthoDDLC) is trivially equal to one because there is only a single parameter to be estimated in this case.
Figure 5.2: (a) Levenberg-Marquardt algorithm for the finite sample concentrated likelihood function $L_n^c(Y^T_t;U^T_t,\tau)$ in (2.77) and the echelon canonical form. The data set was constructed by simulating an output sequence of length $T = 200$ for a univariate exogenous input given by $u = \text{rand}(200, \text{SINE}, [10, 30])$ and a univariate white noise input of variance $\sigma_0^2 = 1$, where the true model is given by $(a_0, b_0, c_0, d_0, k_0) = (0.5, 0.5, 1, 0, 1)$. The initial estimate is chosen as $(a, b, c, d, k) = (0.9, 1, 1, 0, 0.5)$. The estimation is performed by the built-in function "pem"; see section (5.5) for a brief description of the Levenberg-Marquardt method.

(b) Gauss-Newton search (combined with a simple line search) for the finite sample concentrated likelihood function $L_n^c(Y^T_t;U^T_t,\tau)$ in (2.77) and the DDLC parametrization. The data set, the true model and the initial estimates are as in (a). Note that the gradient and the Gauss-Newton approximation to the Hessian are computed analytically here; see sections (5.3.2) and (5.3.3).

(c) Gauss-Newton search (combined with a simple line search) for the asymptotic likelihood function and McGinnie's minimum phase balanced parametrization. Note that $m = 0$, i.e. there are no exogenous inputs. The initial estimate is taken to be $(a, k, c) = (0.9, 0.5, 1)$, the true transfer function is given by $\pi(a_0, k_0, a_0) = \pi(0.9, 1, 1)$ and $\sigma_0^2 = 1$. Note that the gradient and the Gauss-Newton approximation to the Hessian are computed analytically here; see also remark (5.2.4).

(d) Gradient search (combined with a simple line search) for the asymptotic likelihood function and the full state-space parametrization. Note that $m = 0$, i.e. there are no exogenous inputs. The initial estimate, the true transfer function and the true innovations covariance are as in (c). Note that the gradient is computed analytically here; see also remark (5.2.2).
Figure 5.3: The case $n = s = 1$ and $m = 0$ where $S(1) = \mathbb{R}^3$:

(a) Condition number of the Fisher information matrix for DDLC (thick blue curve) as a function of the chosen initial representative $(0.1, t \cdot 1, 1 \cdot t^{-1}), t \in \mathbb{R}^+$. The condition numbers for the echelon form (red), Ober’s Lyapunov balanced form (yellow), McGinnie’s minimum phase balanced form (white) and a simple ARMA parametrization (pink) are given in the legend below the plot and are represented by thin straight coloured lines. Note that the condition number corresponding to the slsDDLC parametrization (and therefore, also for orthoDDLC) is trivially equal to one because there is only a single parameter to be estimated in this case. In case of the ARMA parametrization, the parameters are directly taken to be the coefficients in the numerator and denominator polynomial of $k(z) = \frac{1+\theta z}{1+\xi z}$, respectively. The (partially overlapping) points on the thick blue curve represent the system representatives in the canonical forms considered here.

(b) Condition number of the Fisher information matrix for DDLC (thick blue curve) as a function of the chosen initial representative $(0.9, t \cdot 0.4, 1 \cdot t^{-1}), t \in \mathbb{R}^+$. The condition numbers for the canonical form are given in the legend below the plot and are represented by thin straight coloured lines if the scale of the plot permits this. See (a) for a further description.

(c) Eigenvalues of the Fisher information matrix for DDLC (thick blue curve) as a function of the chosen initial representative $(0.1, t \cdot 1, 1 \cdot t^{-1}), t \in \mathbb{R}^+$. The (partially overlapping) points on the thick blue curve represent the system representatives in the canonical forms considered here.

(d) Eigenvalues of the Fisher information matrix for DDLC (thick blue curve) as a function of the chosen initial representative $(0.9, t \cdot 0.4, 1 \cdot t^{-1}), t \in \mathbb{R}^+$. See (c) for a further description.
Figure 5.4: The case $n = s = 1$ and $m = 0$ where $S(1) = \mathbb{R}^3$:

(a) Gradient search directions for the asymptotic likelihood function for different parametrizations: Note that the true transfer function is represented by the thin hyperbola through the point $(a_0, k_0, c_0) = (0.5, 1, 1)$. The current estimate is represented by the thick hyperbola through the point $(a, k, c) = (0.9, 0.5, 1.5)$. Clearly, the state-space matrices depend on the parameters in a non-linear way in case of Ober’s Lyapunov balanced and McGinnie’s minimum phase balanced form, explaining the curves corresponding to a gradient step for these parametrizations. For the Orthimp parametrization, see remark (5.6.1) above.

(b) Gauss-Newton search directions for the asymptotic likelihood function for different parametrizations: See (a) for a further description.

(c) Projection of the gradient search directions in (a) onto $Q^1_{(0,0.5,1.5)}$, i.e. onto the ortho-complement of the tangent space to the hyperbola $\mathcal{E}(a, k, c)$ at the initial system $(a, k, c) = (0.9, 0.5, 1.5)$. Note the difference between the parametrizations, illustrating the fact that the directions of a (usual) gradient step on the manifold $M(1)$ are different for different parametrizations.

(d) Projection of the Gauss-Newton search directions in (b) onto $Q^1_{(0,0.5,1.5)}$, i.e. onto the ortho-complement of the tangent space to the hyperbola $\mathcal{E}(a, k, c)$ at the initial system $(a, k, c) = (0.9, 0.5, 1.5)$. The figure illustrates the fact that the directions of a Gauss-Newton step on the manifold $M(1)$ are the same for different parametrizations. This is because Gauss-Newton algorithms can also be interpreted as Riemannian gradient algorithms for a particular (data-dependent) Riemannian metric.
Appendix A

Appendix

A.1 Continuous time state-space systems

Consider the following linear, time invariant continuous time state space system:

\[
\begin{align*}
\dot{x}(t) & = A_c x(t) + B_c u(t), \quad x(t_0) = x_0 \quad (A.1) \\
y(t) & = C_c x(t) + D_c u(t) \quad (A.2)
\end{align*}
\]

Here, \(x(t)\) is the \(n\)-dimensional state vector which is not directly observed in general, and \(A_c \in \mathbb{R}^{n \times n}\), \(B_c \in \mathbb{R}^{n \times m}\), \(C_c \in \mathbb{R}^{p \times n}\) and \(D_c \in \mathbb{R}^{p \times m}\) are parameter matrices; \(y(t)\) and \(u(t)\) are the observed outputs and inputs, respectively.

Consider the matrix differential equation in (A.1). Applying the existence and uniqueness result of Picard-Lindelöf for initial value problems and making use of the constructive method of successive approximation, one gets the unique solution to the homogeneous part as the matrix exponential \(x(t) = e^{A_c(t-t_0)}x_0\). A particular solution is obtained by variation of the constant and is given by \(e^{A_c(t-t_0)} \int_{t_0}^{t} e^{-A_c(t-	au)} B_c u(\tau) d\tau = \int_{t_0}^{t} e^{A_c(t-	au)} B_c u(\tau) d\tau\). Therefore, the unique solution to (A.1) is given by

\[
\begin{align*}
x(t) & = e^{A_c(t-t_0)}x_0 + \int_{t_0}^{t} e^{A_c(t-	au)} B_c u(\tau) d\tau \quad (A.3) \\
y(t) & = C_c x(t) + D_c u(t) \quad (A.4)
\end{align*}
\]

Note that under the stability assumption

\[
\text{Re}(\lambda_i(A_c)) < 0 \quad (A.5)
\]

where \(\lambda_i(A_c), i = 1, \ldots, n\) denote the eigenvalues of \(A_c\), we can assume that the system was started at some time instant \(t_0 \to -\infty\) with some bounded, but arbitrary state vector \(x_0 \to x_{-\infty}\) to get the so called steady state solution:
\[
x(t) = \lim_{t_0 \to -\infty} e^{A_c(t-t_0)} x_0 + \int_{t_0}^{t} e^{A_c(t-\tau)} B_c u(\tau) d\tau
\]

\[
= \int_{-\infty}^{t} e^{A_c(t-\tau)} B_c u(\tau) d\tau
\]

\[
= \int_{0}^{\infty} e^{A_c \tau} B_c u(t-\tau) d\tau
\]

(A.6)

\[
y(t) = C_c \int_{0}^{\infty} e^{A_c \tau} B_c u(t-\tau) d\tau + D_c u(t)
\]

(A.7)

The function

\[
\hat{l}_c(t) = C_c e^{A_c t} B_c 1_{[0,\infty)}(t) + D_c \delta(t)
\]

is called the impulse response of the linear system; \(\delta(t)\) denotes the unit impulse and \(1_{[0,\infty)}(t)\) is the unit step function. Under the stability assumption and by setting \(\hat{l}_c(t) = 0\) for \(t \leq 0\), this function is an element of the Hilbert space \(L_2^{\infty}m(\mathbb{R})\), the space of real, matrix valued and square integrable Lebesgue measurable functions on \(\mathbb{R}\).

A wide spread alternative method of solving system (A.1) is via the Laplace transform. The impulse response \(\hat{l}_c(t)\) then becomes the transfer function \(l_c(s)\) of the system:

\[
l_c(s) = \int_{0}^{\infty} \hat{l}_c(t)e^{-st}dt
\]

\[
= C_c \int_{0}^{\infty} e^{(A_c - sI)t} B_c dt + D_c
\]

\[
= C_c(sI - A_c)^{-1} B_c + D_c
\]

where the last expression is a convenient shorthand notation. Note that \(l_c(s)\) will always be a proper rational matrix.

**Definition A.1.1 (Properness).** A rational \(s \times m\) continuous time transfer function \(l_c(s)\) is called proper if \(\lim_{|s| \to \infty} l_c(s)\) is bounded.

We now give a list of definitions of various properties of continuous time rational transfer functions. Note that the Smith McMillan form of a rational matrix is given in theorem (1.2.1):

**Definition A.1.2 (Stability).** A rational and proper \(s \times m\) continuous time transfer function \(l_c(s)\) is said to be stable if all the poles of \(l_c(s)\) are in the open left half plane of the complex plane. Here, the poles are defined as the zeros of the denominator polynomials in the Smith McMillan form of \(l_c(s)\).

**Definition A.1.3 (Positive Realness).** A rational and proper \(s \times s\) continuous time transfer function \(k_c(s)\) is said to be positive real if \(k_c(\infty) + k_c(-\infty)\) \(\geq 0\) and \(k_c(i\omega) + k_c(-i\omega)\) \(\geq 0\) for all \(\omega \in \mathbb{R}\). It is said to be strictly positive real if both inequalities are strict.

**Remark A.1.1.** If \((A_c, B_c, C_c, D_c)\) is a state-space realization of a strictly positive real \(k_c(s)\), then clearly \((D_c + D_c') > 0\).

**Definition A.1.4 (Minimum Phase Property).** A rational and proper \(s \times s\) continuous time transfer function \(k_c(s)\) is said to be minimum phase if all the zeros of \(k_c(s)\) are in the closed left half plane of the complex plane. It is said to be strictly minimum phase if \(k_c^{-1}(s)\) is stable. Here, the zeros are defined as the zeros of the (nonzero) numerator polynomials in the Smith McMillan form of \(k_c(s)\).
Remark A.1.2. If \((A_c, K_c, C_c, E_c)\) is a state-space realization of a strictly minimum phase \(k_c(s)\), then clearly \(E_c\) is invertible.

The terms McMillan degree, degree and order are used interchangeably:

**Definition A.1.5 (McMillan degree of a continuous time transfer function).** A rational and proper \(s \times m\) continuous time transfer function \(l_c(s)\) is said to be of (McMillan) degree or order \(n\) if the sum of the denominator degrees in the Smith McMillan form of \(l_c(s)\) is equal to \(n\).

## A.2 Spectral summands and spectral factors

Let us start with the following definition:

**Definition A.2.1 (Positive realness).** A rational and causal \(s \times s\) discrete time transfer function \(\Phi(z)\) is said to be **positive real** if \(\Phi(e^{i\omega}) + \Phi(e^{-i\omega})' \geq 0\) for all \(\omega \in [0, 2\pi]\). It is said to be **strictly positive real** if the inequality is strict.

Let \(\Phi(z)\) be (strictly) positive real. Put \(f(z) = \Phi(z) + \Phi^*(z)\), where, as usual, \(\Phi^*(z) = \Phi(\frac{1}{z})'\). Then \(\Phi(z)\) is called a **spectral summand** for \(f(z)\).

Let \(f(z) \in L^s_{\times s}([0, 2\pi])\) be a (real) rational \(s \times s\) spectral density, i.e. \(f(e^{i\omega}) = f(e^{-i\omega})'\) and \(f(e^{i\omega}) \geq 0\) \(\omega - a.e\). Such a spectral density may always be factorized as \(f(z) = k(z)k^*(z)\) or, alternatively, as \(f(z) = k_1(z)k_1(z)\); \(k(z)\) is called a right spectral factor of \(f(z)\) and \(k_1(z)\) is called a left spectral factor of \(f(z)\).

We will now explore the relationship between spectral factors and summands:

For this purpose, let us first introduce the set \(\mathcal{M}_n([a, b], n)\) of all rational, causal, stable and (strictly) positive real \(s \times s\) transfer functions \(\Phi(z)\) of order \(n\), where \(\Phi(0) = D\) (and \((D + D^t) \geq (>)0\) by definition).

We start with some (linearly regular) stationary process of the form \(y_t = k(z)\varepsilon_t\) where \((\varepsilon_t)\) is a white noise process with \(E\varepsilon_t\varepsilon_t^t = I\) and \(k(z)\) is rational, causal, stable and (strictly) minimum phase and is given by

\[
k(z) = \begin{bmatrix} A & K \\
C & E \end{bmatrix}
\]

Note that \(k(0)\) need not be the identity matrix in this case. Similar to \((1.12)\) and \((1.13)\), we get

\[
\gamma(0) = CPC + EE' \\
\gamma(j) = CA^{j-1}(APC + KE') \quad \text{for} \; j > 0
\]

where \(P\) is the unique solution to the discrete time Lyapunov equation \(P = APA' + KK'\). As is well known, the spectral density is obtained by applying the Fourier transform to the covariance sequence \((\gamma(j))_{j \in \mathbb{Z}}\) (and extending it from the unit circle \(z = e^{i\omega}, \omega \in [0, 2\pi]\) to the whole complex plane):

\[
2\pi f_y(z) = k(z)k^*(z) = \sum_{j=-\infty}^{\infty} \gamma(j)z^j
\]

As \(\gamma(j) = \gamma(-j)'\), we can put

\[
k(z)k^*(z) = 2\pi f_y(z) = \sum_{j=1}^{\infty} \gamma(j)z^j + \frac{1}{2}\gamma(0) + \frac{1}{2}\gamma(0) + \sum_{j=-\infty}^{-1} \gamma(j)z^j
\]

\[
(A.10)
\]
and the spectral summand $\Phi(z)$ is clearly positive real due to the properties of the spectrum. We will now treat the question of how to obtain a state-space realization for the spectral summand $\Phi(z)$, given a right spectral factor $k(z)$. First, note that the impulse response of $\Phi(z)$ must be equal to the covariance sequence $(\gamma(j))_{j \in \mathbb{N}}$, and from (A.9) we see that the $A$ and $C$ matrices in the state-space representation of $\Phi(z)$ can be chosen to be the corresponding matrices in the state-space representation of $k(z)$. Hence,

$$\Phi(z) = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$  \hspace{1cm} (A.11)

and $B$ and $D$ have to be chosen such that

$$P = APA' + KK'$$  \hspace{1cm} (A.12)\]

$$B = APC' + KE'$$  \hspace{1cm} (A.13)

$$D + D' = CPC' + EE'$$  \hspace{1cm} (A.14)

In particular, $D = \gamma(0)/2$. Note that the Hankel matrices corresponding to $k(z)$ and $\Phi(z)$ are then given by

$$\mathcal{H}_k = \begin{bmatrix} CK & CAK & CA^2K & \cdots \\ CA & CA^2 & CA^3 & \cdots \\ : & : & : & \ddots \end{bmatrix}$$  \hspace{1cm} (A.15)

$$\mathcal{H}_\Phi = \begin{bmatrix} CB & CAB & CA^2B & \cdots \\ CAB & CA^2 & CA^3 & \cdots \\ : & : & : & \ddots \end{bmatrix} = \begin{bmatrix} \gamma(1) & \gamma(2) & \gamma(3) & \cdots \\ \gamma(2) & \gamma(3) & \gamma(4) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$  \hspace{1cm} (A.16)

The converse direction of computing a right and left spectral factor, given a spectral summand $\Phi(z)$ of the form (A.11), will be treated in theorem (A.2.2) below. For this purpose, let us insert $K$ from equation (A.13) into the first Lyapunov equation (A.12). This yields, in case of invertibility of $EE'$ (i.e. strict positive reality; see next theorem), $P = APA' + (B - APC')(EE')^{-1}(B - APC')'$ and by substituting for $(EE')^{-1}$ using the last equation (A.14) we get

$$P = APA' + (B - APC')(D + D' - CPC')^{-1}(B - APC')'$$  \hspace{1cm} (A.17)

which is known as the Positive Real Riccati Equation (PR-ARE).

Remark A.2.1. Given $\Phi(z)$ in (A.11), it is interesting to note that the set of solutions $\mathcal{P}$ to (A.17) is (if nonvoid) a closed, bounded and convex set of positive definite matrices with two extreme points $P^\$ and $P^+$: for all $P \in \mathcal{P}$: $P^\$ $P$ $P^+$. On page 16 of (Fauvre, 1976), the following property is also stated:

$$P^+ > P^\$ \iff \det(2\pi f_p(e^{-i\theta}))) \neq 0 \quad \forall \theta \in [0, 2\pi]$$  \hspace{1cm} i.e. strict positive reality

Let us now state a famous

Theorem A.2.1 (The Positive Real Lemma, Kalman-Popov-Yacubovich). Let

$$\Phi(z) = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathcal{M}_{n, PR}(n)$$

be a minimal realization of $\Phi(z)$. Then
\( \Phi(z) \) is positive real if and only if there exist real matrices \( P \in \mathbb{R}^{n \times n} \), \( K \in \mathbb{R}^{n \times s} \) and \( E \in \mathbb{R}^{s \times s} \) with \( P = P' \) and \( P > 0 \) such that \((A, K)\) is controllable and the three equations

\[
\begin{align*}
P &= APA' + KK' \\
B &= APC' + KE' \\
D + D' &= CPC' + EE'
\end{align*}
\]

are satisfied. If \( \Phi(z) \) is strictly positive real, then \( EE' > 0 \) and \((A - KE^{-1}C)\) is stable.

\( \Phi(z) \) is positive real if and only if there exist real matrices \( Q \in \mathbb{R}^{n \times n} \), \( L \in \mathbb{R}^{s \times n} \) and \( E \in \mathbb{R}^{s \times s} \) with \( Q = Q' \) and \( Q > 0 \) such that \((L, A)\) is observable and the three equations

\[
\begin{align*}
Q &= A'QA + LL' \\
C' &= A'QB + LE' \\
D + D' &= B'QB + EE'
\end{align*}
\]

are satisfied. If \( \Phi(z) \) is strictly positive real, then \( E'E > 0 \) and \((A - BE^{-1}L)\) is stable.

Proof. See theorem 2 and its proof in (Fauvre, 1976). Original references can be found there, too. \( \square \)

Remark A.2.2. Given \( \Phi(z) \) in (A.11), the second part of this theorem can be used to construct a left spectral factor (hence, the notation LSF). It is the dual version of the first part (RSF) which can be used to construct a right spectral factor; see theorem (A.2.2) below.

In general, for given \((A, B, C, D)\) there exists more than one triple \((P, K, E)\) and more than one triple \((Q, L, E)\) satisfying these equations. In case of the right spectral factor, for instance, one \( P \) can correspond to \((K, E)\) and \((KU, EU)\) where \( U \) is an arbitrary orthogonal matrix: The minimal degree spectral factor is only unique up to a right orthogonal multiple. By the remark above, also \( P \) itself is not unique. In fact, there is a one-to-one correspondence between the equivalence classes of minimal degree spectral factors and the set of matrices \( P \) satisfying the three equations (A.12-A.14) along with an equivalence class of \((K, E)\) of respective dimension. However, the smallest \( P^− > 0 \) in the set \( P \) with its associated equivalence class \((K, E)\) corresponds to the minimum phase spectral factor. By taking the uniquely defined positive lower triangular Cholesky factor of \( EE' \) to be \( E \), we can formulate the following theorem:

**Theorem A.2.2 (Unique right and left spectral factors in discrete time).** Let \( \Phi(z) \in \mathbb{M}_{n,apr}(n) \) be a strictly positive real and stable transfer function as given above and let

\( \text{RSF} \quad P = P', \ P > 0 \) be the minimal solution to the Positive Real Riccati Equation

\[
P = APA' + (B - APC')(D + D' - CPC')^{-1}(B - APC')'
\]

(A.18)

If the matrix square root below is taken to be the positive lower triangular Cholesky factor, a unique strictly minimum phase right spectral factor \( k(z) \) is obtained by

\[
k(z) = \begin{bmatrix} A & K \\ C & E \end{bmatrix}
\]

\[
K = (B - APC')E^{-'} \quad \text{and} \quad E = (D + D' - CPC')^{\frac{1}{2}}
\]

(A.19)

(A.20)

\( \text{LSF} \quad Q = Q', \ Q > 0 \) be the minimal solution to the Positive Real Riccati Equation

\[
Q = A'QA + (C' - A'QB)(D + D' - B'QB)^{-1}(C' - A'QB)'
\]

(A.21)

If the matrix square root below is taken to be the positive lower triangular Cholesky factor, a unique strictly minimum phase left spectral factor \( k(z) \) is obtained by
\[ k(z) = \begin{bmatrix} A \\ L \\ E \end{bmatrix} \]
\[ L = [(C' - A'QB)E^{-1}]' \]
\[ E = (D + D' - B'QB)^\frac{1}{2} \]

**Proof.** The result that the minimal solutions \( P^- > 0 \) and \( Q^- \) correspond to the minimum phase spectral factors can be found in (Fauvre, 1976) together with an original reference for the proof. The rest follows from what was said above. \( \square \)

An analogous theorem holds true for the continuous time case:

**Theorem A.2.3** (Unique right and left spectral factors in continuous time). Let \( \Phi_c(s) \in M_{n, spr}(n) \) be a strictly positive real and stable continuous time transfer function and let \( P = P', P > 0 \) be the minimal solution to the Positive Real Riccati Equation

\[ AP + PA' + (B - PC')(D + D')^{-1}(B - PC')' = 0 \quad (A.22) \]

If the matrix square root below is taken to be the positive lower triangular Cholesky factor, a unique strictly minimum phase right spectral factor \( k_c(s) \) is obtained by

\[ k_c(s) = \begin{bmatrix} A & K \\ C & E \end{bmatrix} \]
\[ K = (B - PC')E^{-'} \]
\[ E = (D + D')^\frac{1}{2} \]

**LSF** \( Q = Q', Q > 0 \) be the minimal solution to the Positive Real Riccati Equation

\[ A'Q + QA' + (C' - QB)(D + D')^{-1}(C' - QB)' = 0 \quad (A.23) \]

If the matrix square root below is taken to be the positive lower triangular Cholesky factor, a unique strictly minimum phase left spectral factor \( k_c(s) \) is obtained by

\[ k_c(s) = \begin{bmatrix} A & B \\ L & E \end{bmatrix} \]
\[ L = [(C' - QB)E^{-1}]' \]
\[ E = (D + D')^\frac{1}{2} \]

**Proof.** This theorem is based on the Kalman-Popov-Yacubovich lemma for the continuous time case and is also treated in (Fauvre, 1976) where original references are given. \( \square \)

### A.3 Balancing schemes

We will here give a brief overview of a number of different balancing schemes for different classes of transfer functions. This part will exclusively deal with continuous time transfer functions as they serve as a starting point for the derivation of the discrete time canonical forms in sections (4.6), (4.7) and (4.8) of chapter 4. The subscripts \( c \) to denote continuous time system matrices are omitted in this section.
A.3. BALANCING SCHEMES

In the sequel we will only consider square continuous time transfer functions \( k_c(s) \); see the appendix (A.1). Let \( \mathcal{M}_s^R(n) \) be the set of rational, proper and stable \( s \times s \) transfer functions of order \( n \) where \( k_c(\infty) = D \) for an arbitrary \( D \in \mathbb{R}^{s \times s} \). By \( \mathcal{M}_s^R^{-+}(n) \) we denote the subset of \( \mathcal{M}_s^R(n) \) where additionally \((D + D') > 0\) holds. The set \( \mathcal{M}_s^R_{\text{npr}}(n) \) contains all rational, proper, stable and (strictly) positive real \( s \times s \) transfer functions of order \( n \), where again \( k_c(\infty) = D \) (and \((D + D') \geq (>)0\) by definition). Finally, \( \mathcal{M}_s^R_{\text{mp}}(n) \) is the set of rational, proper, stable and (strictly) minimum phase \( s \times s \) transfer functions of order \( n \) where \( k_c(\infty) = E \) for an arbitrary \( E \in \mathbb{R}^{s \times s} \). For an overview of these classes of transfer functions, see also the appendix (A.7).

The balancing schemes for these classes of continuous time transfer functions are summarized in the table below:

<table>
<thead>
<tr>
<th>( \mathcal{M}_s^R(n) )</th>
<th>Lyapunov balancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{M}<em>s^R</em>{\text{npr}}(n) )</td>
<td>Positive real balancing</td>
</tr>
<tr>
<td>( \mathcal{M}<em>s^R</em>{\text{mp}}(n) )</td>
<td>Stochastic balancing</td>
</tr>
<tr>
<td>( \mathcal{M}<em>s^R</em>{\text{mp}}(n) )</td>
<td>Minimum phase balancing</td>
</tr>
</tbody>
</table>

We want to remark that the unqualified term balanced in the sections below will always be used in the particular sense indicated by the title of the corresponding section.

A.3.1 Lyapunov balancing

The concept of Lyapunov balanced state-space realizations is known for a long time and has first been introduced for continuous time stable systems; see (Moore, 1981). Given a minimal realization of a \( k_c(s) \in \mathcal{M}_s^R(n) \) the calculation of the controllability Gramian \( P \) and the observability Gramian \( Q \) can be done via the Lyapunov equations (LEs)

\[
AP + PA' + BB' = 0 \quad (A.24)
\]

\[
A'Q + QA + C'C = 0 \quad (A.25)
\]

Note that for given \( P, Q \) and \( B \) (respectively \( C \)), the equations are linear in \( A \). This will be used for constructing the canonical form in section (4.6). Moreover, \( P > 0 \) if and only if \((A,B)\) is controllable and \( Q > 0 \) if and only if \((C,A)\) is observable.

For two minimal observationally equivalent systems \((A,B,C,D)\) and \((\bar{A},\bar{B},\bar{C},\bar{D}) = (TAT^{-1},TB,CT^{-1},D), T \in GL(n)\), the Gramians are related by:

\[
\bar{P} = TPT' \quad (A.26)
\]

\[
\bar{Q} = T^{-1}QT^{-1} \quad (A.27)
\]

This shows that the eigenvalues of \( PQ \) and \( \bar{P}\bar{Q} \) stay invariant within the class of observationally equivalent minimal state-space realizations. The positive square roots of the eigenvalues of \( PQ \) are known as the Hankel singular values or second order modes of the corresponding transfer function\(^1\) and will be denoted by \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0 \).

Moreover, it is easy to show the existence of a state transformation \( T \in GL(n) \) such that an arbitrary minimal and stable \((A,B,C,D)\) becomes Lyapunov balanced:

\(^1\)The Hankel singular values of a continuous time transfer function \( k_c(s) \) having no poles on the imaginary axis are defined as the singular values of the associated Hankel operator \( \Gamma_k \). Decomposing \( k_c(s) \) according to its poles into \( k_c(s) = k_{c,s}(s) + k_{c,u}(s) + \lim_{\omega \to \infty} k_c(i\omega) \) it can be shown that the Hankel operator depends only on the stable part \( k_{c,s}(s) \). For instance, an antistable \( k_c(s) \) would yield \( \Gamma_k = 0 \). In case of a stable \( k_c(s) \), the Hankel singular values can be computed as the positive square roots of the eigenvalues of \( PQ \) for any minimal realization.
DEFINITION A.3.1 (LYAPUNOV BALANCING). A minimal state-space representation of $k_c(s) \in \mathbb{M}_e^+(n)$ is called Lyapunov balanced, if the solutions to the LFs (A.24) and (A.25) satisfy $P = Q = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$, where in addition $\sigma_1 \geq \cdots \geq \sigma_n > 0$. The $\sigma_i$ are called second order modes or Hankel singular values.

However, this state transformation $T \in GL(n)$ is in no case unique: there is more than one balanced realization in each equivalence class. In fact, it is easy to see that the class of balanced state-space realizations is only determined up to state transformations of the form

$$T_{\text{bal}} = \begin{pmatrix} T_1 & \cdots & T_k \\ \end{pmatrix}, \quad T_i \in O(n_i) \subset \mathbb{R}^{n_i \times n_i}$$

(A.28)

where $n_1, \ldots, n_k$ denote the multiplicities of the Hankel singular values (hence, $\sum_{i=1}^k n_i = n$) and $O(n_i)$ again denotes the orthogonal group of $n_i \times n_i$ matrices. Note that if all Hankel singular values are distinct, then the set of feasible state transformations $T_{\text{bal}}$ relating balanced realizations consists of all diagonal matrices having only +1 or −1 as diagonal entries. On the other hand, if all second order modes are equal and nonzero, then the set consists of all orthogonal $n \times n$ matrices.

For later reference we want to make a remark at this place: Let $\mathbb{M}_{s,1}^{+,0}(n)$ be the set of stable transfer functions with the greatest second order mode $\sigma_1$ less than one and $(D + D') > 0$. Then $\mathbb{M}_{s,1}^{+,0}(n)$ can easily be linked homeomorphically. Even more, in terms of state-space representations there is a mapping

$$S_s^{(0,1)} : \mathbb{M}_e^{+,0}(n) \to \mathbb{M}_{s,1}^{+,0}(n)$$

(A.29)

such that $S_s^{(0,1)}$ is a homeomorphism preserving system equivalence. Moreover, if $(A, B, C, D)$ is balanced with $P = Q = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$, then $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is balanced with $\tilde{P} = \tilde{Q} = \tilde{\Sigma} = \text{diag}(\sqrt{\frac{\sigma_1}{1+\sigma_1}}, \ldots, \sqrt{\frac{\sigma_n}{1+\sigma_n}})$. Finally, if $(A, B, C, D)$ is in Ober’s Lyapunov balanced canonical form, then $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is obtained by simply replacing the $\sigma_i$ in the canonical form (4.50) by $\rho_i = \frac{\sigma_i}{\sqrt{1+\sigma_i}}$.

A.3.2 Positive real balancing

Given a minimal realization of a $k_c(s) \in \mathbb{M}_{e,spr}(n)$ the continuous time positive real Riccati equations (PR-AREs) (A.22) and (A.23) were given in section (A.2) and are repeated here for ease of reference:

$$AP + PA' + (B - PC')(D + D')^{-1}(B - PC')' = 0$$

(A.30)

$$A'Q + QA' + (C' - QB)(D + D')^{-1}(C' - QB)' = 0$$

(A.31)

DEFINITION A.3.2 (POSITIVE REAL BALANCING). A minimal state-space representation of $k_c(s) \in \mathbb{M}_{e,spr}(n)$ is called positive real balanced, if the minimal (l) solutions to the PR-AREs (A.30) and (A.31) satisfy $P = Q = \Sigma = \text{diag}(\sigma_1, \cdots, \sigma_n)$, where in addition $\sigma_1 \geq \cdots \geq \sigma_n > 0$. The $\sigma_i$ are called positive real singular values.

An interesting observation is the existence of a well defined mapping from the set $\mathbb{M}_{s,1}^{+,0}(n)$ of stable transfer functions with the greatest second order mode $\sigma_1$ less than one and $(D + D') > 0$ to the set of stable and strictly positive real transfer functions which takes the following form (cf. the mapping $I_{XP}$ of definition 5.2 in (Ober, 1996)).
where $P$ and $Q$ are the unique solutions of the Lyapunov equations (A.24) and (A.25) and the matrix square root is taken to be the positive lower triangular Cholesky factor. It can be shown that $S_{s_{n}(0,1)}^{n_{s_{n}(0,1)}}$ is bijective and preserves system equivalence; see theorem 5.1. in (Ober, 1996). Moreover, as the solutions $P$ and $Q$ depend continuously on $(A,B,C,D)$ and $\lambda_{max}(PQ) < 1$, $S_{s_{n}(0,1)}^{n_{s_{n}(0,1)}}$ is continuous, its inverse is also continuous, hence it is a homeomorphism. Finally, if $P$ and $Q$ are the solutions to the Lyapunov equations (A.24) and (A.25), then the minimal solutions to the positive real Riccati equations (A.30) and (A.31) are given by $(I - PQ)P$ and $Q(I - PQ)^{-1}$, respectively (cf. proposition 5.1 in (Ober, 1996)). Hence, we have the following

**Lemma A.3.1.** If $(A,B,C,D)$ is a minimal realization of a stable transfer function with $\sigma_1 < 1$, i.e. $k_c(s) \in \mathbb{M}_{s_{n}(0,1)}^{C_{+}}(n)$ which is Lyapunov balanced with respect to Gramians $P = Q = \Sigma = diag(\sigma_1, \ldots, \sigma_n)$, then the state-space realization obtained from $S_{s_{n}(0,1)}^{n_{s_{n}(0,1)}}(A,B,C,D)$ through $T = (I - PQ)^{-1/2}$ is positive real balanced with respect to the same Gramian $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$.

**Proof.** It is easy to see that the minimal solutions $P$ and $Q$ to the positive real Riccati equations (A.30) and (A.31) also transform according to (A.26) and (A.27) if an observationally equivalent state-space realization is considered. Hence, by applying $T = (I - PQ)^{-1/2}$ to $S_{s_{n}(0,1)}^{n_{s_{n}(0,1)}}(A,B,C,D)$, the solutions $(I - PQ)P$ of (A.30) and $Q(I - PQ)^{-1}$ of (A.31) become $diag(\sigma_1, \ldots, \sigma_n)$.

A.3.3 Minimum phase balancing

Given a minimal realization of a $k_c(s) \in \mathbb{M}_{s_{n}(0,1)}^{C_{+}}(n)$ the controllability (observability) Lyapunov equation corresponding to $k_c(s)$ and the observability (controllability) Lyapunov equation corresponding to $k^{-1}_c(s)$ are given by

\begin{align}
AP + PA' + BB' &= 0 \\
\bar{A}Q + QA' + C' E^{-1} E^{-1} C &= 0 \\
A' \bar{Q} + QA + C'C &= 0 \\
\bar{A} \bar{P} + P \bar{A} + B E^{-1} E^{-1} B' &= 0
\end{align}

where $\bar{A} = (A - BE^{-1}C)$.

**Definition A.3.3 (Right and Left Minimum Phase Balancing).** A minimal state-space representation of $k_c(s) \in \mathbb{M}_{s_{n}(0,1)}^{C_{+}}(n)$ is called right minimum phase balanced, if the solutions to the LEs (A.33) and (A.34) satisfy $P = Q = \Sigma = diag(\sigma_1, \ldots, \sigma_n)$. It is called left minimum phase balanced, if the solutions to the LEs (A.35) and (A.36) satisfy $P = Q = \Sigma = diag(\sigma_1, \ldots, \sigma_n)$. Here, $\sigma_1 \geq \cdots \geq \sigma_n > 0$ holds true. The $\sigma_i$ are called right (left) minimum phase singular values.

A direct consequence from theorem (A.2.3) is the existence of a well defined mapping from the set of stable and strictly positive real to the set of stable and strictly minimum phase transfer functions, which for the unique left spectral factor takes the following form (cf. the mapping $S_{PM}$ of proposition 7.2 in (Ober, 1991)):

\begin{align}
S_{s_{n}(0,1)}^{n_{s_{n}(0,1)}} : \mathbb{M}_{s_{n}(0,1)}^{C_{+}}(n) &\to \mathbb{M}_{s_{n}(0,1)}^{C_{+}}(n) \\
\begin{bmatrix} A & B \\ C & D \end{bmatrix} &\mapsto \begin{bmatrix} A \\ (D + D')^{-1/2}(C - B'Q) \\ (D + D')^{1/2} \end{bmatrix}
\end{align}
where $Q$ is the minimal solution to the dual PR-ARE \((A.23)\) and the matrix square root is taken to be the positive lower triangular Cholesky factor. It can be shown that $S^\text{emp}_{n,\text{spr}}$ is bijective and preserves system equivalence; see (Ober, 1991). Moreover, as the minimal solution $Q$ depends continuously on $(A, B, C, D)$ (see, e.g. (Chou, 1994)), it is obvious that $S^\text{emp}_{n,\text{spr}}$ is continuous. Its inverse is given by

\[
(S^\text{emp}_{n,\text{spr}})^{-1} : M^\text{emp}_{n,\text{spr}}(n) \rightarrow M^\text{emp}_{n,\text{spr}}(n)
\]

\[
\begin{bmatrix}
A & B \\
C & E
\end{bmatrix} \rightarrow \begin{bmatrix}
A & B \\
E' + B'Q & \frac{B}{E'E}
\end{bmatrix}
\]

and is also continuous, hence it is a homeomorphism. Finally, by applying a state transformation $T = (I - PQ)^{1/4}$ on $S^\text{emp}_{n,\text{spr}}(A, B, C, D)$, we have:

**Lemma A.3.2.** If $(A, B, C, D)$ is a minimal realization of a strictly positive real transfer function $k_c(s) \in M^\text{emp}_{n,\text{spr}}(n)$ which is positive real balanced with respect to Gramians $P = Q = \text{diag}(p_1, \ldots, p_n)$, then the state-space realization obtained from $S^\text{emp}_{n,\text{spr}}(A, B, C, D)$ through $T = (I - PQ)^{1/4}$ is left minimum phase balanced with respect to Gramian $\Sigma = \text{diag}(\frac{p_1}{\sqrt{1-p_1^2}}, \ldots, \frac{p_n}{\sqrt{1-p_n^2}})$. Here the matrix square root is normalized such that $(D + D')^{1/2} - (D + D')^{-1/2}(C - B'Q)(I - PQ)^{-1/4}(I + (I - PQ)^{1/4}A(I - PQ)^{-1/4})^{-1}(I - PQ)^{1/4}B$ is positive lower triangular.

**Remark A.3.1.** Note that for $T = (I - PQ)^{1/4}$ being nonsingular it is sufficient that the largest positive real singular value $p_1$ is less than 1.

**Proof.** One has the show that the system matrices in

\[
\begin{bmatrix}
(I - PQ)^{1/4}A(I - PQ)^{-1/4} \\
(D + D')^{-1/2}(C - B'Q)(I - PQ)^{-1/4} \\
(D + D')^{1/2}
\end{bmatrix}
\]

satisfy (A.35) and (A.36) for $P = Q = \Sigma = \text{diag}(\frac{p_1}{\sqrt{1-p_1^2}}, \ldots, \frac{p_n}{\sqrt{1-p_n^2}})$. (A.35) is easy to show: Let us insert the system matrices (A.39) for $(A, B, C, D)$ and $Q = \Sigma$ into the left hand side of (A.35) and pre and post multiply by $(I - PQ)^{1/4}$ to get

\[
A(I - PQ)^{1/4}\Sigma(I - PQ)^{1/4} + (I - PQ)^{1/4}\Sigma(I - PQ)^{1/4}A + (C - B'Q)'(D + D')^{-1/2}(C - B'Q) = A(I - PQ)^{1/4} + (I - PQ)^{1/4}A + (C - B'Q)'(D + D')^{-1/2}(C - B'Q) = A(I - PQ)^{1/4} + (I - PQ)^{1/4}A + (C - B'Q)'(D + D')^{-1/2}(C - B'Q)
\]

The last equality follows from the fact that $(A, B, C, D)$ is positive real balanced. For the proof that $\Sigma$ also satisfies (A.36), see theorem 5.10 (respectively 4.51) in (McGinnie, 1993).

**A.3.4 Stochastic balancing**

The notion of stochastic balancing has been introduced in (Desai and Pal, 1982) for the continuous time case and in (Desai and Pal, 1984) for the discrete time case. According to our definition below, the term *stochastically balanced* actually refers to a triple of state-space realizations. However, as will be shown below, we may also use the term *stochastically balanced* for a state-space realization of a transfer function in $M^\text{emp}_{n,\text{spr}}$ only.
Definition A.3.4 (Stochastic balancing). Let a triple of minimal realizations of a spectral summand and the corresponding left and right spectral factors be given as in theorem (A.2.3) above, i.e.

\[
\Phi_c(s) = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad k_c^L(s) = \begin{bmatrix} A & K \\ C & E \end{bmatrix}, \quad k_c^R(s) = \begin{bmatrix} A & B \\ L & E \end{bmatrix}
\]

with \( K, L \) and \( E \) given in the same theorem. This triple of state-space realizations is said to be stochastically balanced with Gramian \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \) if \((A, K, L)\) is Lyapunov balanced with the same Gramian. Here, \( 1 > \rho_1 \geq \ldots \geq \rho_n > 0 \) holds true. The \( \rho_i \) are called canonical correlation coefficients.

Whereas for a positive real balanced realization of a given \( k_c(s) \in M_{s, \text{apr}}^P \), we had to apply an additional (diagonal) state transformation in order to make \( S_{s, \text{apr}}^P(A, B, C, D) \) left minimum phase balanced, this state transformation need not be applied in the stochastically balanced case:

Lemma A.3.3. If \((A, B, C, D)\) is a minimal realization of a strictly positive real transfer function \( k_c(s) \in M_{s, \text{apr}}^P(n) \), then \((A, B, C, D)\) is positive real balanced with respect to Gramians \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \) and \( \rho_1 < 1 \) if and only if the triple of state-space realizations \((A, B, C, D), (A, B, L, E) = S_{s, \text{apr}}^P(A, B, C, D) \) and \((A, K, C, E)\) (corresponding to the right spectral factor according to theorem (A.2.3)) is stochastically balanced with respect to the same Gramian.

Proof. \((A, B, C, D)\) is positive real balanced if \( \text{diag}(\rho_1, \ldots, \rho_n) \) is the minimal solution to (A.30) and (A.31). The mapping (A.37) leaves \( A \) and \( B \) unchanged, \( C \) is mapped to \( L = (D + D')^{-1/2}(C - B'Q) \) and for the right spectral factor we get \( K = (B - PC')(D + D')^{-1/2} \). The state-space triple is therefore stochastically balanced if

\[
\begin{align*}
AP + PA' + KK' &= AP + PA' + (B - PC')(D + D')^{-1}(B - PC')' \\
A'Q + AQ' + L'L &= A'Q + QA' + (C - B'Q)'(D + D')^{-1}(C' - QB)'
\end{align*}
\]

Clearly, these equations coincide with (A.30) and (A.31).

This lemma shows that we may alternatively define stochastic balancing involving only (strictly) minimum phase transfer function in \( M_{s, \text{mp}}^P \) as it is shown in the table at the beginning of this section A.3:

Definition A.3.5 (Right and left stochastic balancing). A minimal state-space representation \((A, B, C, E)\) of \( k_c(s) \in M_{s, \text{mp}}^P(n) \) is called right stochastically balanced with Gramian \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \) with \( \rho_1 < 1 \), if \((A, BE' + PC', C, 1 \frac{1}{2} EE')\) is positive real balanced, i.e. \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \) is the minimal solution to the equations

\[
\begin{align*}
AP + PA' + (BE' + PC' - PC')(EE')^{-1}(BE' + PC' - PC')' &= AP + PA' + BB' = 0 \\
A'Q + QA' + [C'Q(BE' + PC')]'(EE')^{-1}[C'Q(BE' + PC')]' &= 0
\end{align*}
\]

It is called left stochastically balanced with Gramian \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \) with \( \rho_1 < 1 \), if \((A, B, E'C + B'Q, \frac{1}{2} EE') = (S_{s, \text{mp}}^P)^{-1}(A, B, C, E)\) is positive real balanced, i.e. \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \) is the minimal solution to the equations

\[
\begin{align*}
AP + PA' + [B - P(C'E + QB)](EE')^{-1}[B - P(C'E + QB)]' &= 0 \quad \text{(A.40)} \\
A'Q + QA' + (C'E + QB - QB)(EE')^{-1}(C'E + QB - QB)' &= A'Q + QA' + C'C = 0 \quad \text{(A.41)}
\end{align*}
\]

From this definition it is also clear that right (left) stochastically balanced and right (left) minimum phase balanced state-space realizations of \( k_c(s) \in M_{s, \text{mp}}^P \) are related by a diagonal state-transformation of the form \((I - PQ)^{-1/4}\) (respectively \((I - PQ)^{1/4}\) where, of course, \( P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \). Clearly, then the right (=left) minimum phase singular values \( \sigma_i \) and the right (=left) canonical correlations \( \rho_i \) satisfy \( \sigma_i = \frac{\rho_i}{\sqrt{1 - \rho_i^2}} \); see also lemma (A.3.2).
A.3.5 Summary of homeomorphisms relating balanced realizations

The following diagram summarizes the homeomorphic mappings between the different model classes considered in this section. Note that $\rho_i = \sqrt{\frac{\sigma_i}{1+\sigma_i}}$ and $\sigma_i = \sqrt{1-\rho_i^2}$:

\[
\begin{align*}
M_c^{\mu^+}(n) \longrightarrow & \text{ Lyapunov balanced: (A.24), (A.25) with } P = Q = \text{diag}(\sigma_1, \ldots, \sigma_n) \\
S_{\mu^+}^{(0,1)} \downarrow & \text{ } \\
M_c^{\mu^+}(n) \longrightarrow & \text{ Lyapunov balanced: (A.24), (A.25) with } P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \\
S_{\mu^+}^{(0,1)} \downarrow & \text{ } \\
M_c^{s,spr}(n) \longrightarrow & (I - PQ)P \text{ solves (A.30) and } Q(I - PQ)^{-1} \text{ solves (A.31)} \\
S_{(I - rQ)^{-1/2}} \downarrow & \text{ } \\
M_c^{s,spr}(n) \longrightarrow & \text{ Positive real balanced: (A.30), (A.31) with } P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \\
S_{s,spr}^{+} \downarrow & \text{ } \\
M_c^{s,mp}(n) \longrightarrow & \text{ Left stoch. balanced: (A.40), (A.41) with } P = Q = \text{diag}(\rho_1, \ldots, \rho_n) \\
S_{(I - rQ)^{1/2}} \downarrow & \text{ } \\
M_c^{s,mp}(n) \longrightarrow & \text{ Left minimum phase balanced: (A.35), (A.36) with } P = Q = \text{diag}(\sigma_1, \ldots, \sigma_n)
\end{align*}
\]

A.4 Relation between discrete and continuous time state-space systems

The following mapping is a simple Moebius transformation which is known under the term bilinear transformation:

\[
\rho_1 : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto \frac{1-z}{1+z} = s
\]  

(A.42)

Note that $\rho_1$ is a bijection on the compactified complex plane (with inverse $\rho_1^{-1}(s) = z = \frac{1-s}{1+s}$) mapping the complement of the closed unit disk onto the open left half plane; $z = 1$ is mapped to $s = 0$, both corresponding to frequency zero if $z$ and $s$ are interpreted as complex variables in continuous time and discrete time transfer functions.

From these observations it is clear that we can link continuous time transfer functions $k_c(s)$ and discrete time transfer functions $k(z)$ through

\[
k(z) = k_c(\rho_1(z))
\]

(A.43)

preserving the stability and minimum-phase property. It can be shown that (A.43) also preserves the $l_\infty$-norm, whereas the $l_2$-norm of the transfer functions is changed under (A.43).

In terms of state-space representations $(A_c, B_c, C_c, D_c)$ for $k_c(s)$ and $(A, B, C, D)$ for $k(z)$ the transformation (A.43) can be chosen to be of the form (A.44) given in the theorem below. For the definitions of the various continuous time and discrete time balancing schemes and the corresponding Gramians see sections (A.3) and (3.4), (4.7) and (4.8), respectively. The rest of the terms appearing in the theorem below have already been defined:
Theorem A.4.1. Let \((A_c, B_c, C_c, D_c)\) be a (not necessarily minimal) state-space representation of some (not necessarily stable) continuous time transfer function \(k_c(s)\) and let \(\lambda_i(A_c) \neq 1, i = 1, \ldots, n\) hold true. Then \((A, B, C, D) = \rho(A_c, B_c, C_c, D_c)\) is a state-space representation of the discrete time transfer function \(k(z)\) given in (A.43) where

\[
\begin{align*}
\rho(A_c, B_c, C_c, D_c) &= (A, B, C, D) \\
A &= (I + A_c)(I - A_c)^{-1} \\
B &= \sqrt{2}(I - A_c)^{-1}B_c \\
C &= \sqrt{2}C_c(I - A_c)^{-1} \\
D &= D_c + C_c(I - A_c)^{-1}B_c
\end{align*}
\]

(i) \(\rho\) is a homeomorphism when considered as a mapping from the sets of

\[
\begin{align*}
\{(A_c, B_c, C_c, D_c) : (A_c, B_c, C_c, D_c) \text{ is minimal and stable} &\} &\to& \{(A, B, C, D) : (A, B, C, D) \text{ is minimal and stable} &\} \\
\{(A, B, C, D) : (A, B, C, D) \text{ is minimal, stable and strictly positive real} &\} &\to& \{(A, B, C, D) : (A, B, C, D) \text{ is minimal, stable and strictly positive real} &\}
\end{align*}
\]

(ii) \(\rho\) preserves observational equivalence for minimal systems, i.e.

\[
(A_c, B_c, C_c, D_c) \sim (A_{c,1}, B_{c,1}, C_{c,1}, D_{c,1}) \iff \rho(A_c, B_c, C_c, D_c) \sim \rho(A_{c,1}, B_{c,1}, C_{c,1}, D_{c,1})
\]

(iii) \(\rho\) preserves balancing: \((A_c, B_c, C_c, D_c)\) is continuous time Lyapunov balanced (positive real balanced, stochastically balanced, minimum phase balanced) if and only if \((A, B, C, D)\) is discrete time Lyapunov balanced (positive real balanced, stochastically balanced, minimum phase balanced). Moreover, the corresponding pair of Gramians stays invariant under the transformation \(\rho\).

Proof. From (A.43) we get

\[
k(z) = k_c(s)_{s = \frac{1}{1 + z}} = C\left(\frac{1}{1 + z}I - A_c\right)^{-1}B_c + D_c
\]

which, after tedious manipulations, can be written as

\[
k(z) = \sqrt{2}C_c(I - A_c)^{-1}(z^{-1}I - (I + A_c)(I - A_c)^{-1})\sqrt{2}(I - A_c)^{-1}B_c + D_c + C_c(I - A_c)^{-1}B_c
\]

yielding (A.44).

To show (i), one can easily see that for a (continuous-time) stable \(A_c\), the matrix \((I + A_c)(I - A_c)^{-1}\) is always well defined and (discrete-time) stable. Moreover, if \((A_c, B_c, C_c, D_c)\) is stable and minimal, then one can show by simple algebra that the unique solutions \(P > 0\) and \(Q > 0\) to the continuous time Lyapunov equations (A.24) and (A.24) in section (A.3) also satisfy the discrete-time Lyapunov equations (3.16) and (3.16) in section (4.6). As \(A\) is known to be stable from what was said above, this implies that \((A, B, C, D)\) is minimal: The image of \(\rho\) of all minimal and (continuous-time) stable systems solely consists of minimal and (discrete-time) stable systems. Furthermore, it is straightforward to see that the mapping

\[
\begin{align*}
\hat{\rho}(A, B, C, D) &= (A_c, B_c, C_c, D_c) \\
A_c &= (I + A)^{-1}(A - I) \\
B_c &= \sqrt{2}(I + A)^{-1}B \\
C_c &= \sqrt{2}C(I + A)^{-1} \\
D_c &= D - C(I + A)^{-1}B
\end{align*}
\]
satisfies an analogous property: The image of $\hat{\rho}$ of all minimal and (discrete-time) stable systems only consists of minimal and (continuous-time) stable systems. Straightforward calculations show that $\hat{\rho} = \rho^{-1}$, i.e. $\rho$ is a bijection on the set of minimal and stable systems. Continuity of $\rho$ and $\rho^{-1}$ is evident because matrix additions and multiplications are trivially continuous and the inversion is continuous for stable $A_c$. This shows the first statement of (i), and the others can be derived in an analogous manner; see (Ober, 1991), for instance.

Statement (ii) follows by simple calculations.

For the case of minimal and stable systems, it has already been mentioned that the controllability and observability Gramians are invariant under $\rho$, which in particular implies that $\rho$ preserves balancing. Invariance of the corresponding Gramians (A.33) - (A.36) and (4.68) - (4.71) in the case of minimal, stable and strictly minimum phase systems is proved in lemma 4.9 (and its dual version on page 106) in (McGinnie, 1993). Remark 8.1 in (Ober, 1991) and the references there also deal with the analogous statement for minimal, stable and strictly positive real systems. 

Remark A.4.1. From (i) and (ii) it follows that any canonical form e.g. for stable and minimal continuous time state-space representations directly induces a canonical form for stable and minimal discrete time state-space representations. Due to (iii), a continuous time balanced canonical form will yield a discrete time balanced canonical form.

A.5 Partitioned Sylvester equations

Given the $n \times n$ matrices $A^{(k)}$, $A^{(l)}$ and $Q$, equations of the form $P = A^{(k)} PA^{(l)\prime} + Q$ are known as Sylvester equation in the literature. Clearly, for $A = A^{(k)} = A^{(l)\prime}$, we obtain a Lyapunov equation of the form (1.11).

Assume we are given a Sylvester equation $P = A^{(k)} PA^{(l)\prime} + Q$ where both $A^{(k)}$ and $A^{(l)}$ have a block triangular structure and $Q$ is of a particular form:

$$ P = \begin{pmatrix} A_{11}^{(k)} & 0 \\ A_{21}^{(k)} & A_{22}^{(k)} \end{pmatrix} P \begin{pmatrix} A_{11}^{(l)} & 0 \\ A_{21}^{(l)} & A_{22}^{(l)} \end{pmatrix} \prime + \begin{pmatrix} K_{1}^{(k)} \\ K_{2}^{(k)} \end{pmatrix} \Sigma \begin{pmatrix} K_{1}^{(l)\prime} & K_{2}^{(l)\prime} \end{pmatrix} $$ (A.46)

Partitioning $P$ according to

$$ P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} $$

and multiplying the submatrices we immediately get the following set of equations for the blocks $P_{ij}$:

$$ P_{11} = A_{11}^{(k)} P_{11} A_{11}^{(l)\prime} + K_{1}^{(k)} \Sigma K_{1}^{(l)\prime} $$ (A.47)

$$ P_{12} = A_{11}^{(k)} P_{12} A_{22}^{(l)\prime} + A_{11}^{(k)} P_{11} A_{21}^{(l)\prime} + K_{1}^{(k)} \Sigma K_{2}^{(l)\prime} $$

$$ P_{21} = A_{22}^{(k)} P_{21} A_{11}^{(l)\prime} + A_{22}^{(k)} P_{22} A_{21}^{(l)\prime} + K_{2}^{(k)} \Sigma K_{1}^{(l)\prime} $$

$$ P_{22} = A_{22}^{(k)} P_{22} A_{22}^{(l)\prime} + A_{22}^{(k)} P_{11} A_{22}^{(l)\prime} + A_{22}^{(k)} P_{12} A_{21}^{(l)\prime} + A_{22}^{(k)} P_{21} A_{21}^{(l)\prime} + K_{2}^{(k)} \Sigma K_{2}^{(l)\prime} $$

Note that the first equation contains $P_{11}$ only. Given $P_{11}$, the second and third equation are again usual Sylvester equations etc. Thus, the solution of the original Sylvester equation (A.46) can be given in terms of the solutions of "nested" and "smaller" Sylvester equations in (A.47).

A.6 Real analytic – and Riemannian manifolds

In this section, some basic definitions are presented:
A.6. REAL ANALYTIC - AND RIEMANNIAN MANIFOLDS

**Definition A.6.1 (Real analytic manifold of dimension n).** A real analytic manifold of dimension n is a separable Hausdorff space \( \mathbb{M} \) together with a family \( (U_\alpha, \phi_\alpha), \alpha \in I \) where \( I \) is some index set. \( U_\alpha \) is an open subset of \( \mathbb{M} \) and \( \phi_\alpha \) is a homeomorphism from \( U_\alpha \) onto an open subset of \( \mathbb{R}^n \) satisfying the following properties:

(i) \( U_\alpha \) is an open covering of \( \mathbb{M} \)

(ii) If \( U_\alpha \cap U_\beta \neq \emptyset \), then \( \phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta) \) is an analytic function, i.e. has a convergent power series expansion in some neighborhood of each point in \( \phi_\alpha(U_\alpha \cap U_\beta) \subseteq \mathbb{R}^n \).

\( U_\alpha \) is called a coordinate neighborhood, \( \phi_\alpha \) a coordinate mapping, \( ((U_\alpha, \phi_\alpha), \alpha \in I) \) a system of local coordinates, and finally \( \phi_\beta \circ \phi_\alpha^{-1} \) is called a coordinate transformation.

**Remark A.6.1.** Each pair \( (U_\alpha, \phi_\alpha) \) is also called a (coordinate) chart, and the family \( ((U_\alpha, \phi_\alpha), \alpha \in I) \) may be called an atlas.

**Remark A.6.2.** A space is called Hausdorff if for any two different points there are disjoint neighborhoods for each point. This implies the uniqueness of limits of convergent series. A space is called separable if it has a countable dense subset.

In any of the following definitions, \( \mathbb{M} \) is assumed to be a real analytic manifold of dimension \( n \):

**Definition A.6.2 (Differentiability).** Let \( p \in \mathbb{M} \) and let \( U \) be an open neighborhood of \( p \). A function \( f : U \to \mathbb{R} \) is said to be differentiable in \( p \) if for one (and thus for every) chart \( (U_\alpha, \phi_\alpha) \) with \( p \in U_\alpha \) the mapping \( f \circ \phi_\alpha^{-1} \) is differentiable at \( \phi_\alpha(p) \). If the derivative satisfies \( (f \circ \phi_\alpha^{-1})'(\phi_\alpha(p)) = 0 \), then \( p \) is said to be a stationary point of \( f \).

Using this definition we can now consider the set \( \mathcal{F}_p \) of all functions that are defined in some (possibly varying) open neighborhood of \( p \) and are differentiable in \( p \). If \( f, g \in \mathcal{F}_p \) and \( c \in \mathbb{R} \), then \( (f + g) \in \mathcal{F}_p \) and \( c \cdot f \in \mathcal{F}_p \), where the domain of definition of \( f + g \) is the intersection of the domains of definition of \( f \) and \( g \), respectively:

**Definition A.6.3 (Tangent vector and tangent space).** A mapping \( X_p : \mathcal{F}_p \to \mathbb{R} \) is called a tangent vector to \( \mathbb{M} \) at \( p \) if it satisfies the following properties:

(i) If \( p \) is a stationary point of \( f \), then \( X_p(f) = 0 \).

(ii) \( X_p(f + g) = X_p(f) + X_p(g) \).

(iii) \( X_p(c \cdot f) = c \cdot X_p(f) \).

The set of all tangent vectors to \( \mathbb{M} \) at \( p \) is called the tangent space to \( \mathbb{M} \) at \( p \) and is denoted by \( Q^\mathbb{M}_p \). Defining the operations \( (X_p + Y_p)(f) = X_p(f) + Y_p(f) \) and \( (c \cdot X_p)(f) = c \cdot (X_p(f)) \), \( Q^\mathbb{M}_p \) is a linear vector space of dimension \( n \).

**Remark A.6.3.** For any chart \( (U_\alpha, \phi_\alpha) \) with \( p \in U_\alpha \) and all \( f \in \mathcal{F}_p \), the mappings \( E_{i_p} \) defined by

\[
E_{i_p}(f) := \frac{\partial f \circ \phi^{-1}_\alpha}{\partial x^i}(\phi_\alpha(p)), \quad i = 1, \ldots, n
\]

satisfy \( E_{i_p} \in Q^\mathbb{M}_p \), where \( \partial / \partial x^i \) denotes differentiation with respect to the \( i \)-component. They are called coordinate frames and form a basis of \( Q^\mathbb{M}_p \).

Note that the notion of the tangent space is independent of the local coordinates being used at \( p \). Similarly, it does not depend on an embedding of \( \mathbb{M} \) into some "larger" vector space or manifold.

**Definition A.6.4 (Real analytic submanifold).** Let \( \mathbb{M} \) be a real analytic manifold of dimension \( n \). A subset \( \mathbb{A} \subseteq \mathbb{M} \) is called a k-dimensional real analytic submanifold of \( \mathbb{M} \) if every point \( a \in \mathbb{A} \) possesses a coordinate chart \( (U, \phi) \) around \( a \in U \) such that \( \phi(U \cap \mathbb{A}) = \phi(U) \cap (\mathbb{R}^k \times \{0\}) \) for some integer \( 0 \leq k \leq n \). The number \( n - k \) is called the codimension of the submanifold \( \mathbb{A} \).

**Remark A.6.4.** In particular, real analytic submanifolds \( \mathbb{A} \subseteq \mathbb{M} \) are also real analytic manifolds. If \( \mathbb{A} \subseteq \mathbb{M} \) is a submanifold, then at each point \( p \in \mathbb{A} \), the tangent space \( Q^\mathbb{A}_{p} \subseteq Q^\mathbb{M}_p \) is considered as a sub vector space of \( Q^\mathbb{M}_p \).
Remark A.6.5. Note that any open subset of a real analytic manifold is a real analytic submanifold of the same dimension. However, there are subsets $A$ of a real analytic manifold $M$ such that $A$ is a real analytic manifold but not a real analytic submanifold of $M$.

If $V$ is a $n$ dimensional vector space over $\mathbb{R}$, a mapping $\Phi : V \times V \to \mathbb{R}$ is called a bilinear form if it is linear in each argument separately. Once a basis for $V$ is chosen, there is a one-to-one correspondence between bilinear forms on $V$ and matrices in $\mathbb{R}^{n \times n}$. A bilinear form is called symmetric and positive definite if the corresponding matrix has these properties.

Definition A.6.5. A field $\Phi$ of differentiable bilinear forms on a manifold $M$ consists of a function assigning to each point $p \in M$ a bilinear form $\Phi_p$ on $Q_p^M$ such that for any coordinate neighborhood $(U_\alpha, \phi_\alpha)$ the functions $\alpha_{i,j} = \Phi(E_i, E_j)$ defined by $\Phi$ and the coordinate frames $E_1, \ldots, E_n$ are differentiable.

Next we will briefly introduce the concept of a Riemannian manifold. The idea here is to endow the tangent spaces of a real analytic manifold $M$ with a Euclidean geometry that may vary from point to point:

Definition A.6.6 (Riemannian manifold and Riemannian metric tensor). Let $\Phi$ be a differentiable field of symmetric and positive definite bilinear forms on $M$. Then $\Phi$ is called a Riemannian metric (tensor) and the pair $(M, \Phi)$ is called a Riemannian manifold.

This definition turns each tangent space $Q_p^M$ into a Euclidean vector space with inner product $\Phi_p$ which depends (as $Q_p^M$) on $p \in M$.

Remark A.6.6. Let $(U_\alpha, \phi_\alpha)$ be a coordinate neighborhood where $p = \phi^{-1}_\alpha(p) \in U_\alpha$ and let us denote by $F$ a bijective and real analytic change of local coordinates, say from $\tau_\alpha$ to $F(\tau_\alpha) = \tau_{\alpha'}$. If $R(\tau_{\alpha'}) \in \mathbb{R}^{n \times n}$ denotes the symmetric and positive definite matrix corresponding to $\Phi_{\tau_{\alpha'}}$, then the matrix $R(\tau_{\alpha'})$ transforms according to

$$R(\tau_{\alpha'}) = \frac{dF}{d\tau_{\alpha'}}(\tau_{\alpha'})^{-1} R(\tau_{\alpha'}) \frac{dF}{d\tau_{\alpha'}}(\tau_{\alpha'})^{-1}$$

It can be shown that each real analytic manifold can be endowed with a differentiable Riemannian metric (tensor).

A.7 Acronyms

This section summarizes the notation used in the thesis. The symbols are grouped according to the chapter where they appear for the first time. The list only contains symbols which appear regularly throughout the thesis:

Chapter 1

- $y$: outputs
- $u$: exogenous inputs
- $\varepsilon$: white noise
- $x_i$: state vector
- $s$: dimension of outputs
- $m$: dimension of exogenous inputs
- $n$: dimension of state vector, order of a rational transfer function
- $(l, k) = C(z^{-1}I - A)^{-1}(B, K) + (D, I)$: transfer function
- $(L_j, K_j)_{j \in \mathbb{N}} = ((D, I), C(B, K), CA(B, K), \ldots)$: impulse response corresponding to $(l, k)$
- $\gamma(k) = \mathbb{E}(y_{t+k} - \mathbb{E}y_{t+k})(y_t - \mathbb{E}y_t)'$: autocovariance matrix
- $\gamma_y(k) = \mathbb{E}y_{t+k}y_t'$: matrix of second moments
- $U_A$: set of all rational and causal $s \times (m+s)$ transfer functions $(l, k)$ where $(l(0), k(0)) = (D, I)$, $D$ arbitrary
- $M(n)$: subset of $U_A$ where $(l, k)$ is of fixed order $n$
- $M_g(n)$: subset of $M(n)$ where $(l, k)$ is stable
- $M_{g|m}(n)$: subset of $M_g(n)$ where $k$ is (strictly) minimum phase
$S_n$: set of all $(A, B, C, D, K)$ for fixed $m$ and $s$ and variable $n$

$S_{n}(n)$: set of all $(A, B, C, D, K)$ for fixed $m$, $s$, $n$ $S_{n}(n)$: subset of $S(n)$ containing only minimal $(A, B, C, D, K)$

$\tilde{B} = (B, K)$

$C_n = (\tilde{B}, \tilde{A}B, \ldots, A^{2n-1}\tilde{B})$: controllability matrix

$O_n = (C', (C\tilde{A})', \ldots, (C\tilde{A}^{n-1})')'$: observability matrix

$C = (B, A\tilde{B}, \ldots)$: infinite controllability matrix

$O = (C', (C\tilde{A})', \ldots)'$: infinite observability matrix

$\mathcal{H} = OC$: Hankel matrix corresponding to transfer function $(l, k)$

$\pi$: mapping from $(A, B, C, D, K)$ (or parameters or ARMAX systems) to transfer functions

$\mathcal{E}(A, \tilde{B}, C, D); (l, k)$-equivalence class in $S_{n}(n)$

$(a, d, b)$: ARMAX system, triple of polynomial matrices

**Chapter 2**

$\hat{z}_{t|k}$: best linear least squares estimate of $z_t$ based on the finite past $\{u_1, \ldots, u_k\}$ and $\{y_1, \ldots, y_k\}$

$\hat{y}_{t|k}$: see $\hat{z}_{t|k}$

$\hat{u}_{t|k}$: see $\hat{z}_{t|k}$

$\epsilon_t = y_t - \hat{y}_{t|t-1}$: random variable the realization of which is obtained from the Kalman filter if initialized correctly

$\Sigma_t = \Sigma_{t|t-1}$: equal to $\mathbb{E}\epsilon_t \epsilon_t'$ under certain assumptions; see (2.20)

$(\tilde{A}, \tilde{B}, C, \tilde{D}, K) = (A - KC, B - KD, -C, -D, K)$: inverse system

$\mathcal{K} = (\tilde{B}, K, \tilde{A}(\tilde{B}, K), \ldots)$: infinite controllability matrix of inverse system

$O_J$: first $J$ block rows of $O$

$K_p$: first $p$ block columns of $K$

$\sigma$: parameter vector of on- and above-diagonal elements of $\Sigma$

$\Theta$: set of transfer functions $(l, k)$ and covariance parameters $\sigma$

$\hat{\Theta} \cup \{ (l, k) \text{ with poles on unit circle, singular } \Sigma \}$

$\mathcal{O}^*$: $\hat{\Theta} \cup \{ k \text{ with zeros on unit circle } \}$

$\theta = (l, k, \sigma)$: element of $\Theta$

$\theta_0 = (l_0, k_0, \sigma_0)$: element of $\Theta$ corresponding to the true $(l_0, k_0)$ and the true $\Sigma_0$

$Y_t = (y_1, \ldots, y_T)'$: vector of stacked outputs

$U_t = (u_1, \ldots, u_T)'$: vector of stacked exogenous inputs

$\tau$: vector of free parameters for transfer function $(l, k)$

$\tau^*$: vector of free parameters for $(\tilde{A}, \tilde{C}) = (\tilde{A}(\tau^*), \tilde{C}(\tau^*))$

$\tau^{**}$: vector of free parameters for $(\tilde{D}, \tilde{B}, \tilde{K}) = (\tilde{D}(\tau^{**}), \tilde{B}(\tau^{**}), \tilde{K}(\tau^{**}))$

$L_{T}(Y_t^T; U_t^T, \theta)$: likelihood function; either original form (see (2.67)) or prediction error form (see (2.70) and theorem (2.3.1)) or approximate form (see (2.76))

$L_{T}^{(a)}(Y_t^T; U_t^T, (l, k))$: concentrated likelihood function; see (2.77)

$L_{T}^{(a)}(Y_t^T; U_t^T, \tau^*)$: double concentrated "likelihood function"; see (2.85)

$L(\theta)$: asymptotic form of the likelihood function; see (2.86) or (2.87)

**Chapter 3**

$O(n)$: set of real orthogonal $n \times n$ matrices

$GL(n)$: set of real non singular $n \times n$ matrices

$Q_{(A, B, C, D)}$: tangent space to $\mathcal{E}(A, \tilde{B}, C, D)$ at $(A, B, C, D)$, viewed as affine subspace of $S(n)$

$Q$: matrix the columns of which span $Q_{(A, B, C, D)}$ when shifted by $(\tilde{A}, \tilde{B}, C, D)$

$\mathcal{E}^{\text{MFN}}(A, \tilde{B}, C, D) \subset \mathcal{E}(A, \tilde{B}, C, D)$: observationally equivalent systems of minimum Frobenius norm

$Q_{(A, B, C, D)}^{\text{MFN}}$: tangent space to $\mathcal{E}^{\text{MFN}}(A, B, C, D)$ at $(A, B, C, D)$, viewed as affine subspace of $S(n)$

$\mathcal{M}(n)$: set of all minimal realizations having minimum Frobenius norm within their $(l, k)$-equivalence classes

$\mathcal{E}^{\text{BSA}}(A, B, C, D) \subset \mathcal{E}(A, B, C, D)$: observationally equivalent balanced stable allpass systems

$Q_{(A, B, C, D)}^{\text{BSA}}$: tangent space to $\mathcal{E}^{\text{BSA}}(A, B, C, D)$ at $(A, B, C, D)$, viewed as affine subspace of $S(n)$

$\mathcal{BS}(n)$: set of all balanced stable allpass realizations in $S(n)$

$Q_{(A, B, C, D)}^{\text{BS}}$: tangent space to $\mathcal{BS}(n)$ at $(A, B, C, D)$, viewed as affine subspace of $S(n)$

$\Delta Y$: mapping from square system $(\tilde{A}, B, C, D)$ to original system $(A, B, C, D, K)$

$\mathcal{E}^{\text{CC}}(A, B, C, D)$: set of $L_{T}^{(a)}$-equivalent systems; see definition (3.6.1)

$Q_{(A, B, C, D)}^{\text{CC}}$: tangent space to $\mathcal{E}^{\text{CC}}(A, B, C, D)$ at $(A, B, C, D)$, viewed as affine subspace of $S(n)$

$Q^{\text{CC}}$: matrix the columns of which span $Q_{(A, B, C, D)}^{\text{CC}}$ when shifted by $(A, B, C, D)$

**Chapter 4**
Echelon state-space parametrization
\( V_\alpha \): set of transfer functions with Kronecker index \( \alpha \)
\( \tau_\alpha \in T_\alpha \): parameter vector, parameter space
\( \psi_\alpha : V_\alpha \to T_\alpha \): parametrization
\( \varphi_\alpha : T_\alpha \to S_m(n) \): mapping from parameters to system matrices

Overlapping state-space parametrization
\( U_\alpha \): set of transfer functions with structure index
\( \tau^{(1)}_\alpha \in T^{(1)}_\alpha \): parameter vector, parameter space
\( \psi^{(1)}_\alpha : U_\alpha \to T^{(1)}_\alpha \): parametrization
\( \varphi^{(1)}_\alpha : T^{(1)}_\alpha \to S_m(n) \): mapping from parameters to system matrices

Echelon ARMAX parametrization
\( \tau_\alpha \in T_\alpha \): parameter vectors, parameter spaces
\( \psi_\alpha : V_\alpha \to T_\alpha \): parametrizations

Overlapping ARMAX parametrizations
\( \tau^{(1)}_\alpha \in T^{(1)}_\alpha \): parameter vectors, parameter spaces
\( \psi^{(1)}_\alpha : U_\alpha \to T^{(1)}_\alpha \): parametrizations

Ober’s Lyapunov balanced parametrization
\( V_\delta \): set of stable transfer functions with multi index \( \delta \)
\( \tau_\delta \in T_\delta \): parameter vector, parameter space
\( \psi_\delta : V_\delta \to T_\delta \): parametrization
\( \varphi_\delta : T_\delta \to S_m(n) \): mapping from parameters to system matrices

Ober’s stochastically balanced parametrization
\( V^{(1)}_\delta \): set of stable and strictly miniphase transfer functions with multi index \( \delta \)
\( \tau^{(1)}_\delta \in T^{(1)}_\delta \): parameter vector, parameter space
\( \psi^{(1)}_\delta : V^{(1)}_\delta \to T^{(1)}_\delta \): parametrization
\( \varphi^{(1)}_\delta : T^{(1)}_\delta \to S_m(n) \): mapping from parameters to system matrices

McGinnie’s miniphase balanced parametrization
\( V^{(2)}_\delta \): set of stable and strictly miniphase transfer functions with multi index \( \delta \)
\( \tau^{(2)}_\delta \in T^{(2)}_\delta \): parameter vector, parameter space
\( \psi^{(2)}_\delta : V^{(2)}_\delta \to T^{(2)}_\delta \): parametrization
\( \varphi^{(2)}_\delta : T^{(2)}_\delta \to S_m(n) \): mapping from parameters to system matrices

DDLC
\( V_D = \pi(T_D) \): set of transfer functions of order \( n \) described by DDLC
\( \tau_D \in T_D \): parameter vector, parameter space
\( \psi_D^{\text{loc}} : V_D^{\text{loc}} \to T_D^{\text{loc}} \): local parametrization
\( \varphi_D : T_D \to S_m(n) \): mapping from parameters to system matrices

\( s1s\text{DDLC} \)
\( V_{D} = \pi(T_{D}) \): set of transfer functions of order \( n \) described by \( s1s\text{DDLC} \)
\( \tau_D \in T_D^{\text{loc}} \): parameter vector, parameter space
\( \psi_D^{\text{loc}} : V_D^{\text{loc}} \to T_D^{\text{loc}} \): local parametrization
\( \varphi_D : T_D \to S_m(n) \): mapping from parameters to system matrices

Chapter 5
\( (A_0, B_0, C_0, D_0, K_0) \): system corresponding to the true transfer function \((l_0, k_0)\)
\( (A, B, C, D, K) \): derivatives of system matrices w.r.t. system parameters \( \tau \)

\( L(\tau, \sigma) \): asymptotic form of the likelihood function, parametrized in \((\tau, \sigma)\)
\( L_T(\tau, \sigma) \): finite sample likelihood function, parametrized in \((\tau, \sigma)\)
\( L_T(\tau) \): finite sample concentrated likelihood function, parametrized in \( \tau \)
\( L_T^{\text{CC}}(\tau^o) \): finite sample double concentrated "likelihood function", parametrized in \( \tau^o \)

\( \hat{L} \): derivative of criterion function \( \mathcal{L} \) w.r.t. system parameters \( \tau \)
\( \mathcal{L}^{(kl)} \): (approximate) second derivative of criterion function \( \mathcal{L} \) w.r.t. \( k \) and \( l \) entry in \( \tau \)

\( \{ \psi_{\tau^T, T} \in GL(n) \} \): family of DDLC parametrizations at initial systems \((T A T^{-1}, TB, CT^{-1}, D), T \in GL(n) \)
\( \{ \psi_{D, Q}^\delta, Q \in O(n) \} \): family of \( s1s\text{DDLC} \) parametrizations at initial systems \((Q \hat{A} Q', QB, CQ', D), Q \in O(n) \)

Appendix A
\( \mathcal{M}_{b,|s|\text{pr}}(n) \): set of all rational, causal, stable and (strictly) positive real \( s \times s \) transfer functions \( \Phi(z) \) of order \( n \), where \( \Phi(0) = D \) and thus \((D + D') \geq (>) 0 \)
\( \mathcal{M}_{b,|s|}\text{pr}(n) \): set of all rational, proper and stable \( s \times s \) c.t. transfer functions of order \( n \) where \( k_c(\infty) = D, D \) arbitrary
\( \mathcal{M}^+_b(n) \): subset of \( \mathcal{M}_{b,|s|\text{pr}}(n) \) where \((D + D') > 0 \)
\( \mathcal{M}_{b,|s|}\text{pr}(n) \): subset of \( \mathcal{M}_{b}(n) \) where \( k \) is (strictly) positive real and thus \((D + D') \geq (>) 0 \)
\( \mathcal{M}_{b,|s|}\text{mp}(n) \): set of all rational, proper, stable and (strictly) minimum phase \( s \times s \) c.t. transfer functions of order \( n \) where \( k_c(\infty) = E, E \) arbitrary
Bibliography


Curriculum vitae

NAME UND ANSCHRIFT:
Thomas Ribarits
Kollonitschgasse 14/3/11
2700 Wiener Neustadt

PERSÖNLICHE DATEN:
Geburtsdatum 5. März 1975
Geburtsort Wiener Neustadt
Nationalität Österreich
Familienstand ledig

AUSBILDUNG:
1981 - 1985 Volksschule in Wiener Neustadt
1985 - 1993 Bundesrealgymnasium in Wiener Neustadt
Reifeprüfung mit Auszeichnung
1994 - 1999 Technische Universität Wien: Studium der Technischen Mathematik
Studienzweig: Wirtschaftsmathematik
Diplomarbeit: *Dynamic Principal Component Models*
2000/01 - 07 University of Cambridge, Großbritannien: Forschungsaufenthalt
2000 - 2002 Technische Universität Wien: Doktoratsstudium; Dissertation:
*The Role of Parametrizations in Identification of Linear Dynamic Systems*
2002/06 - 07 Vrije Universiteit Brüssel, Belgien: Forschungsaufenthalt
1993 - 1994 Präsenzdienst an der Theresianischen Militärakademie Wiener Neustadt