DIPLOMARBEIT

Estimation of Large Covariance Matrices using Results on Large Deviations

Ausgeführt am Institut für
Wirtschaftsmathematik
der Technischen Universität Wien

unter der Anleitung von
Em.O.Univ.Prof. Dipl.-Ing. Dr.techn. Manfred Deistler

durch
Mirsad Tulic

Februar 2010 Unterschrift (Student)

Contents

1 Introduction

One might wonder about the title of this thesis, since it is always possible to estimate the population covariance matrix from samples of multivariate data with the sample covariance matrix

$$Q = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T, \tag{1.1}$$

where we assume that we observe $\mathbf{X}_1, \ldots, \mathbf{X}_n$, i.i.d. p-variate random variables with mean $\mathbf{0}$ and population covariance matrix Σ_p.
If we assume $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ to be multivariate normal, the maximum likelihood estimator of Σ_p is

$$\hat{\Sigma}_p = Q\frac{n-1}{n}.$$

If p is fixed, the MLE of Σ_p behaves asymptotically optimally, converging to Σ_p at rate $n^{-\frac{1}{2}}$. In recent years data with a high dimension relative to the sample size, i.e. p much larger than n, have emerged from many fields of application, such as gene expression arrays, fMRI (functional magnetic resonance imaging) data, numerical weather forecasting and in portfolio optimization via Markowitz's rule for portfolio selection. In the seminal paper [20] written by Marcenko and Pastur in 1967 the authors showed that for a sample of size n from a p-variate Gaussian distribution with population covariance matrix $\Sigma_p = I$ (the identity matrix) the empirical distribution of the eigenvalues of the sample covariance matrix $\hat{\Sigma}_p$ is supported on the interval $((1-\sqrt{c})^2, (1+\sqrt{c})^2)$ for $p/n \to c > 0$. They studied the limiting distribution of the empirical distribution of the eigenvalues of random unitary and hermitian operators. The behaviour is observed as $n \to \infty$, while $p = p(n)$ with $p/n \to c > 0$ and under the notion of empirical distribution of eigenvalues they mean a function $\nu(\lambda; B_n(p))$ giving the ratio of the number of eigenvalues of a specific self-adjoint operator $B_n(p)$ lying to the left of λ to the dimension of the space. This specific operator is of the form

$$B_n(p) = A_n + \sum_{i=1}^{p}\tau_i q^{(i)}(\cdot, q^{(i)}), \tag{1.2}$$

where A_n is a nonrandom self-adjoint operator, the τ_i are independent, identically distributed real random variables and the $q^{(i)}$ are mutually independent random vectors in a n-dimensional unitary space H_n, independent also of the τ_i (see [20]). Here a unitary space U is a complex vector space equipped with

a positive definite Hermitian form $(\cdot,\cdot) : U \times U \to \mathbb{C}$, which serves as the inner product on U. They showed that $\nu(\lambda; B_n(p))$ converges to a nonrandom function $\nu(\lambda; c)$ under some assumptions. This thesis is built on the results presented in Bickel, Levina [2]. Its purpose is to generalize and illuminate some results of that article. The work of Bickel, Levina should have a fairly strong influence in many application fields where it is of significant importance to have a reliable estimator of the population covariance matrix for both p and n large enough, especially in the field of portfolio optimization. There are some published works in the last mentioned field. Ledoit and Wolf ([17], [18]) considered a shrinkage estimator of the population covariance matrix (the notion of shrinkage was first introduced by C. Stein in 1955). They compared a convex, linear combination $\delta F + (1-\delta)S$ (where $0 < \delta < 1$) of the sample covariance matrix S and a highly structured estimator F with the sample covariance matrix (highly structured means that it involves only a small number of free parameters), another shrinkage estimator with a constant correlation model as the shrinkage target F and a multi-factor model based on statistical factors. Their highly structured estimator is the single-factor matrix of Sharpe. After having read [17] and [18] nobody should use the sample covariance matrix for the purpose of portfolio optimization anymore, they concluded. That inference is based on the results they obtained by comparing the out-of-sample performance of the mentioned estimators. The shrinkage estimator tends to pull the most extreme coefficients towards more central values and thus reduces estimation error where it matters most. The shrinkage estimator with a constant correlation model yields in all scenarios the highest (average) information ratio, the lowest (average) standard deviation of excess return and yields in most scenarios (different scenarios match with different number of stocks considered) the highest (average) mean excess return. The regularized estimator in [2] may be used in their model as the shrinkage target and/or may be compared solely with the best performer of [17] and [18]. This is a topic for further research. My contribution to the topic is on the one hand to identify some flaws of [2] and on the other to give an indication how the results of [2] could eventually be generalized to stationary processes instead of only assuming an identically and independently distributed Gaussian process. In the last chapter of this thesis I will present some theorems with different assumptions which all allow to use a bound for large deviations for the standardized sum of non independent random variables if the cumulants of that standardized sum can be appropriately bounded. This bound is presented in the Lemma 4.2.

Before we introduce the model and come to the main results that can be deducted from our assumptions, we will need to deal with an important concept in statistics, important especially when it is about proving a limiting

distribution for certain random variables or vectors. It is the concept of cumulants which is a modification of the moment problem or to more precisely, a modification of a theorem that allows us to prove convergence in distribution of a sequence of distribution functions from the convergence of the corresponding sequence of moments (under certain conditions).

2 Cumulants

The characteristic and moment generating function of a random variable is familiar, but we will mention it again for the purpose of displaying the analogy to cumulants. This section uses the notation of the book of Saulis and Statulevicius [24], furthermore we will cite some results from the book that are essential to this thesis. The characteristic function of a random variable (r.v.) ξ is given by

$$f_\xi(t) = \mathbb{E}e^{it\xi} = \int_\Omega e^{itx} dF_\xi(x). \tag{2.1}$$

If ξ has absolute moments up to order k, then its characteristic function has k-th order derivatives and

$$\alpha_\nu = \frac{1}{i^\nu} \frac{d^\nu}{dt^\nu} f_\xi(t)\big|_{t=0} \tag{2.2}$$

and

$$f_\xi(t) = \sum_{\nu=0}^{k} \frac{i^\nu \alpha_\nu}{\nu!} t^\nu + o(|t|^k). \tag{2.3}$$

Remark It is actually not necessary to distinguish between absolute and simple moments. For later theorems we will need the existence of absolute moments, because of the effort to find upper bounds for expressions of centered moments, among others. In the book of Bisgaard and Sasvari, [3], they proof a theorem where it is enough to assume existence of the moment of a r.v. for some $k \geq 1$ to obtain the k times differentiability of the characteristic functions and to obtain the boundedness and uniform continuity of $f^{(k)}$ as well as the identity $f^{(k)}(0) = i^k \alpha_k$.

Remark But the existence of the k-th derivative of f does not imply the existence of the moment α_k in general. However, it can be shown that if for some integer k the real part of the characteristic function f_ξ is $2k$ times differentiable at 0, then the moment α_{2k} of ξ exists.

If there exists an l such that the l-th absolute moment $\beta_l = \mathbb{E}\left|\xi^l\right|$ exists, then for sufficiently small t, $\log f_\xi(t)$ can be written as

$$\log f_\xi(t) = \sum_{k=1}^{l} \frac{1}{k!}\gamma_k (it)^k + o(|t|^l), \tag{2.4}$$

$$\gamma_k = \frac{1}{i^k}\frac{d^k}{dt^k}(\log f_\xi(t))\big|_{t=0}. \tag{2.5}$$

The γ_k is called the k-th cumulant of the random variable ξ. The existence of β_k implies the existence of γ_k. This follows from the fact that

$$\left|\gamma_k\right| = \left|\frac{1}{i^k}\frac{d^k}{dt^k}(\log f_\xi(t))\big|_{t=0}\right| = \left|\frac{d^k}{dt^k}\log\int_\Omega e^{itx}dF_\xi(x)\right| \leq ?$$

Now we use the well-known Taylor-series expansion of $\log(1+z)$ for $|z|<1$:

$$\log(1+z) = \sum_{s=1}^{\infty}\frac{(-1)^{s+1}}{s}z^s \quad \text{for} \quad |z|<1, \tag{2.6}$$

to obtain the formal equality

$$\log f_\xi(t) = \log\left(1+\sum_{\nu=1}^{\infty}\frac{i^\nu\alpha_\nu}{\nu!}t^\nu\right) = \sum_{s=1}^{\infty}\frac{(-1)^{s+1}}{s}\left(\sum_{\nu=1}^{\infty}\frac{i^\nu\alpha_\nu}{\nu!}t^\nu\right)^s. \tag{2.7}$$

Using (2.4) we can write $\log f_\xi(t)$ formally as

$$\log f_\xi(t) = \sum_{k=1}^{\infty}\frac{1}{k!}\gamma_k(it)^k. \tag{2.8}$$

By equating the terms in sums in (2.7) and (2.8) we obtain identities where the cumulants γ_k are expressed via the moments α_i:

$$\gamma_k = \sum_{\nu=1}^{k}\frac{(-1)^{\nu-1}}{\nu}\sum_{k_1+k_2+\cdots+k_\nu=k}\frac{k!}{k_1!\cdots k_\nu!}\alpha_{k_1}\cdots\alpha_{k_\nu}, \tag{2.9}$$

where under the summation $k_1+k_2+\cdots+k_\nu = k$ we mean all possible partitions of k in ν summands, where the order also matters. From (2.9) we obtain the first few identities between cumulants and moments:

$$\gamma_1 = \alpha_1, \quad \gamma_2 = \alpha_2 - \alpha_1^2,$$

$$\gamma_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3,$$

$$\gamma_4 = \alpha_4 - 4\alpha_3\alpha_1 - 3\alpha_2^2 + 12\alpha_2\alpha_1^2 - 6\alpha_1^4$$
$$\vdots$$

Let us from now on denote the k-th cumulant γ_k of the random variable ξ by $\Gamma_k(\xi)$. The concept of cumulants reveals its first advantage when dealing with independent random variables $\xi_1, \xi_2, \ldots, \xi_n$. Let $S_n := \xi_1 + \cdots + \xi_n$, then for the k-th cumulant of S_n

$$\Gamma_k(S_n) = \sum_{j=1}^{n} \Gamma_j(\xi_j)$$

holds. Thus, the k-th cumulant of the sum of n independent random variables conveniently becomes the sum of the k-th cumulant of the separate random variables. This can be seen from the fact that the characteristic function of S_n, f_{S_n} in the case of independent random variables ξ_i can be represented by the product of separate characteristic functions:

$$f_{S_n}(t) = \prod_{j=1}^{n} f_{\xi_j}(t),$$

The last relation, in turn, is easily seen just from the fact that the distribution function of a sum of independent random variables is the convolution of the marginal distribution functions. Namely, let ξ and η be independent random variables with the distribution functions F_ξ and F_η. Then the joint distribution function is calculated as follows:

$$F_{\xi+\eta}(x) = \int_{\mathbb{R}} F_\xi(x-y)\, dF_\eta(y) = \int_{\mathbb{R}} F_\eta(x-y)\, dF_\xi(y),$$

where $(\Omega, \mathfrak{A}, \mathbb{P})$ is an appropriate probability space. The last equation can be written in short $F_{\xi+\eta} = F_\xi * F_\eta$. Here we emphasize two important facts; first, the Fourier transform (and also the inverse Fourier transform) of the convolution of two functions is the multiplication of the Fourier transforms (inverse Fourier transforms) of the separate functions and second, the characteristic function of a random variable ξ is the inverse Fourier transform of the density function of ξ (if the density function exists, if not, it is the generalized inverse Fourier transform of F_ξ).
The second advantage of the cumulant concept is displayed when dealing with random variables for which we want to show convergence to a normal distribution for all points of continuity of the distribution function. Let η be a $\mathrm{N}(\mu, \sigma^2)$ distributed random variable, then as it is can be easily computed and as it is well known, the characteristic function of η is

$$f_\eta(t) = \exp(-i\mu t - \frac{1}{2}\sigma^2 t^2),$$

so the central moments are equal to

$$\begin{aligned} \mu_{2k+1} &= 0, && \forall k \geq 0, \\ \mu_{2k} &= 1 \cdot 3 \cdot \ldots \cdot (2k-1)\sigma^{2k} && \forall k \geq 1. \end{aligned}$$

The cumulants of η are

$$\begin{aligned} \Gamma_1(\eta) &= \mu, & \Gamma_2(\eta) &= \sigma^2, \\ \Gamma_k(\eta) &= 0, & & \forall k \geq 3. \end{aligned}$$

Here we see the advantage of the cumulants over the moments when you are about to show convergence to a normal distribution. It is much easier to show that a sequence converges to zero than converging to certain constants. Here for a normal distributed r.v. the cumulants of higher order than k are equal to zero, as is easily seen from the form of the characteristic function of η.

2.1 The moment problem

The following theorem formulates precisely the idea that was mentioned in the end of the introduction. The theorem lets us sense how the concept of cumulants can be used in the way we have described in the previous section. We omit the proof, as though it can be found in many books on advanced probability theory.

Theorem 2.1 *Suppose there is distribution function F, uniquely determined by its moments $\{m_n\}_{n=1}^{\infty}$ and let $\{F_k\}_k$ be a sequence of distribution functions each of which has all its moments finite:*

$$m_n^{(k)} = \int_\Omega x^n \, dF_k < \infty \qquad \forall n \geq 1.$$

Further suppose that $\forall n \geq 1$

$$\lim_{k\to\infty} m_n^{(k)} = m_n.$$

Then $F_k \xrightarrow{d} F$, i.e. $\lim_{k\to\infty} F_k(x) = F(x)$ for all continuity points x of F.

In the vast probability and statistics literature it is also said that F_k converges in distribution to F. Note that since cumulants of a r.v. can be expressed through moments, as we have seen above, theorem 2.1 can be modified to

Corollary 2.2 *Let a r.v. ξ_k depend on a parameter k and there exist all absolute moments of ξ_k $\mathbb{E}\,|\xi_k^n| < \infty$ for all $n \geq 1$. If*

$$\lim_{k\to\infty} \Gamma_n(\xi_k) = \Gamma_n(\xi)$$

for every n, where ξ is a random variable whose distribution function is uniquely determined, then

$$\xi_k \xrightarrow{d} \xi \qquad as \;\; k \to \infty,$$

i.e. the r.v. ξ_k converges to the r.v. ξ in distribution.

Remark The cumulants are also called *semi-invariants* in the literature on probability theory, as in [22].

Remark In order for a probability distribution F to be uniquely determined by its moments, as it is required by the preceding theorem, the *Carleman's condition* is sufficient, namely

$$\sum_{n=1}^{\infty} (\alpha_{2n})^{-\frac{1}{2n}} = \infty.$$

In other words, it is sufficient that the sum of the even moments has an appropriate decay.

Remark Otherwise there are some known conditions under which a distribution is definitely not determined by its moments. Among those conditions one of them is: If F satisfies the condition

$$\int_{\Omega} \frac{\log(F'(x))}{1+x^2}\, dx > -\infty,$$

where F' is the Radon-Nikodym derivative of F.

2.2 Mixed cumulants of the random vector X

In the investigation of random processes we shall make use of finite-dimensional distributions of a process, i.e. the distribution of a random vector $\mathbf{X} = (X_{t_1}, \ldots, X_{t_k})$, $(t_1, \ldots, t_k) \in T \subset \mathbb{R}^k$. If $\mathbb{E}|X_t^m| < \infty$, $t \in T$, then for all $k \leq m$ the functions

$$m_X(t_1 \ldots, t_k) := \mathbb{E} X_{t_1} \ldots X_{t_k}$$

are well defined. The function $m_X(t_1 \dots, t_k)$ is called the $k-$th moment function or the simple moment of the k-th order of the random process $\mathbf{X}_t$. Let

$$f_X(u_1, \dots, u_k) := \mathbb{E} \exp\{i\langle u, \mathbf{X}\rangle\}$$

be the characteristic function of the random vector $\mathbf{X}$, where $\langle \cdot, \cdot \rangle$ denotes the euclidean scalar product of two vectors from $\mathbb{R}^k$. Analogously to the one-dimensional case, if $\mathbb{E}|X_t^m| < \infty$, then for all k and $\nu = (\nu_1, \dots, \nu_k)$, where $\nu_i \geq 0$, $k|\nu| \leq m$ and $|\nu| := |\nu_1| + \dots + |\nu_k|$, there exist mixed cumulants of the random vector $\mathbf{X}$

$$\Gamma_\nu(\mathbf{X}) := \frac{1}{|\nu|} \frac{\partial^{\nu_1 + \dots + \nu_k}}{\partial u_1^{\nu_1} \dots \partial u_k^{\nu_k}} \big(\ln f_X(u_1, \dots, u_k) \big) \Big|_{u_1 = 0, \dots, u_k = 0}.$$

Sometimes we write shortly $\Gamma_\nu(\mathbf{X})$ instead of $\Gamma_\nu(X_{t_1}, \dots, X_{t_k})$.
If $\nu = (1, \dots, 1)$, then the corresponding cumulant $\Gamma_\nu(X_{t_1}, \dots, X_{t_k})$ will be denoted by $\Gamma(\mathbf{X})$ or $\Gamma(X_{t_1}, \dots, X_{t_k})$. When we deal with $S_n = \sum_{t=1}^n X_t$, then it follows from the definition that

$$\Gamma_k(S_n) = \sum_{1 \leq t_1, \dots, t_k \leq n} \Gamma(X_{t_1}, \dots, X_{t_k}).$$

We will make use of the following notation: if $\nu = (\nu_1, \dots, \nu_k)$ is an integer nonnegative vector and $a = (a_1, \dots, a_k)$ is a real vector, then

$$a^\nu := a_1^{\nu_1} \dots a_k^{\nu_k}, \quad \nu! := \nu_1! \dots \nu_k!, \quad |\nu| := \nu_1 + \dots + \nu_k.$$

Furthermore, denote

$$\mathbb{E}_\nu(\mathbf{X}) = \mathbb{E} X_{t_1}^{\nu_1} \dots X_{t_k}^{\nu_k}.$$

If $\mathbb{E}|X_t^m| < \infty$, $t = (t_1, \dots, t_k)$, for some integers $m \geq 1$, then the function $f_{\mathbf{X}}(u)$ can be expanded into a Taylor series as follows

$$f_{\mathbf{X}}(u) = \sum_{|\nu| \leq m} \frac{i^{|\nu|}}{\nu!} \mathbb{E}_\nu(\mathbf{X}) u^\nu + o(|u|^m),$$

where $\sum_{|\nu| \leq m}$ is taken over all nonnegative collections of $(\nu_1, \dots, \nu_k)$ such that $|\nu| \leq m$. Similarly, as we have obtained in the one-dimensional case, we have

$$\log f_{\mathbf{X}}(u) = \sum_{|\nu| \leq m} \frac{i^{|\nu|}}{\nu!} \Gamma_\nu(\mathbf{X}) u^\nu + o(|u|^m)$$

in the neighbourhood $|u| < \delta$, $\delta > 0$. It is possible to derive formulas, connecting $\mathbb{E}_\nu(\mathbf{X})$ and $\Gamma_\nu(\mathbf{X})$, analogously to the one-dimensional case:

$$\mathbb{E}_\nu(\mathbf{X}) = \sum_{\lambda^{(1)} + \dots + \lambda^{(q)} = \nu} \frac{1}{q!} \frac{\nu!}{\lambda^{(1)}! \dots \lambda^{(q)}!} \prod_{p=1}^q \Gamma_{\lambda^{(p)}}(\mathbf{X}),$$

$$\Gamma_\nu(\mathbf{X}) = \sum_{\lambda^{(1)}+\dots+\lambda^{(q)}=\nu} \frac{(-1)^{q-1}}{q} \frac{\nu!}{\lambda^{(1)}!\dots\lambda^{(q)}!} \prod_{p=1}^{q} \mathbb{E}_{\lambda^{(p)}}(\mathbf{X}),$$

where $\sum_{\lambda^{(1)}+\dots+\lambda^{(q)}=\nu}$ is shortly denoted for the summation over all ordered collections of integer nonnegative vectors $\lambda^{(p)}$, $|\lambda^{(p)}| > 0$, which equal ν in sum.

Let $I = \{t_1, \dots, t_k\}$ be a set of indices of vector $\mathbf{X}$. An unordered collection of disjoint nonempty sets I_p, such that $\bigcup_p I_p = I$, is called a partition of I. In this notation the last two expressions can be rewritten as

$$\mathbb{E}X_{t_1}\dots X_{t_k} = \sum_{\bigcup_{p=1}^{q} I_p = I} (q-1)! \prod_{p=1}^{q} \Gamma(X_{I_p}),$$

$$\Gamma(X_{t_1}\dots X_{t_k}) = \sum_{\bigcup_{p=1}^{\nu} I_p = I} (-1)^{q-1}(q-1)! \prod_{p=1}^{q} \mathbb{E}(X_{I_p}).$$

The last four formulas were first introduced and proved by A.N. Shiryaev and V.P. Leonov, [19].

Let the letter "c" as subscript of a random variable ξ denote the centered random variable:

$$\xi_c := \xi - \mathbb{E}\xi.$$

In the investigation and estimation of cumulants $\Gamma_k(S_n)$, where S_n is the sum of n random variables ξ_i, it will be more convenient for us to express $\Gamma(X_{t_1}, \dots, X_{t_k})$ through centered moments.

$$\mathbb{E}_c(X_I) := \mathbb{E}\big\{X_{t_1}(X_{t_2}\dots(X_{t_{m-1}}(X_{t_m})_c)_c)_c\big\},$$

where $X_I := (X_{t_1}, \dots, X_{t_m})$ for $(t_1, \dots, t_m)$ is a partition of I.

Sometimes the notation $E_c(X_I)$ will be replaced by $\mathbb{E}_c X_{t_1}\dots X_{t_m}$.

$\Gamma(X_{t_1}, \dots, X_{t_k})$ does not change after any kind of permutation of $t_1, \dots, t_k$, we can w.l.o.g. assume $t_1 \le t_2 \le \dots \le t_k$ in $\mathbb{E}_c X_{t_1}\dots X_{t_k}$. The next formula gives a relationship between the centered moments and the moments of a random process:

$$\mathbb{E}_c X_{t_1}\dots X_{t_k} = \sum_{\nu=1}^{k} (-1)^{\nu-1} \sum_{\bigcup_{p=1}^{\nu} I_p = I}^{*} \prod_{p=1}^{\nu} \mathbb{E}(X_{I_p}),$$

where $\mathbb{E}(X_{I_p}) = \mathbb{E}X_{t_1^{(p)}}\dots X_{t_{k_p}^{(p)}}$ and the summation $\sum^{*}_{\bigcup_{p=1}^{\nu} I_p = I}$ is taken over partitionings $\{I_1, \dots, I_\nu\}$ of the set I such that $\max I_p \le \min I_{p+1}$, $1 \le p \le$

$\nu - 1$.
Note that in the case of independent r.v. $X_{t_1}, \ldots, X_{t_k}$, $\mathbb{E}_c X_{t_1} \ldots X_{t_k}$ does not vanish only if $t_1 = t_2 = \ldots = t_k$. The same is true for $\Gamma(X_{t_1}, \ldots, X_{t_k})$.
The following lemma will be presented without a proof, since its proof would exceed the scope of this thesis. It is given in the Appendix of [24].

Lemma 2.3 *The representation*

$$\Gamma(X_{t_1}, \ldots, X_{t_k}) = \sum_{\nu=1}^{k} (-1)^{\nu-1} \sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1, \ldots, I_\nu) \prod_{p=1}^{\nu} \mathbb{E}_c X_{I_p} \tag{2.10}$$

holds. The sum $\sum_{\bigcup_{p=1}^{\nu} I_p = I}$ *denotes the summation over all* ν*-block partitionings* $\{I_1, \ldots, I_k\}$ *of the set* I*. The integers* $N_\nu(I_1, \ldots, I_\nu)$ *depend on* $\{I_1, \ldots, I_\nu\}$ *only, and if* $N_\nu(I_1, \ldots, I_\nu) > 0$*, then*

$$\sum_{p=1}^{\nu} \max_{t_i, t_j \in I_p} (t_j - t_i) \geq \max_{1 \leq i,j \leq k} (t_j - t_i). \tag{2.11}$$

Furthermore,

$$0 \leq N_\nu(I_1, \ldots, I_\nu) \leq (\nu - 1)! \tag{2.12}$$

is valid.

The structure of $N_\nu(I_1, \ldots, I_\nu)$ will be explained in more details in chapter 5.2, to be more precisely, in the proof of Theorem 5.7. Another explicit expression of it, which requires lots of time to get through it, is given in the Appendix of [24]. Let us give some of the first mixed cumulants using the above lemma:

$$\Gamma(X_t) = \mathbb{E}_c X_t = \mathbb{E} X_t, \quad \Gamma(X_s, X_t) = \mathbb{E}_c X_s X_t,$$
$$\Gamma(X_{t_1}, X_{t_2}, X_{t_3}) = \mathbb{E}_c X_{t_1} X_{t_2} X_{t_3} - \mathbb{E}_c X_{t_2} \mathbb{E}_c X_{t_1} X_{t_3},$$

$$\begin{aligned}\Gamma(X_{t_1}, X_{t_2}, X_{t_3}, X_{t_4}) = {} & \mathbb{E}_c X_{t_1} X_{t_2} X_{t_3} X_{t_4} - \mathbb{E} X_{t_2} \mathbb{E}_c X_{t_1} X_{t_3} X_{t_4} - \\ & - \mathbb{E} X_{t_3} \mathbb{E}_c X_{t_1} X_{t_2} X_{t_4} - \mathbb{E}_c X_{t_1} X_{t_3} \mathbb{E}_c X_{t_2} X_{t_4} - \\ & - \mathbb{E}_c X_{t_1} X_{t_4} \mathbb{E}_c X_{t_2} X_{t_3} + \mathbb{E} X_{t_2} \mathbb{E} X_{t_3} \mathbb{E}_c X_{t_1} X_{t_4}, \ldots .\end{aligned}$$

3 The statistical model and regularized estimates

As before, we assume that we observe $\mathbf{X}_1, \ldots, \mathbf{X}_n$ identical and independent distributed p-variate random variables with mean $\mathbf{0}$ and population covariance matrix Σ_p. Let $\mathbf{X}_i$ denote the vector $(X_{i1}, \ldots, X_{ip})$ and let $\mathbf{X}$ denote a

$p \times n$ matrix of n observations on a system of p random variables. If $\mathbf{X}_i$ have mean vector $\mu \neq \mathbf{0}$, then it is clearly seen why the sample covariance matrix is singular for $p > n$, even if the population covariance matrix is regular: the sample mean vector is given by

$$\hat{\mu} = \frac{1}{n}\mathbf{X}\mathbf{1},$$

where $\mathbf{1}$ denotes a $n \times 1$ vector, where every component is 1 and the sample covariance matrix by

$$Q = \frac{1}{n-1} \underbrace{\mathbf{X}}_{p\times n} \underbrace{\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)}_{n\times n} \underbrace{\mathbf{X}^T}_{n\times p}.$$

Whether the scaling factor is $\frac{1}{n}$ or $\frac{1}{n-1}$ does not is irrelevant for the fact that Q is singular for $p > n$. Indeed, the matrix Q can have at most the rank of the matrix $\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)$, yet this matrix is idempotent, i.e.

$$\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right).$$

It can easily be shown that for idempotent matrices the rank equals the trace. Thus,

$$\operatorname{tr}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) = \sum_{i=1}^{n} 1 - \sum_{i=1}^{n} \frac{1}{n} = n - 1.$$

Therefore when the dimension p exceeds $n-1$, the sample covariance matrix is singular, independently whether the true covariance matrix is regular or not. The same is valid also for a zero mean random vector, but can not be seen from decomposition from above. This time it can be concluded from the fact that for $p > n$ the last $p - n$ eigenvalues of Q (and also of $\hat{\Sigma}_p$) are zero. This is seen as follows: the matrix $\mathbf{X}^T : \mathbb{R}^p \to \mathbb{R}^n$ has rank at most equal to n, due to basic linear algebra. After applying another map onto $\mathbf{X}^T$, namely $\mathbf{X}$ the rank can not increase. Thus, $\operatorname{rank}(\frac{1}{n}\mathbf{X}\mathbf{X}^T) \leq n$. Bickel and Levina suggest two types of regularization to avoid the above discussed flaws of the sample covariance matrix in [2].

3.1 Banding the sample covariance matrix

Define the banding operator for any matrix $M = [m_{ij}]_{p\times p}$ and any $0 \leq k < p$ as follows:

$$B_k(M) = \left[m_{ij}\mathbf{1}_{[|i-j|\leq k]}\right]. \tag{3.1}$$

We apply this operator to the sample covariance matrix and denote it $\hat{\Sigma}_{k,p} := \hat{\Sigma}_k := B_k(\hat{\Sigma}_p)$. This kind of regularization is ideal when you regularize a sample covariance matrix of an finite inhomogeneous moving average process $Y_t = \sum_{j=1}^{k} a_{j,t-j}\varepsilon_j$ and ε_j are IID$(0, \sigma^2)$, where for the true covariance matrix $|i-j| > k \Rightarrow \sigma_{ij} = 0$ holds. The covariance matrix has two essential features that make it easier to handle than arbitrary matrices, namely it is symmetric and positive definite. Symmetry will clearly be preserved by the banding operator, but as one's intuition may tell very quickly, the positive definiteness will not. Consider the following symmetric and positive definite matrix Σ_3 (thus, there exist a unique stochastic process for which Σ_3 is a covariance matrix), where the 3 represents the dimension:

$$\Sigma_3 = \begin{pmatrix} 10.8 & -3.1 & -2.3 \\ -3.1 & 1.4 & 2 \\ -2.3 & 2 & 4.2 \end{pmatrix}.$$

The main minors of Σ_3 are $H_1 = 10.8$, $H_2 = 5.51$ and $H_3 = 1.056$, therefore Σ_3 is positive definite. Let us apply the banding operator on Σ_3 with $k = 1$:

$$\Sigma_{1,3} = \begin{pmatrix} 10.8 & -3.1 & 0 \\ -3.1 & 1.4 & 2 \\ 0 & 2 & 4.2 \end{pmatrix}.$$

This matrix is not positive definite anymore (H_1 and H_2 stay the same, but $H_3 = -20.058$). The inverse of the banded matrix $\Sigma_{1,3}$ is not banded:

$$\Sigma_{1,3}^{-1} = \begin{pmatrix} -0.094 & -0.649 & 0.309 \\ -0.649 & -2.261 & 1.077 \\ 0.309 & 1.077 & -0.275 \end{pmatrix}.$$

3.2 Banding the inverse

There is an estimator for the inverse of the covariance matrix suggested by Wu and Pourahmadi [28] and Huang et al. [13]. It is based on the Cholesky decomposition of the inverse and we will show its construction in what follows. Consider a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and the collection C of all random variables X_i defined on Ω with finite variance. Introduce an inner product on this vector space (that C is a vector space is easily checked) by defining for any two elements $X,\ Y \in C$

$$(X, Y) = \mathbb{E}(XY).$$

Since from $(X, X) = 0$ it does not follow $X(\omega) = 0$ for all ω, but only that $\mathbb{P}(X = 0) = 1$, we need to consider the equivalence classes of C by saying that

X and Y are equivalent if $\mathbb{P}(X = Y) = 1$. The collection of these equivalence classes of C is the space $L_2(\Omega, \mathfrak{A}, \mathbb{P})$. In the sequel $\hat{X}_j$ denotes the $L_2(\Omega, \mathfrak{A}, \mathbb{P})$ projection of X_j on the linear span of $X_1, \ldots, X_{j-1}$, if not otherwise specified. Let us suppose that we have $\mathbf{X} = (X_1, \ldots, X_p)^T$ a random vector distributed $\mathrm{N}(\mathbf{0}, \Sigma_p)$ and denote $\Sigma_p = [\sigma_{ij}]$. We can model every X_j as

$$X_j = \sum_{t=1}^{j-1} a_{jt} X_t + \varepsilon_j,$$

where ε_j is the error term independent and identically distributed with mean zero and variance σ^2, also independent of the preceding X_t for $0 < t \leq j-1$. Then we can conduct a regression of X_j on X_t for all $t > 0$ up to $j-1$. Let $\mathbf{Z}_j := (X_1, \ldots, X_{j-1})^T$ and $\mathbf{a}_j = (a_{j1}, \ldots, a_{j,j-1})^T$, then we can write

$$\hat{X}_j = \sum_{t=1}^{j-1} a_{jt} X_t = \mathbf{Z}_j^T \mathbf{a}_j. \tag{3.2}$$

For $j = 1$ let $\hat{X}_1 = 0$. Every $\mathbf{a}_j^T$ can be computed as

$$\mathbf{a}_j = \left(\mathbb{V}(\mathbf{Z}_j)\right)^{-1} \mathbb{C}\left(X_j, \mathbf{Z}_j\right). \tag{3.3}$$

Now construct a lower triangular matrix A with zeros on the diagonal containing the coefficients of $\mathbf{a}_j$ arranged in rows and let $\varepsilon_j = X_j - \hat{X}_j$, $d_j^2 = \mathbb{V}(\varepsilon_j)$ and $D = \mathrm{diag}(d_1^2, \ldots, d_p^2)$. According to regression theory, the residuals ε_j are independent. Let us now apply the covariance operator to the identity $\varepsilon = X - AX = (I - A)X$

$$\mathbb{C}(\varepsilon) = (I - A)\mathbb{C}(X)(I - A)^T,$$

which can also be written as

$$\Leftrightarrow D = (I - A)\Sigma_p(I - A)^T.$$

Thus, we obtain the modified Cholesky decomposition

$$\Sigma_p = (I - A)^{-1} D [(I - A)^{-1}]^T, \tag{3.4}$$

$$\Sigma_p^{-1} = (I - A)^T D^{-1} (I - A). \tag{3.5}$$

Note that $(I - A)$ indeed is invertible, since it is lower triangular with 1 as the entry on every diagonal element (A itself has zeros on the diagonal). In order to apply the banding operator, equation (3.2) suggests itself to approximate Σ_p by taking $\mathbf{Z}_j^{(k)} = \left(X_{\max(j-k,1)}, \ldots, X_{j-1}\right)$, thus, obtaining A_k and D_k for

$k < p$ and by taking A_k and D_k and inserting them into (3.4) and (3.5) for A and D, where A_k are k-banded lower triangular matrices containing $\mathbf{a}_j^{(k)}$ and $D_k = \text{diag}(d_{j,k}^2)$ are diagonal matrices with the new residual variances. In other words, we regress each X_j on its closest k predecessors only. The next step would be to replace $\mathbf{Z}_j$ by $\mathbf{Z}_j^{(k)}$ and thus, obtain $\mathbf{a}_j^{(k)}$ instead of $\mathbf{a}_j$. For a given sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the estimates of A_k and D_k are obvious. Equation (3.3) becomes the ordinary least squares estimate of $\mathbf{a}_j$:

$$\hat{\mathbf{a}}_j = ((\mathbf{Z}_j^{(k)})^T \mathbf{Z}_j)^{-1} (\mathbf{Z}_j^{(k)})^T X_j$$

and D_k become the corresponding residual variances. Let us denote these sample estimates with $\tilde{A}_k = [\tilde{a}_{jt}^{(k)}]$ and $\tilde{D}_k = \text{diag}(\tilde{d}_{j,k}^2)$. By inserting them into (3.4) and (3.5) we obtain the estimates for Σ_p^{-1} and Σ_p and we denote them by $\tilde{\Sigma}_{k,p}^{-1}$ and $\tilde{\Sigma}_{k,p}$ respectively. Since $\tilde{A}_k$ is a k-banded lower triangular matrix, $\tilde{\Sigma}_{k,p}^{-1}$ is k-banded and non-negative definite. Its inverse $\tilde{\Sigma}_k$ is in general not banded as we have indicated above with the counterexample and is different from $\hat{\Sigma}$. Note also that $\tilde{\Sigma}_k^{-1}$ is not the same as $B_k(\hat{\Sigma}^{-1})$, which is not well-defined for $p > n$.

4 The main theorems

Now we examine the asymptotic behaviour (asymptotic in the sense when both p and n tend to infinity and the ratio p/n tends to $c \in (0, 1)$) of the banded sample covariance matrix and we will introduce another concept of estimating the population covariance matrix, which is a generalization of banding and which preserves the positive definiteness, namely the concept of tapering a matrix. The following results show convergence of estimators in the matrix L_2 norm, $\|M\| := \sup\{\|M\mathbf{x}\| : \|\mathbf{x}\| = 1\} = \lambda_{\max}^{1/2}\left(M^T M\right)$, which for symmetric matrices reduces to $\|M\| = \max_i |\lambda_i(M)|$, see [12]. It will be shown that the convergence of certain estimators is uniform on sets of covariance matrices which we introduce in the sequel.
We define a set of covariance matrices, Σ_p, which we will refer to as *well-conditioned covariance matrices*, as follows:

$$\left\{\Sigma_p : \ 0 < \varepsilon \leq \lambda_{\min}(\Sigma_p) \leq \lambda_{\max}(\Sigma_p) \leq 1/\varepsilon < \infty\right\},$$

where $\lambda_{\min}(\Sigma_p)$, $\lambda_{\max}(\Sigma_p)$ are the minimum and maximum eigenvalues of Σ_p and ε is independent of p. Why are we interested in especially such covariance matrices? For numerical computations that are of essential importance in fields like portfolio optimization or other fields, we already mentioned in the introduction, the condition of such large matrices are important for not

obtaining misleading estimates of the desired parameters. For the numerical condition of a matrix the norm and the norm of the inverse play an indispensable role. Let us introduce a subset $\mathcal{U}$ of the set of well-conditioned matrices, where σ_{ij} denote the elements of Σ.

$$\mathcal{U}(\varepsilon_0, \alpha, C) := \Big\{\Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Ck^{-\alpha} \quad \forall k > 0, \\ \text{and} \;\; 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma_p) \leq \lambda_{\max}(\Sigma_p) \leq 1/\varepsilon_0\Big\}. \quad (4.1)$$

In $\mathcal{U}(\varepsilon_0, \alpha, C)$ contained are only well conditioned covariance matrices such that the maximum of the row sums of those elements which would disappear after applying the banding operator with parameter k on it can be bounded by a multiple of a certain power of k.

4.1 Banding

For (deterministic) sequences $\{k_n\}_n$ and $\{a_n\}_n$ we write

$$k_n \asymp a_n, \text{if} \lim_{n\to\infty} \frac{k_n}{a_n} = 1.$$

Let us yet recapitulate the o_P and O_P notation. Suppose $\{X_n : \; n = 1, 2, \ldots\}$ is a sequence of random variables all defined on the same probability space with probability measure $\mathbb{P}$. We denote (according to [5])

$X_n := o_P(1)$, if $\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}(|X_n| > \varepsilon) = 0.$

$X_n := o_P(a_n)$, if $\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\left(|a_n^{-1} X_n| > \varepsilon\right) = 0.$

$X_n := O_P(1)$, if $\forall\, \varepsilon > 0 \; \exists\, \delta(\varepsilon) \in (0, \infty) : \; \mathbb{P}(|X_n| > \delta(\varepsilon)) < \varepsilon \;\; \forall n \geq 1.$

$X_n := O_P(a_n)$, if $\forall\, \varepsilon > 0 \; \exists\, \delta(\varepsilon) \in (0, \infty) : \; \mathbb{P}\left(|a_n^{-1} X_n| > \delta(\varepsilon)\right) < \varepsilon \;\; \forall n \geq 1.$

The relation between these two concepts, namely O_P and o_P, is clarified by the following equivalent characterization of convergence in probability to zero, that is to say $X_n = o_P$ if and only if for every ε there exists a sequence $\delta_n(\varepsilon)$ converging to 0 such that

$$\mathbb{P}(|X_n| > \delta(\varepsilon)) < \varepsilon \;\; \forall n \geq 1.$$

The first theorem, taken from [2], determines rates of convergence for the banded covariance estimator. It shows that in the case of a Gaussian stochastic process with a covariance matrix from our mentioned set $\mathcal{U}$ the difference between the banded estimator and the covariance matrix can be bounded in probability by $\log p/n$. Its proof requires some results from [24], which we will present after stating the theorem.

Theorem 4.1 *Suppose that* $\mathbf{X}$ *is a Gaussian stochastic process, i.e.* $\mathbf{X}_i$ *is i.i.d.* $N(\mu, \Sigma_p)$ *with mean vector* μ *and population covariance matrix* $\Sigma_p \in \mathcal{U}(\varepsilon_0, \alpha, C)$*. If we also assume* $k_n \asymp (\frac{\log p}{n})^{-1/2(\alpha+1)}$*, then*

$$\|\hat{\Sigma}_{k_n,p} - \Sigma_p\| = O_P\left(\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right) = \|\hat{\Sigma}^{-1}_{k_n,p} - \Sigma_p^{-1}\| \tag{4.2}$$

uniformly on $\mathcal{U}(\varepsilon_0, \alpha, C)$*.*

Note that the optimal k_n in general depends also (additionally to p and n) on the dependence structure of the model, expressed by α.

4.1.1 Some results for large deviations

The main ingredient for the proof of theorem 4.1 is a lemma from [2], namely Lemma 4.4, whose proof relies on results for large deviations derived by Saulis and Statulevicius in [24]. The next lemma and its proof are taken from [24].

Lemma 4.2 *(Bentkus, Rudzkis [1]). Let for an arbitrary r.v.* ξ *with* $\mathbb{E}\xi = 0$ *there exist* $\gamma \geq 0$*,* $H > 0$ *and* $\bar{\Delta} > 0$ *such that*

$$|\Gamma_k(\xi)| \leq \left(\frac{k!}{2}\right)^{1+\gamma} \frac{H}{\bar{\Delta}^{k-2}}, \qquad k = 2, 3, \ldots. \tag{4.3}$$

Then for all $x \geq 0$

$$\mathbb{P}(\pm\xi \geq x) \leq \exp\left\{\frac{x^2}{2(H + (x/\bar{\Delta}^{1/(1+2\gamma)}))^{(1+2\gamma)/(1+\gamma)}}\right\}. \tag{4.4}$$

Proof. Let us introduce a function $g_m(x)$ for a nonnegative integer m and a real-valued function $g(x),\ x \in \mathbb{R}$ with existing m derivatives at $x = 0$ as follows:

$$g_m(x) = \sum_{k=0}^{m} \frac{1}{k!} g^{(k)}(0) x^k,\ x \in \mathbb{R}.$$

In particular

$$\exp_m(x) = \sum_{k=0}^{m} \frac{1}{k!} x^k,\ x \in \mathbb{R}, \quad m = 0, 1, \ldots.$$

Now we claim that $\forall n \geq 0\ \exp_{2n}(x) > 0$ holds. For $x \geq 0$ this is easily seen. Now set

$$a_k = \frac{x^{2k-1}}{(2k-1)!} + \frac{x^{2k}}{(2k)!}, \quad k \in \mathbb{N}$$

and examine a_k for $x < 0$:

$$a_k \geq 0 \Leftrightarrow x^{2k-1}(2k + x) \geq 0$$

If $x \leq -2n$, then $a_k \geq 0 \; \forall k \geq 1$ and therefore

$$\exp_{2n}(x) = 1 + \sum_{k=1}^{n} a_k > 0.$$

For $-2n < x < 0$ $a_k < 0$, but now let $k = n + 1, n + 2, \ldots$. Since

$$e^x = \exp_{2n}(x) + \sum_{k=n+1}^{\infty} a_k > 0,$$

it follows that $\exp_{2n}(x) > 0 \; \forall x \in \mathbb{R}$. Now we can apply Chebyshev's inequality as follows, since $\exp_{2n}(x)$ is nonnegative on $\mathbb{R}$ and monotonically increasing in the interval $[0, \infty)$. For all $h \geq 0$ and $x \geq 0$

$$\mathbb{P}(\xi \geq x) \leq \mathbb{P}(\exp_{2n}(h\xi) \geq \exp_{2n}(hx)) \leq (\exp_{2n}(hx))^{-1} \sum_{k=0}^{2n} \frac{1}{k!} m_k h^k. \quad (4.5)$$

The second inequality is due to Chebyshev's inequality and the first inequality follows due to the monotonically increase of the function $\exp_{2n}(x)$, since every ω of the set $\{\omega \in \Omega : \xi(\omega) \geq x\}$ is also an element of the set $\{\omega \in \Omega : \exp_{2n}(h\xi(\omega)) \geq \exp_{2n}(hx)\}$. Thus, we have

$$\mathbb{P}(\xi \geq x) \leq \inf_{h \geq 0} \left\{ (\exp_{2n}(hx))^{-1} \sum_{k=0}^{2n} \frac{1}{k!} m_k h^k \right\}. \quad (4.6)$$

Denote

$$g(x) = \exp_n \left(\sum_{r=2}^{2n} \frac{1}{r!} \gamma_k x^r \right), \quad x \in \mathbb{R}. \quad (4.7)$$

From the previous discussion on cumulants, we have the relation between moments and cumulants

$$m_k = \sum_{r_1 + \cdots + r_q = k} \frac{1}{q!} \frac{k!}{r_1! \cdots r_q!} \gamma_{r_1} \cdots \gamma_{r_q},$$

for any $k \in \mathbb{N}$, where the summation is taken over all ordered partitions $r_1 + \cdots + r_q = k$, $r_j \geq 1$ and $q = 1, 2, \ldots, k$. Under the condition $\mathbb{E}\xi = 0$

(which implies $\gamma_1 = 0$) and (4.7) we obtain

$$m_k = \sum_{q=1}^{k} \frac{1}{q!} \sum_{\substack{r_1+\cdots+r_q=k,\\ r_j\geq 2}} \frac{k!}{r_1!\cdots r_q!}\gamma_{r_1}\cdots\gamma_{r_q} =$$

$$= \sum_{q=1}^{k} \frac{1}{q!}\frac{d^k}{dx^k}\left(\sum_{r=2}^{2n}\frac{1}{r!}\gamma_k x^r\right)^q\Bigg|_{x=0} = \frac{d^k}{dx^k}g(x)\Big|_{x=0},$$

for any $k \in \{2, 3, \ldots, 2n\}$. Now we can write

$$\sum_{k=0}^{2n}\frac{1}{k!}m_k h^k = g_{2n}(h).$$

For $h \geq 0$

$$g_{2n}(h) \leq \exp_n\left(\sum_{r=2}^{2n}\frac{1}{r!}|\gamma_r|h^r\right) \tag{4.8}$$

holds. Substituting (4.8) into (4.6) we have for any $x \geq 0$

$$\mathbb{P}(\xi \geq x) \leq \inf_{h\geq 0}\left\{\left(\exp_{2n}(hx)\right)^{-1}\exp_n\left(\sum_{k=2}^{2n}\frac{1}{k!}|\gamma_k|h^k\right)\right\}. \tag{4.9}$$

Furthermore we need the inequality

$$\exp_n(x) \,/\, \exp_{2n}(2x) \leq e^{-x} \tag{4.10}$$

valid for $0 \leq x \leq 0.6$ if $n = 1$, $0 \leq x \leq 1.4$ if $n = 2$ and $0 \leq x \leq 0.8n$ if $n = 3, 4, \ldots$.
Denote $\varepsilon = x/n$ and

$$\Delta_n(\varepsilon) = e^{-2\varepsilon n}\sum_{k=0}^{2n}\frac{1}{k!}(2\varepsilon n)^k - e^{-\varepsilon n}\sum_{k=0}^{n}\frac{1}{k!}(\varepsilon n)^k.$$

Equation (4.10) is equivalent to the inequality $\Delta_n(\varepsilon) \geq 0$ for all $0 \leq \varepsilon \leq 0.6$ if $n = 1$, for all $0 \leq \varepsilon \leq 0.7$ if $n = 2$ and for all $0 \leq \varepsilon \leq 0.8$ if $n = 3, 4, \ldots$.
Let us now examine the derivative of $\Delta_n(\varepsilon)$

$$\frac{d}{d\varepsilon}\Delta_n(\varepsilon) = e^{-\varepsilon n}\frac{(\varepsilon n)^n}{(n-1)!}\left(1 - \frac{2^{2n+1}n!}{(2n)!}(\varepsilon n)^n e^{-\varepsilon n}\right).$$

Applying Stirling's formula

$$\sqrt{2\pi n}n^n e^{-n+\frac{1}{12n+1}} < n! < \sqrt{2\pi n}n^n e^{-n+\frac{1}{12n}},$$

we conclude

$$\frac{2^{2n+1}n!}{(2n)!} \leq n^{-n}e^n\sqrt{2}\exp\left\{\frac{12n+1}{12n(24n+1)}\right\}.$$

Thus, for all $\varepsilon \geq 0$ and $n = 1, 2, \ldots$

$$\frac{d\Delta_n(\varepsilon)}{d\varepsilon} \geq e^{-\varepsilon n}\frac{(\varepsilon n)^n}{(n-1)!}\left(1 - (\varepsilon e^{1-\varepsilon})^n\sqrt{2}\exp\left\{\frac{12n+1}{12n(24n+1)}\right\}\right) \tag{4.11}$$

The function $f(\varepsilon) = \varepsilon e^{1-\varepsilon}$ is strictly increasing in the interval $[0,1]$ (due to $f'(\varepsilon) = (1-\varepsilon)e^{1-\varepsilon}$) and $f(0) = 0,\ f(1) = 1$. Thus, for fixed n the right-hand side of (4.11) depend on $\varepsilon \in [0,1]$ as follows: it is positive in the interval $(0, \varepsilon_0)$, where $\varepsilon_0 \in [0,1]$ is the unique zero of $\frac{d\Delta_n(\varepsilon)}{d\varepsilon}$ and it is negative for $\varepsilon \in (\varepsilon_0, 1)$. From these facts we can conclude that for every $n \geq 1$ the function Δ_n has the following property:
if $0 < \varepsilon < 1$ and $\Delta_n(\varepsilon) \geq 0$, then $\Delta_n(\varepsilon')$ for all $0 < \varepsilon' < \varepsilon$. To prove (4.10), it suffices to show that $\Delta_1(0.6) \geq 0$, $\Delta_2(0.7) \geq 0$ and $\Delta_n(0.8) \geq 0$ for $n \geq 3$. In the case where $n < 17$, $\Delta_n(0.8) \geq 0$ can be verified to be valid by direct calculations and if $n \geq 17$, then the right-hand side of (4.11) is positive for all $\varepsilon \in (0, 0.8]$.
It suffices to consider the case $H = 1$, since if $H \neq 1$ we can introduce the random variable $\tilde{\xi} = \xi/H$ for which condition (4.3) is fulfilled with $\tilde{H} = 1$ and $\tilde{\Delta} = \bar{\Delta}\sqrt{H}$. The bound for $\mathbb{P}(\xi > x)$ can be obtained from the bound for $\mathbb{P}(\tilde{\xi} > \tilde{x})$, by substituting $x/\sqrt{H}$ and $\bar{\Delta}\sqrt{H}$ for $\tilde{x}$ and $\tilde{\Delta}$, respectively. Using (4.9) and the condition of the lemma, namely (4.3), we obtain that for all $x \geq 0$

$$\mathbb{P}(\xi \geq x) \leq \inf_{\substack{h\geq 0,\\ n\in\mathbb{N}}} \frac{\exp_n\left(\frac{1}{2}h^2\sum_{k=2}^{2n}\left(\frac{h}{\bar{\Delta}}\right)^{k-2}\left(\frac{k!}{2}\right)^{\gamma}\right)}{\exp_{2n}(hx)}. \tag{4.12}$$

On the other hand, applying Chebyshev's inequality, for all $x \geq 0$ we have

$$\mathbb{P}(\xi \geq x) \leq \mathbb{P}((1+\xi x)^2 \geq (1+x^2)^2) \leq \frac{1 + 2x\mathbb{E}\xi + x^2\mathbb{E}\xi^2}{(1+x^2)^2} \leq \frac{1}{1+x^2}, \tag{4.13}$$

where the last equality is due to the fact that $\mathbb{E}\xi = 0$ and $\mathbb{E}\xi^2 \leq H = 1$. The first inequality follows in the same way we obtained the first inequality in (4.5) from the fact that $(1+y^2)^2$ is an increasing function for $y \geq 0$. Set

$$h := \frac{x\left(x\bar{\Delta}\right)^{1/(1+\gamma)}}{x^2 + \left(x\bar{\Delta}\right)^{1/(1+\gamma)}} = x\underbrace{\frac{\left(x\bar{\Delta}\right)^{1/(1+\gamma)}}{x^2 + \left(x\bar{\Delta}\right)^{1/(1+\gamma)}}}_{<1} < x$$

and choose n so that the condition $0.8(n-1) < hx/2 \le 0.8n$ is fulfilled. Let us first consider the case $hx > 6.4$, i.e. $n \ge 5$. We claim that in this case

$$\frac{k!}{2} \le (hx)^{k-2}, \quad k = 2, 3, \ldots, 2n. \tag{4.14}$$

To this conclusion we came by observing that (4.14) follows from the easily verifiable inequality $(2n)!/2 \le (1.6(n-1))^{2n-2}$ for $n \ge 5$, since $hx > 1.6(n-1)$. Using (4.14) and the chosen values of h and n we obtain

$$\frac{1}{2}\sum_{k=2}^{2n}\left(\frac{h}{\bar{\Delta}}\right)^{k-2}\left(\frac{k!}{2}\right)^{\gamma} \le \frac{1}{2}\sum_{k=2}^{2n}\left(\frac{h}{\bar{\Delta}}(hx)^{\gamma}\right)^{k-2} \le \frac{h^2}{2}\frac{1}{1-q} = \frac{hx}{2}, \tag{4.15}$$

since

$$q = \left(\frac{h}{\bar{\Delta}}(hx)^{\gamma}\right)^{1/(1+\gamma)} = \frac{x^2}{x^2 + (x\bar{\Delta})^{1/(1+\gamma)}} < 1.$$

We prove (4.4) by substituting (4.15) into (4.12) and using (4.10), but only for $hx > 6.4$. Let us now consider the case $0 < hx \le 6.4$. We will split this case into two subcases: 1) when the condition

$$\frac{1}{1+x^2} \le \exp\left\{-\frac{hx}{2}\right\} \tag{4.16}$$

is fulfilled and 2) it is not fulfilled. In the first case (4.4) follows from (4.13). If the condition is not satisfied, then

$$\exp\left\{\frac{hx}{2}\right\} > 1 + x^2 > 1 + hx, \tag{4.17}$$

since $x > h$, as we have seen before. From (4.17) it follows that $hx > 2.5$. Thus, it remains to consider the case $2.5 < hx \le 6.4$ under the condition (4.17). In this case (4.4) is obtained from (4.12) by putting $n = 4$. Set $f(t) = (1/t) - (e^{t/2} - 1)^{-1}$, $t > 0$. According to (4.17) we have $x^2 < \exp hx/2 - 1$. Therefore,

$$\frac{h}{\bar{\Delta}} = q\left(\frac{1}{hx} - \frac{1}{x^2}\right)^{\gamma} < q\,(f(h(x)))^{\gamma}$$

holds. Since $e^{t/2} - 1 < t(t+1)/2$ holds in the interval $2.5 < t \le 6.4$, we can bound $f(t)$ for all $2.5 < t \le 6.4$ as follows:

$$f(t) < \frac{t-1}{t(t+1)} < \left.\frac{t-1}{t(t+1)}\right|_{t=2.5} \le 0.172.$$

Consequently, $(h/\bar{\Delta}) \le q(0.172)^{\gamma}$. Thus, we conclude that

$$\frac{1}{2}h^2 \sum_{k=2}^{2n} \left(\frac{h}{\bar{\Delta}}\right)^{k-2} \left(\frac{k!}{2}\right)^{\gamma} \le \frac{1}{2}h^2 \sum_{k=2}^{2n} q^{k-2} \left(0.172^{k-2}\frac{k!}{2}\right)^{\gamma} \le \frac{h^2}{2(1-q)} = \frac{hx}{2}, \tag{4.18}$$

since $0.172^{k-2}(k!/2) < 1$ for all $k = 2, 3, \ldots, 8$. Finally we can substitute (4.18) into (4.12) and obtain the assertion of the lemma using (4.4).

Before we state the next important theorem, we need to focus on notation. For $n \ge 1$ let $\xi_1, \xi_2, \ldots, \xi_n$ be independent random variables with $\mathbb{E}\xi_j = 0$ and $\sigma_j^2 = \mathbb{V}\xi_j > 0$, $j = 1, 2, \ldots$. Set

$$S_n = \sum_{j=1}^{n} \xi_j, \quad B_n^2 = \sum_{j=1}^{n} \sigma_j^2, \quad Z_n = S_n/B_n. \tag{4.19}$$

We say that a random variable ξ_j ($j \in \mathbb{N}$) satisfy condition (P), if there exist positive constants $A, C, c_1, c_2, \ldots$, such that

$$\left|\frac{\ln \mathbb{E}(\exp z\xi_j)}{z^2}\right| \le c_j^2, \quad |z| < A \quad (j \in \mathbb{N}) \tag{P}$$

and

$$\limsup_{n\to\infty} \frac{1}{B_n^2} \sum_{j=1}^{n} c_j^2 \le C. \tag{4.20}$$

Now we can state the theorem:

Theorem 4.3 *Let a random variable ξ_j with $\mathbb{E}\xi_j = 0$ and $\sigma_j^2 = \mathbb{V}\xi_j > 0$ satisfy condition (P). Then*

$$|\Gamma_k(Z_n)| \le \frac{k!\ C}{(AB_n)^{k-2}}, \quad \forall k \ge 3,$$

and for the r.v. $\xi = Z_n$ the relation of large deviations (4.4) holds with

$$\gamma = 0, \quad H = 2C \quad \textit{and} \quad \bar{\Delta} = AB_n.$$

Proof. Using the fact that

$$\Gamma_k(\xi_j) = \frac{d^k}{dz^k} = \ln \mathbb{E}(\exp(z\xi_j))\Big|_{z=0},$$

and using condition (P) and the Cauchy inequality for derivatives of analytical functions we obtain

$$|\Gamma_k(\xi_j)| \le k! c_j^2/A^{k-2}, \quad \forall k \ge 3.$$

As $\xi_j,\ j = 1, 2, \ldots$ are independent we can use the nice features of cumulants explained previously and we find

$$|\Gamma_k(S_n)| \leq k! \sum_{j=1}^{n} c_j^2 / A^{k-2}, \quad \forall k \geq 3,$$

and from (4.20) it follows that

$$|\Gamma_k(Z_n)| \leq k! C / (AB_n)^{k-2}, \quad \forall k \geq 3,$$

holds. Now, by using Lemma 4.2, we obtain the assertion of the theorem.

4.1.2 Proof of theorem concerning convergence of the banded matrix

To prove the next lemma we needed the previous results on limit theorems for large deviations. The next lemma is, as already mentioned, the main ingredient for proving Theorem 4.1

Lemma 4.4 *Let* $\mathbf{Z}_i$ *be i.i.d.* $N(\mathbf{0}, \Sigma_p)$ *and* $\lambda_{max}(\Sigma_p) \leq \varepsilon_0^{-1} < \infty$. *Let* σ_{ab} *denote the individual entries of* Σ_p. *Then*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} (Z_{ij} Z_{ik} - \sigma_{jk})\right| \geq n\nu\right] \leq C_1 \exp(-C_2 n \nu^2) \quad \textit{for } |\nu| < \delta, \tag{4.21}$$

where $C_1,\ C_2$ *and* δ *only depend on* ε_0.

Proof. It holds that

$$\begin{aligned} &\mathbb{P}\left[\left|\sum_{i=1}^{n} (Z_{ij} Z_{ik} - \sigma_{jk})\right| \geq n\nu\right] \\ &\qquad = \mathbb{P}\left[\left|\sum_{i=1}^{n} (Z_{ij}^* Z_{ik}^* - \rho_{jk})\right| \geq \frac{n\nu}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right], \end{aligned} \tag{4.22}$$

where $\rho_{jk} = \sigma_{jk}(\sigma_{jj}\sigma_{kk})^{1/2}$ and $(Z_{ij}^*, Z_{ik}^*) \sim \mathrm{N}_2(0, 0, 1, 1, \rho_{jk})$. Let us now apply an easy trick to simplify the sum in (4.22):

$$\begin{aligned} &\sum_{i=1}^{n} (Z_{ij}^* Z_{ik}^* - \rho_{jk}) = \\ &\quad = \frac{1}{4}\left[\sum_{i=1}^{n} [(Z_{ij}^* + Z_{ik}^*)^2 - 2(1 + \rho_{jk})] - \sum_{i=1}^{n} [(Z_{ij}^* - Z_{ik}^*)^2 - 2(1 - \rho_{jk})]\right], \end{aligned}$$

so we can write

$$\mathbb{P}\left[\frac{1}{4}\left|\sum_{i=1}^{n}\Big\{[(Z_{ij}^*+Z_{ik}^*)^2-2(1+\rho_{jk})]\right.\right.$$
$$\left.\left.-[(Z_{ij}^*-Z_{ik}^*)^2-2(1-\rho_{jk})]\Big\}\right|\geq\frac{n\nu}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right]\leq$$
$$\leq\mathbb{P}\left[\frac{1}{4}\left|\sum_{i=1}^{n}\Big\{[(Z_{ij}^*+Z_{ik}^*)^2-2(1+\rho_{jk})]\right|\geq\frac{n\nu}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right]+$$
$$+\mathbb{P}\left[\frac{1}{4}\left|\sum_{i=1}^{n}[(Z_{ij}^*-Z_{ik}^*)^2-2(1-\rho_{jk})]\right|\geq\frac{n\nu}{(\sigma_{jj}\sigma_{kk})^{1/2}}\right]=$$
$$=\mathbb{P}\left[\frac{1}{4}\left|\sum_{i=1}^{n}\left(\frac{(Z_{ij}^*+Z_{ik}^*)^2}{2(1+\rho_{jk})}-1\right)\right|\geq\frac{n\nu}{2(1+\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right]+$$
$$+\mathbb{P}\left[\frac{1}{4}\left|\sum_{i=1}^{n}\left(\frac{(Z_{ij}^*-Z_{ik}^*)^2}{2(1-\rho_{jk})}-1\right)\right|\geq\frac{n\nu}{2(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right]$$

Now we need only to find an upper bound for the following term, where $V_i\sim\mathrm{N}(0,1)$ i.i.d. and thus $V_i^2\sim\chi_1^2$:

$$\mathbb{P}\left[\frac{1}{4}\left|\sum_{i=1}^{n}\left(V_i^2-1\right)\right|\geq\frac{n\nu}{2(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}\right].$$

The cause for the simplification is going to become evident. Recall the fact that for a linear transformation of a random vector $\mathbf{X}$ (with p components) which is normally distributed with mean vector μ and covariance matrix Σ, the distribution of $\mathbf{Y}=C\mathbf{X}$ is $\mathrm{N}(C\mu,C\Sigma C^T)$. In our case

$$\mathbf{X}=\begin{pmatrix}Z_{ij}^*\\Z_{ik}^*\end{pmatrix},\quad C=\begin{pmatrix}1&1\\1&-1\end{pmatrix}\text{ and }\Sigma=\begin{pmatrix}1&\rho_{jk}\\\rho_{jk}&1\end{pmatrix}.$$

Σ is of this form, because, as we have mentioned, $(Z_{ij}^*,Z_{ik}^*)\sim\mathrm{N}_2(0,0,1,1,\rho_{jk})$ holds, so σ_{jj} and σ_{kk} are equal to 1. Thus,

$$\mathbf{Y}=\begin{pmatrix}Z_{ij}^*+Z_{ik}^*\\Z_{ik}^*-Z_{ik}^*\end{pmatrix}\sim\mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}2(1+\rho_{jk})&0\\0&2(1-\rho_{jk})\end{pmatrix}\right),$$

from where we conclude that

$$\frac{(Z_{ij}^*+Z_{ik}^*)^2}{2(1+\rho_{jk})}\sim\chi_1^2,\quad\frac{(Z_{ij}^*-Z_{ik}^*)^2}{2(1-\rho_{jk})}\sim\chi_1^2.$$

To be able to apply the result for large deviations we need to verify whether the random variable $V_i^2 - 1 \sim \chi_1^2 - 1$ satisfies condition (P). The lemma would then follow using theorem 4.3 and (4.4). Thus, we calculate the characteristic function of $X := V_i^2$:

$$f_X(t) = \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} \int_0^\infty e^{itx} e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{-\frac{1}{2}} dx = \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} \int_0^\infty e^{-\frac{x}{2}(1-2it)} \left(\frac{x}{2}\right)^{-\frac{1}{2}} dx.$$

To find a primitive for the last integral, we introduce a substitution of variables, namely $-\frac{x}{2}(1-2it) =: -\frac{y}{2}$, from where we conclude that $dx = \frac{dy}{1-2it}$. We now get

$$\begin{aligned} f_X(t) &= \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} \int_0^\infty e^{-\frac{y}{2}} \left(\frac{y}{2(1-2it)}\right)^{-\frac{1}{2}} \frac{dy}{1-2it} = \\ &= (1-2it)^{-\frac{1}{2}} \underbrace{\int_0^\infty \frac{e^{-\frac{y}{2}} \left(\frac{y}{2}\right)^{-\frac{1}{2}}}{\sqrt{2}\Gamma(\frac{1}{2})}}_{=1}, \end{aligned}$$

implying $f_X(t) = (1-2it)^{-\frac{1}{2}}$. The above integral equals to 1, because it represents the probability density of a χ_1^2 distributed random variable. Note that the characteristic function of a χ_n^2 distributed r.v. is $(1-2it)^{-\frac{n}{2}}$. Let us now determine the cumulants of V_i^2.

$$\log f_X(t) = -\frac{1}{2} \log(1-2it),$$

Substituting the Mercator series for $\log(1-2it)$ with $x = -2it$, we obtain

$$\sum_{k=0}^{\infty} \frac{1}{k!} \gamma_k (it)^k = \frac{1}{2} \sum_{k=1}^{\infty} \frac{(2it)^k}{k},$$

where for the left hand side we used the Taylor expansion for the logarithm of the characteristic function of r.v. X for a sufficiently small annulus region $|t| < \frac{1}{2}$, since the right-hand side converges for $|t| < \frac{1}{2}$. Comparing the coefficients of t^j on both sides, we have that for the cumulants of $V_i^2 \sim \chi_1^2$

$$\gamma_k = \frac{1}{2} \frac{k!}{k} 2^k = 2^{k-1}(k-1)!$$

holds. For a χ_n^2 distributed r.v. the cumulants equal to

$$\gamma_k = n \cdot 2^{k-1}(k-1)!.$$

For the random variables $V_i^2,\ i = 1, 2, \ldots, n$ to satisfy condition (P), we need to show that there exist positive constants A, C, c_1 such that

$$\left|\frac{\ln \mathbb{E}(\exp z\xi)}{z^2}\right| \leq c^2 \quad \text{for} \quad |z| < A$$

and

$$\limsup_{n\to\infty} \frac{1}{B_n^2} \sum_{j=1}^{n} c_1^2 \leq C,$$

where $\xi \sim \chi_1^2 - 1$. In our case, this reduces to

$$\left|\frac{1}{z^2}\frac{1}{2}(1 - \ln(1-2z))\right| < c_1^2 \text{ for } \ |z| < A$$

and

$$\limsup_{n\to\infty} \frac{1}{2n} n \cdot c_1^2 \leq C,$$

since $\mathbb{V}\xi = 2$ and thus $B_n^2 = \sum_{j=1}^{n} \sigma^2 = 2n$. The function $\ln(1-2z)$ can be expanded into a Maclaurin series for $|z| < \frac{1}{2}$ and we obtain:

$$-\frac{1}{2z^2}(\ln(1-2z)-1) = -\frac{1}{2z^2}\left(-1 - \sum_{k=1}^{\infty} \frac{(2z)^k}{k}\right) = \frac{1+2z}{z^2} + 2 + \frac{8}{3}z + a_1 z^2 + \ldots,$$

where $a_i,\ i \geq 1$ are finite constants, whose exact values are not important for further inference. The complex valued function $\frac{1+2z}{z^2} + 2 + \frac{8}{3}z + a_1 z^2 + \ldots$ has a double pole at $z = 0$ (it explodes to ∞ for z near 0), thus, it can not be bounded in $|z| < A$. You would have to remove a small circle around zero. Here at this place the authors of [2] make a mistake by saying that condition (P) is fulfilled by the mentioned random variable. For an annulus $a < |z| < A$, $\left| -\frac{1}{2z^2}(1 - \ln(1-2z))\right|$ is bounded by a positive constant C. Since the cumulants of a chi-squared distributed r.v. are $\gamma_k = 2^{k-1}(k-1)!$, we compare the found expression for γ_k with (4.3) to find that

$$\gamma = 0, \quad H = 4n, \quad \bar{\Delta} = \frac{1}{2}.$$

Set

$$x = \frac{n\nu}{2(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}$$

and plug it together with H, γ and $\bar{\Delta}$ into (4.4), in consideration of

$$(\sigma_{jj}\sigma_{kk})^{1/2}|1-\rho_{jk}| = |(\sigma_{jj}\sigma_{kk})^{1/2} - (\sigma_{jj}\sigma_{kk})^{1/2}\rho_{jk}| \leq (\sigma_{jj}\sigma_{kk})^{1/2} + |\sigma_{jk}| \leq 2\varepsilon_0^{-1}.$$

The last inequality holds due to the symmetry and positive definiteness of Σ_p, since in that case the spectral theorem can be applied, implying that there exists an orthonormal basis $\{\mathbf{v}_i\}_i$ consisting of eigenvectors of the matrix. Now every $\mathbf{x} \in \mathbb{R}^p$ can be written as $\mathbf{x} = \sum_{i=1}^p \alpha_i \mathbf{v}_i$. The expression $\mathbf{x}^T \Sigma_p \mathbf{x}$ (which is always positive, since Σ_p is positive definite) then equals to

$$\mathbf{x}^T \Sigma_p \mathbf{x} = \big(\sum_{i=1}^p \alpha_i \mathbf{v}_i\big)^T \Sigma_p \sum_{i=1}^p \alpha_i \mathbf{v}_i = \big(\sum_{i=1}^p \alpha_i \mathbf{v}_i\big)^T \sum_{i=1}^p \alpha_i \Sigma_p \mathbf{v}_i = \big(\sum_{i=1}^p \alpha_i \mathbf{v}_i\big)^T \sum_{i=1}^p \alpha_i \lambda_i \mathbf{v}_i = \sum_{i=1}^p \alpha_i^2 \lambda_i \mathbf{v}^T \mathbf{v} \leq \max_i |\lambda_i| \|\mathbf{x}\|.$$

Thus, all diagonal elements of Σ_p are bounded by the maximum eigenvalue, since every diagonal element is obtainable from $\mathbf{x}^T \Sigma_p \mathbf{x}$ by setting $\mathbf{x}$ equal to a vector with one entry equal to 1 and all the others equal to 0. Using the fact that all principal minors are bigger than zero and applying an appropriate permutation of the columns, it can be shown that the biggest element (by absolute value) of the matrix lies on the main diagonal.
After simple algebraic transformations, we obtain

$$\begin{aligned}
\mathbb{P}(\pm\xi \geq x) &\leq \exp\left\{-\frac{n\nu^2}{8(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}\left[4(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}+\nu\right]}\right\} \\
&\leq \exp\left\{-\frac{n\nu^2}{\frac{16}{\varepsilon_0}(\frac{8}{\varepsilon_0}+\nu)}\right\} = \exp\left\{-\frac{\varepsilon_0^2 n\nu^2}{16(8+\varepsilon_0\nu)}\right\} \\
&= \exp(-C_2 n\nu^2) \quad \text{for} \quad |\nu| \leq \delta,
\end{aligned}$$

since $16(8+\varepsilon_0\nu) < 16(8+\varepsilon_0\delta)$ and therefore

$$\exp\left\{-\frac{\varepsilon_0^2 n\nu^2}{16(8+\varepsilon_0\nu)}\right\} \leq \exp\left\{-\frac{\varepsilon_0^2 n\nu^2}{16(8+\varepsilon_0\delta)}\right\}.$$

Note that C_2, δ depend on ε_0 only. In the proof above we used the fact that $|\sigma_{jk}| \leq \varepsilon_0^{-1}$ for all j, $k \in \{1, \ldots, p\}$, where Σ_p is of dimension $p \times p$.

In addition to the operator norm $\|M\|$ from l_2 to l_2 we have already defined, there are some other matrix norms, too, which we will present in what follows. For a vector $\mathbf{x} = (x_1, \ldots, x_p)^T$, let

$$\|x\|_1 = \sum_{j=1}^p |x_j|, \quad \|x\|_\infty = \max_{j=1} |x_j|.$$

For a matrix M, the corresponding matrix norms induced from the vector norms from l_1 to l_1 and from l_∞ to l_∞ are, respectively,

$$\|M\|_{(1,1)} := \sup\{\|M\mathbf{x}\|_1 : \|\mathbf{x}\|_1\} = \max_j \sum_i |m_{ij}|,$$

$$\|M\|_{(\infty,\infty)} := \sup\{\|M\mathbf{x}\|_\infty : \|\mathbf{x}\|_\infty\} = \max_j \sum_i |m_{ij}|.$$

In addition to the mentioned norms we will use $\|M\|_\infty := \max_{i,j} |m_{ij}|$. The l_1 to l_1 norm arises naturally through the inequality

$$\|M\| \leq \left[\|M\|_{(1,1)}\|M\|_{(\infty,\infty)}\right]^{1/2} = \|M\|_{(1,1)} \text{ for } M \text{ symmetric}, \tag{4.23}$$

since $\|M\|_{(1,1)} = \|M\|_{(\infty,\infty)}$ for symmetric matrices M.

Proof of Theorem 4.1. The preceding equations concerning matrix norms imply

$$\|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\| \leq \|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\|_{(1,1)} \tag{4.24}$$

The right-hand side equals to

$$\max_j \sum_i \left|\hat{\sigma}_{ij} \cdot \mathbf{1}_{[|i-j|\leq k]} - \sigma_{ij}\right| \leq (2k+1)\|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\|_\infty,$$

where the last inequality is implied by the fact that the banded matrix $B_k(\hat{\Sigma}_p)$ contains at most $2k+1$ elements different from zero, since the banding operator preserves k elements from each diagonal plus the diagonal element. Thus, we can write

$$\|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\| = O_P\left(k\|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\|_\infty\right). \tag{4.25}$$

Let $\hat{\Sigma}^0 := \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T$ and w.l.o.g. $\mathbb{E}\mathbf{X}_i = \mathbf{0}$ for all $i \in \{1,\ldots,n\}$. Furthermore denote $\mathbf{X}_i = (x_{i1},\ldots,x_{ip})$ and denote I the set of all indices (m,j) such that $|m-j| \leq k$.

$$\begin{aligned}
\mathbb{P}\Big[\|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\|_\infty \geq t\Big] &= \mathbb{P}\left[\max_{m,j}\Big|\sum_{i=1}^n x_{ij}x_{im} - n\sigma_{jm}\Big| \geq t\right] \\
&= \mathbb{P}\Big[\bigcup\bigcup\Big|\sum_{i=1}^n x_{ij}x_{im} - \sigma_{jm}\Big| \geq nt\Big] \\
&\leq \sum\sum \mathbb{P}\Big[\Big|\sum_{i=1}^n x_{ij}x_{im} - \sigma_{jm}\Big| \geq nt\Big] \\
&\leq (2k+1)p \cdot \mathbb{P}\Big[\Big|\sum_{i=1}^n x_{ij}x_{im} - \sigma_{jm}\Big| \geq nt\Big],
\end{aligned}$$

where the we take the double union and the sum over I. The second inequality was obtained by applying the union sum inequality for a measure function (in this case, the probability measure, of course) over non disjoint sets. The last inequality follows due to the fact that the index set I contains at most $(2k+1)p$ elements different from zero. Now Lemma 4.4 can be applied upon $\mathbf{X}_i$ ($\mathbb{E}\mathbf{X}_i = 0$) to obtain

$$\mathbb{P}\Big[\|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\|_\infty \geq t\Big] \leq (2k+1)p \exp\{-nt^2\gamma(\varepsilon_0, \lambda)\} \tag{4.26}$$

for $|t| \leq \lambda \equiv \lambda(\varepsilon_0)$. By choosing $t = M(\frac{\log(pk)}{n})^{1/2}$ for an arbitrary M, we conclude that, uniformly on $\mathcal{U}$,

$$\|B_k(\hat{\Sigma}_p) - B_k(\Sigma_p)\|_\infty = O_P\left(\left(\frac{\log(pk)}{n}\right)^{\frac{1}{2}}\right) = O_P\left(\left(\frac{\log p}{n}\right)^{\frac{1}{2}}\right),$$

since $k < p$ ($\log(pk) = \log p + \log k < 2\log p$). On the other hand, by (4.1)

$$\|B_k(\hat{\Sigma}_p) - \Sigma_p\|_\infty \leq Ck^{-\alpha}$$

for $\Sigma_p \in \mathcal{U}(\varepsilon_0, \alpha, C)$.
Combining the last two inequalities, we can bound the desired expression as follows:

$$\begin{aligned}
\|B_k(\hat{\Sigma}_p^0) - \Sigma_p\|_\infty &\leq \|B_k(\hat{\Sigma}_p^0) - B_k(\Sigma_p)\|_\infty + \|B_k(\Sigma_p) - \Sigma_p\| \\
&= O_P\left(\min\left\{\left(\frac{\log p}{n}\right)^{1/2},\ k^{-\alpha}\right\}\right) \\
&= O_P\left(\min\left\{\left(\frac{\log p}{n}\right)^{1/2},\ \left(\frac{\log p}{n}\right)^{\alpha/2(\alpha+1)}\right\}\right) \\
&= O_P\left(\left(\frac{\log p}{n}\right)^{\alpha/2(\alpha+1)}\right)
\end{aligned}$$

Thus, the assertion of Theorem 4.1 follows for $B_k(\Sigma^0)$. Let us yet generalize the results for stochastic processes which mean is non zero. Denote $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_p)^T$,

$$\begin{aligned}
\|B_k(\hat{\Sigma}_p^0) - B_k(\hat{\Sigma})_p\| &= \left\|B_k\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\bar{\mathbf{X}}^T + \bar{\mathbf{X}}\mathbf{X}_i^T - \bar{\mathbf{X}}\bar{\mathbf{X}}^T\right)\right\| \\
&= \|B_k\left(\bar{\mathbf{X}}\bar{\mathbf{X}}^T\right)\| \leq (2k+1)\max_{1\leq j\leq p}|\bar{X}_j|^2 \\
&= O_P\left(k\sqrt{\frac{\log p}{n}}\right) = O_P\left(\left(\frac{\log p}{n}\right)^{\alpha/2(\alpha+1)}\right),
\end{aligned}$$

where we have used the well-known fact that $\bar{X} = O_P(1/\sqrt{n})$. Since

$$\|[B_{k_n}(\hat{\Sigma}_p)]^{-1} - \Sigma_p^{-1}\| = \Omega_P(\|B_{k_n}(\hat{\Sigma}_p) - \Sigma_p\|),$$

uniformly on $\mathcal{U}(\varepsilon_0, \alpha, C)$ the result follows.

4.1.3 Choice of the banding parameter

The previous results give us the rate of $k = k_n$ that guarantees convergence of the banded estimator $\hat{\Sigma}_k$, but they do not tell us much about how to chose a value for k, if we have given a sample, i.e. a dataset. The obvious choice for k would be to minimize the risk

$$R(k) = \mathbb{E}\|\Sigma_{k,p} - \Sigma_p\|_{(1,1)} \tag{4.27}$$

and then to obtain the optimal k (called the "oracle" k) with

$$k_0 = \arg\min_k R(k), \tag{4.28}$$

where under the $(1,1)$-norm for a matrix $M = [m_{ij}]$ we mean $\|M\|_{(1,1)} := \sup\{\|M\mathbf{x}\|_1 : \|\mathbf{x}\|_1 = 1\} = \max_j \sum_i |m_{ij}|$. The choice of the matrix norm in (4.27) is somehow arbitrary. Bickel and Levina ([2]) found out that the choice of k is not sensitive to the choice of the matrix norm. Further they found out that the l_1 matrix norm performed slightly better than other in simulations and was also easier to compute. To estimate the risk and thus k_0 they proposed a resampling scheme as follows: divide the original sample into two samples at random and use the sample covariance matrix of one sample as the "target" to choose the best k for the other sample. Let n_1, $n_2 = n - n_1$ be the two sample sizes for the random split, and let $\hat{\Sigma}_1^{(\nu)}$, $\hat{\Sigma}_2^{(\nu)}$ be the two sample covariance matrices from the ν-th split, for $\nu = 1, \ldots, N$. Then the risk can be estimated by

$$\hat{R}(k) = \frac{1}{N}\sum_{\nu=1}^{N} \|B_k\big(\hat{\Sigma}_1^{(\nu)}\big) - \hat{\Sigma}_2^{(\nu)}\|_{(1,1)} \tag{4.29}$$

and k is selected by

$$\hat{k} = \arg\min_k \hat{R}(k)$$

Little sensitivity was found to the choice of n_1 and n_2, so they used $n_1 = n/3$ throughout the paper.
The oracle k_0 provides the best choice in terms of expected loss, whereas $\hat{k}$

tries to adapt to the data at hand. Another comparison is that of $\hat{k}$ to the best band choice for the sample in question:

$$k_1 = \arg\min_k \|\hat{\Sigma}_{k,p} - \Sigma_p\|_{1,1}.$$

Here k_1 is a random quantity, and its loss is always smaller than that of k_0. Numerical simulations show that $\hat{k}$ generally agrees very well with both k_0 and k_1, which are quite close for normal data. For heavier-tailed data, one would expect more variability; in that case, the agreement between $\hat{k}$ and k_1 is more important than that between $\hat{k}$ and k_0. Since in (4.29) $\hat{\Sigma}_2$ is known to be a very noisy estimate of Σ_p, it may be surprising that $\hat{\Sigma}_2$ as the target works at all. Nevertheless, it is an unbiased estimate and numerical results showed that (4.29) tend to overestimate the actual value of the risk, but even though, it gives very good results for choosing k, see [2] (page 212-217). Criterion (4.29) can be used to select k for the Cholesky-based $\tilde{\Sigma}_{k,p}$ as well. One has to keep in mind that while $\hat{\Sigma}_{k,p}$ is always well-defined, $\tilde{\Sigma}_{k,p}$ in only well-defined for $k < n$, since otherwise regressions become singular. Thus, if $p > n$, k can only be chosen from the range $0, \ldots, n-1$, not $0, \ldots, p-1$.

4.2 General tapering of the covariance matrix

As we have noted before in Chapter 3.1, one problem with simple banding of the covariance matrix is the eventual loss of positive definiteness. Furrer and Bengtsson [11] examine the estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, namely the ensemble Kalman filter and the ensemble square-root Kalman filter. We will explain these variants in more details. The best known filtering (data assimilation) algorithm is in the context of Gaussian distributions and linear system dynamics, where the prior and posterior probability density functions are described by the Kalman filter recursion [16]. To address the heavy computational expense of the Kalman filter recursions in very large-scale problems, Evensen [10] proposes the ensemble Kalman filter. Conceptually, the ensemble Kalman filter implements Bayes theorem by perturbing a (Gaussian) forecast sample to produce a posterior sample with the correct first two moment structures. The ensemble Kalman filter is optimal only in Gaussian settings, but because it samples using the empirical forecast distribution the method is known to have excellent non-Gaussian properties in various settings. In particularly, they apply their methods to an important application area where the employed sample sizes are several orders of magnitude smaller than the system dimension, that is to say in numerical weather prediction. To reduce necessary ensemble size requirements and to address rank-deficient sample covariances, covariance-shrinking (tapering) based on the Schur product of the

prior sample covariance and a positive definite function is demonstrated to be a simple, computationally feasible, and very effective technique. The positive definiteness can be preserved by tapering the sample covariance matrix, that is, replacing $\hat{\Sigma}_p$ with $\hat{\Sigma}_p * R$, where $*$ denotes Schur (coordinate-wise) matrix multiplication and where $R = [r_{ij}]$ is a positive definite, symmetric matrix. The Schur product of positive definite matrices is again positive definite. An elegant proof of this claim in a probabilistic context is as follows: if X and Y are independent, mean $\mathbf{0}$ random vectors with $\mathbb{C}(X) = A$ and $\mathbb{C}(Y) = B$, then

$$\mathbb{C}(X * Y) = A * B.$$

Schur proved this fact in a general setting of bounded bilinear forms with infinitely many variables, see [25]. Now we will show how this positive definite and symmetric matrix R can be obtained. Let A be a countable set of labels of cardinality $|A|$. Let $\rho : A \times A \to \mathbb{R}^+$, $\rho(a, a) = 0$ for all a, be a function that can be interpreted as a distance function of the point (a, b) from the diagonal. One example of ρ is obvious, namely $\rho(a, b) = |a - b|$, where a and b are identified with points in $\mathbb{R}^+$ and $|\cdot|$ is a norm in $\mathbb{R}^+$.
Let $R = [r_{ab}]_{a,b\in A}$ be a symmetric, positive definite matrix with $r_{ab} = g(\rho(a, b))$, $g : \mathbb{R}^+ \to \mathbb{R}^+$. We assume further $g(0) = 1$ and g is decreasing to 0. Then $R * M$ is a regularization of M. The familiar banding operator is obtained by $\rho(i, j) = |i - j|$, $g(t) = \mathbf{1}_{[t\leq k]}$ (which is not negative definite).
In general, let $R_\sigma = [r_\sigma(a, b)]$, where

$$r_\sigma(a, b) = g\left(\frac{\rho(a, b)}{\sigma}\right), \quad \sigma \geq 0.$$

In preparation of the next theorem, we will need an assumption on the mentioned function g.

Assumption A. g is continuous, $g(0) = 1$, g is nonincreasing and

$$\lim_{x\to\infty} g(x) = 0.$$

One example of such R_σ is

$$r_\sigma(i, j) = \left(1 - \frac{|i - j|}{\sigma}\right)_+,$$

and another one would be

$$r_\sigma(i, j) = e^{-|i-j|/\sigma}.$$

Now define

$$R_\sigma(M) := [m_{ab} r_\sigma(a,b)]$$

with $R_0(M) = M$. Note that

$$\lim_{\sigma\to\infty} R_\sigma(M) = M.$$

Furthermore, denote the range of $g_\sigma(\rho(a,b))$ by $\{g_\sigma(\rho_1),\ldots,g_\sigma(\rho_L)\}$ where $\{0<\rho_1<\cdots<\rho_L\}$ is the range of $\rho(a,b)$, $a,b\in A$. Note that L depends on $|A|=p$.

Theorem 4.5 *Let $\Delta(\sigma)=\sum_{l=1}^{L} g_\sigma(\rho_l)$. Note that Δ depends on $|A|=p$ and the range of ρ. Suppose Assumption A holds. Then if*

$$\Delta \asymp (n^{-1}\log p)^{-1/2(\alpha+1)},$$

the conclusion of Theorem 4.1 holds for $R_\sigma(\hat{\Sigma}_p)$, namely

$$\|R_\sigma(\hat{\Sigma}_p)-\Sigma_p\| = O_P\left(\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right) = \|[R_\sigma(\hat{\Sigma}_p)]^{-1}-\Sigma_p^{-1}\|.$$

Proof. The proof of this theorem is closely related to the proof of Theorem 4.1. There is only one modification, namely the following lemma substitutes for (4.25)

4.3 Banding the Cholesky factor of the inverse

Theorem 4.1 gives the rate of convergence of the banded sample covariance matrix to the population covariance matrix. The next theorem proposes that very similar results can be obtained by banding the Cholesky factor of the inverse. Let us firstly define an appropriate space for covariance matrices analogously to $\mathcal{U}(\varepsilon_0,\alpha,C)$: For $\Sigma^{-1}=T(\Sigma)^T D^{-1}(\Sigma)T(\Sigma)$ with $T(\Sigma)$ lower triangular and $T(\Sigma)\equiv[t_{ij}(\Sigma)]$, let

$$\mathcal{U}^{-1}(\varepsilon_0,\alpha,C) := \Bigg\{\Sigma : 0<\varepsilon_0\le\lambda_{\min}(\Sigma_p)\le\lambda_{\max}(\Sigma_p)\le\varepsilon_0^{-1},$$
$$\max_i \sum_{j<i-k}|t_{ij}(\Sigma)|\le Ck^{-\alpha}\quad\forall k\le p-1\Bigg\}. \quad (4.30)$$

Theorem 4.6 *Suppose* $\mathbf{X}$ *is a Gaussian stochastic process which population covariance matrix* Σ_p *is an element of* $\mathcal{U}^{-1}(\varepsilon_0, \alpha, C)$. *If* $k_n \asymp (\frac{\log p}{n})^{-1/2(\alpha+1)}$ *and* $n^{-1}\log p = o(1)$, *then*

$$\|\tilde{\Sigma}_{k_n,p}^{-1} - \Sigma_p^{-1}\| = O_P\left(\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right) = \|\tilde{\Sigma}_{k_n,p} - \Sigma_p\|, \tag{4.31}$$

uniformly for $\Sigma \in \mathcal{U}^{-1}$.

To prove Theorem 4.6 we will need an additional lemma. The reason that the argument of Theorem 4.1 cannot be applied simply to this theorem is that, as we have already mentioned, $\tilde{\Sigma}^{-1}$ is not the same as $B_k(\hat{\Sigma}^{-1})$, which is not well defined if $p > n$.

Lemma 4.7 *Under the conditions of Theorem 4.6, the following assertions hold uniformly on* $\mathcal{U}$:

$$\max\left\{\|\tilde{\mathbf{a}}_j^{(k)} - \mathbf{a}_j^{(k)}\|_\infty : \ 1 \le j \le p\right\} = O_P(n^{-1/2}\log^{1/2} p), \tag{4.32}$$

$$\max\left\{|\tilde{d}_{j,k}^2 - d_{j,k}^2| : \ 1 \le j \le p\right\} = O_P((n^{-1}\log p)^{\alpha/(2(\alpha+1))}), \tag{4.33}$$

and

$$\|A_k\| = \|D_k^{-1}\| = O(1), \tag{4.34}$$

where $\tilde{\mathbf{a}}_j^{(k)} = (\tilde{a}_{j1}^{(k)}, \ldots, \tilde{a}_{j,j-1}^{(k)})$ *are, as we noted before, the empirical estimates of the vectors* $\mathbf{a}_j^{(k)} = (a_{j1}^{(k)}, \ldots, a_{j,j-1}^{(k)})$ *and* $\tilde{d}_{j,k}^2$ *are the empirical estimates of the* $d_{j,k}^2$ *for* $1 \le j \le p$.

Proof. The first claim needed to prove Lemma 4.7 can be showed by using our crucial lemma, namely Lemma 4.4. To see that

$$\|\mathbb{V}\mathbf{X} - \hat{\mathbb{V}}\mathbf{X}\|_\infty = O_P(n^{-1/2}\log^{1/2} p), \tag{4.35}$$

note that the entries of $\hat{\mathbb{V}}\mathbf{X} - \mathbb{V}\mathbf{X}$ can be bounded by

$$\frac{1}{n}\Big|\sum_{i=1}^n X_{ia}X_{ib} - \sigma_{ab}\Big| + \frac{1}{n^2}\Big|\sum_{i=1}^n X_{ia}\Big|\Big|\sum_{i=1}^n X_{ib}\Big|,$$

since $\hat{\mathbb{V}}\mathbf{X} = \frac{1}{n}\sum_{i=1}^n (X_{ia}X_{ib} - X_{ia}\bar{X}_b - \bar{X}_a + \bar{X}_b\bar{X}_a)$, where w.l.o.g. we assumed $\mathbb{E}\mathbf{X} = \mathbf{0}$. Lemma 4.4 ensures that

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^n X_{ia}X_{ib} - \sigma_{ab}\right| \ge \nu\right] \le C_1\exp(-C_2 n\nu^2)$$

for $|\nu| < \delta$. The last inequality implies

$$\mathbb{P}\left[\max_{a,b}\frac{1}{n}\left|\sum_{i=1}^{n}X_{ia}X_{ib}-\sigma_{ab}\right|\geq\nu\right]\leq C_1p^2\exp(-C_2n\nu^2)$$

for $|\nu| < \delta$. Now choose $\nu = M(\frac{\log p^2}{nC_2})^{1/2}$ for M arbitrary. Thus, it follows that

$$\mathbb{P}\left[\|\mathbb{V}\mathbf{X}-\hat{\mathbb{V}}\mathbf{X}\|_\infty\geq M\sqrt{\frac{2\log p}{nC_2}}\right]\leq C_1p^2\exp(-M^2\log p^2)$$

holds. And the last inequality is equivalent to

$$\begin{aligned}\mathbb{P}\left[\left(\sqrt{\frac{\log p}{n}}\right)^{-1}\|\mathbb{V}\mathbf{X}-\hat{\mathbb{V}}\mathbf{X}\|_\infty\geq M\sqrt{\frac{2}{C_2}}\right]&\leq C_1p^2\exp(-M^2\log p^2)\\&=C_1p^2p^{-2M^2}.\end{aligned}$$

Since M can be chosen arbitrarily, (4.35) follows. Analogously, we conclude

$$\max_j\|\mathbb{V}^{-1}\mathbf{Z}_j^{(k)}-\hat{\mathbb{V}}^{-1}\mathbf{Z}_j^{(k)}\|_\infty=O_P(n^{-1/2}\log^{1/2}p).$$

Claim (4.32) and $\|A_k\| = O_P(1)$ follow from (3.3), (4.35) and the last equation. For proving (4.33), we firstly note that

$$d_{jk}^2=\mathbb{V}\varepsilon_j=\mathbb{V}(X_j-\hat{X}_j)=\mathbb{V}X_j-\mathbb{V}\left(\sum_{t=j-k}^{j-1}a_{jt}^{(k)}X_t\right),$$

since X_j is orthogonal onto $\hat{X}_j$ according to standard regression theory. Thus, analogously

$$\tilde{d}_{jk}^2=\hat{\mathbb{V}}X_j-\hat{\mathbb{V}}\left(\sum_{t=j-k}^{j-1}\tilde{a}_{jt}^{(k)}X_t\right),$$

and since the covariance operator is linear, we conclude

$$\begin{aligned}|\tilde{d}_{jk}^2-d_{jk}^2|\leq&|\mathbb{V}X_j-\hat{\mathbb{V}}X_j|\\&+\left|\hat{\mathbb{V}}\sum_{t=j-k}^{j-1}\left(\tilde{a}_{jt}^{(k)}-a_{jt}^{(k)}\right)X_t\right|\\&+\left|\hat{\mathbb{V}}\sum_{t=j-k}^{j-1}a_{jt}^{(k)}X_t-\mathbb{V}\sum_{t=j-k}^{j-1}a_{jt}^{(k)}X_t\right|.\end{aligned}\tag{4.36}$$

By the sum $\sum_{t=j-k}^{j-1}$ we mean $\sum_{t=\max(1,j-k)}^{j-1}$ and we omitted the longer notation due to its cumbersomeness. Let us now examine the limiting behaviour of these three terms in the last inequality. For the first term we have already showed that it is of the form $O_P(n^{-1/2}\log^{1/2} p)$ in the first part of this lemma. The second one can be written as

$$\begin{aligned}
&\Big|\sum\big\{\big(\tilde{a}_{js}^{(k)}-a_{js}^{(k)}\big)\big(\tilde{a}_{jt}^{(k)}-a_{jt}^{(k)}\big)\mathbb{C}(X_s,X_t): j-k\le s,\ t\le j-1\big\}\Big| \\
&\quad\le\left(\sum_{t=j-k}^{j-1}\big|\tilde{a}_{jt}^{(k)}-a_{jt}^{(k)}\big|\hat{\mathbb{V}}^{1/2}(X_t)\right)^2 \\
&\quad\le k^2\max_t(\tilde{a}_{jt}^{(k)}-a_{jt}^{(k)})^2\max_t\hat{\mathbb{V}}(X_t) \\
&\quad= O_P(k^2n^{-1}(\log p)^2)=O_P\left((n^{-1}\log p)^{\alpha/2(\alpha+1)}\right).
\end{aligned} \tag{4.37}$$

The first inequality follows from the Cauchy-Schwartz inequality in $L_2(\Omega,\mathfrak{A},\mathbb{P})$, namely $\mathbb{C}(X_s,X_t)\le(\mathbb{V}(X_s)\mathbb{V}(X_t))$, whereas the last equality follows from (4.32) and the fact that $\|\Sigma_p\|\le\varepsilon_0^{-1}$ and from the assumption $n^{-1}\log p=o(1)$ that we required in Theorem 4.6 only to avoid a cumbersome rate of convergence. The third term in (4.37) is similarly bounded, since for a series of random variables Y_i,

$$\hat{\mathbb{V}}\left(\sum_{i=1}^n Y_i\right)=\sum_{i=1}^n\hat{\mathbb{V}}(Y_i)+\sum_{i=1}^n\sum_{\substack{j=1\\ i\ne j}}^n\hat{\mathbb{C}}(Y_i,Y_j)$$

holds. Thus, (4.33) follows. After observing that

$$\begin{aligned}
d_{jk}^2&=\mathbb{V}\left(X_j-\sum\left\{a_{jt}^{(k)}X_t:\max(1,j-k)\le t\le j-1\right\}\right)\\
&\ge\varepsilon_0\left(1+\sum\left(a_{jt}^{(k)}\right)^2\right)\ge\varepsilon_0
\end{aligned}$$

holds, (4.34) follows and thus the lemma is completely proved.

Proof of Theorem 4.6 The proof of this theorem is in some points identical to the proof of Theorem 4.1. As before we need only to show that

$$\|\tilde{\Sigma}_{k,p}^{-1}-\Sigma_{k,p}^{-1}\|=O_P(n^{-1/2}\log^{1/2}p) \tag{4.38}$$

and

$$\|\Sigma_{k,p}^{-1}-\Sigma_p^{-1}\|=O(k^{-\alpha}). \tag{4.39}$$

By definition, see section 3.2,

$$\tilde{\Sigma}_{k,p}^{-1}-\Sigma_{k,p}^{-1}=(I-\tilde{A}_k)\tilde{D}_k^{-1}(I-\tilde{A}_k)^T-(I-A_k)D_k^{-1}(I-A_k)^T, \tag{4.40}$$

where $\tilde{A}_k$ and $\tilde{D}_k$ are the empirical versions of A_k and D_k. Now we apply a standard inequality, which can be easily verified, but whose verification we will omit, since it is elementary, namely

$$\begin{aligned}\|A^{(1)}A^{(2)}A^{(3)} - B^{(1)}B^{(2)}B^{(3)}\| \\ \leq \sum_{j=1}^{3} \|A^{(j)} - B^{(j)}\| \prod_{k\neq j} \|B^{(k)}\| \\ + \sum_{j=1}^{3} \|B^{(j)}\| \prod_{k\neq j} \|A^{(k)} - B^{(k)}\| + \prod_{j=1}^{3} \|A^{(j)} - B^{(j)}\|.\end{aligned}$$

Take $A^{(1)} = [A^{(3)}]^T = I - \tilde{A}_k$, $B^{(1)} = [B^{(3)}]^T = I - A_k$, $A^{(2)} = \tilde{D}_k^{-1}$ and $B^{(2)} = D_k^{-1}$, substitute them into the last inequality and (4.38) will follow after applying Lemma 4.7. To show (4.39), we need to use the fact that for an arbitrary matrix M,

$$\begin{aligned}&\|MM^T - B_k(M)B_k(M^T)\| \\ &\quad = \| - \left(2MB_k(M) - 2MM^T + B_k(M)B_k(M^T) - 2MB_k(M) + MM^T\right) \| \\ &\quad \leq 2\|M\|\|B_k(M) - M\| + \|B_k(M) - M\|^2\end{aligned}$$

holds. This fact can now be applied as follows. Since $\Sigma_p^{-1} = T(\Sigma)^T D^{-1}(\Sigma)T(\Sigma)$ and since, the entries of the diagonal matrix D are all positive, thus, there exist a diagonal matrix $D^{1/2}$ such that $D = DD^{1/2} = DD^{1/2T}$. The same is valid for D^{-1}. Now set $M = T(\Sigma)^T D^{-1/2}$. The expression $\|B_k(M) - M\|$ is bounded by $Ck^{-\alpha}$ according to the assumption $\Sigma \in \mathcal{U}^{-1}$, since it contains only elements on entries (i, j) of the matrix for which $k < i - j$.

5 Theorems of large deviations for sums of dependent random variables

Theorem 4.3 states in which case for a sum of independent, not necessarily identical distributed, random variables expressions can be obtained for large deviations in the form of (4.4). In this chapter we will give some theorems of large deviations for sums of dependent random variables. Therefore, we again need to focus on notation.
Let X_t be a random process on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\left\{\mathcal{F}_s^t, 1 \leq s \leq t < \infty\right\}$ be a family of σ-algebras such that

1) $\mathcal{F}_s^t \subset \mathcal{F}, \quad \forall s \leq t;$

2) $\mathcal{F}_{s_1}^{t_1} \subset \mathcal{F}_{s_2}^{t_2}, \quad \forall [s_1,t_1] \subset [s_2,t_2]$;

3) $\mathcal{F}_s^t \supset \sigma\big\{X_u : s \le u \le t\big\}$.

Now we introduce the concepts of $\alpha-$, $\varphi-$ and $\psi-$mixing that are used to establish upper bounds for mixed cumulants and/or moments. These concepts go back to Andrey Nikolayevich Kolmogorov who contributed to many areas of pure and applied mathematical research, especially to the fields of probability theory and topology. In addition to his work on the foundations of probability, he contributed profound papers on stochastic processes, especially Markov processes. The above mentioned concepts are defined as follows:

$$\alpha(s,t) = \sup_{A\in\mathcal{F}_1^s, B\in\mathcal{F}_t^\infty} |\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)|$$

(Rosenblatt, 1956, [23]),

$$\varphi(s,t) = \sup_{A\in\mathcal{F}_1^s, B\in\mathcal{F}_t^\infty} \left|\frac{\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)}\right|$$

(Ibragimov, 1959, [14]),

$$\psi(s,t) = \sup_{A\in\mathcal{F}_1^s, B\in\mathcal{F}_t^\infty} \left|\frac{\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)\mathbb{P}(B)}\right|$$

(Blum, Hanson, Koopmans, 1963, [4]). The idea of this concept is after bounding a random variable with one of the mixing functions, by assuming that the underlying r.v. is e.g. α-mixing, you obtain that the bounded expression tends to zero for $|t-s| \to \infty$, since $\alpha(s,t) \to 0$ for $|t-s| \to \infty$. Recall that the letter "c" as subscript of a random variable ξ denotes the centered r.v.:

$$\xi_c := \xi - \mathbb{E}\xi.$$

In the investigation and estimation of cumulants $\Gamma_k(S_n)$, where S_n is the sum of n random variables ξ_i, it will be more convenient for us to express $\Gamma(X_{t_1},\dots,X_{t_k})$ through centered moments.

$$\mathbb{E}_c(X_I) := \mathbb{E}\big\{X_{t_1}(X_{t_2}\dots(X_{t_{m-1}}(X_{t_m})_c)_c)_c\big\},$$

where $X_I := (X_{t_1},\dots,X_{t_m})$ for $(t_1,\dots,t_m)$ is a partition of I. Sometimes the notation $E_c(X_I)$ will be replaced by $\mathbb{E}_c X_{t_1}\dots X_{t_m}$. We have

$$\begin{aligned} &\mathbb{E}_c X_t = \mathbb{E}X_t, \quad \mathbb{E}_c X_s X_t = \mathbb{E}X_s X_t - \mathbb{E}X_s\mathbb{E}X_t, \\ &\mathbb{E}_c X_{t_1}X_{t_2}X_{t_3} = \mathbb{E}X_{t_1}X_{t_2}X_{t_3} - \mathbb{E}X_{t_1}\mathbb{E}X_{t_2}X_{t_3} \\ &\qquad\qquad - \mathbb{E}X_{t_1}X_{t_2}\mathbb{E}X_{t_3} + \mathbb{E}X_{t_1}\mathbb{E}X_{t_2}\mathbb{E}X_{t_3}. \end{aligned} \tag{5.1}$$

5.1 Bounds of the k-th order centered moments of random processes with mixing

The upper bounds of $\mathbb{E}_c X_s X_t = \mathbb{E}X_s X_t - \mathbb{E}X_s \mathbb{E}X_t$, expressed through α, φ or ψ, are very important in limit theorems for sums

$$S_n = \sum_{t=1}^{n} X_t$$

of dependent random variables under different mixing conditions. Two basic ones are

$$|\mathbb{E}_c X_s X_t| \leq 4C^2 \alpha(s,t), \tag{A}$$

if $|X_s| \leq C$ and $|X_t| \leq C$ with probability 1 (Volkonskii, Rozanov, [27]);

$$|\mathbb{E}_c X_s X_t| \leq 6\alpha^{1-\frac{1}{u}-\frac{1}{v}}(s,t)\mathbb{E}^{\frac{1}{u}}|X_s|^u \mathbb{E}^{\frac{1}{v}}|X_t|^v$$

for any $u \geq 1$, $v \geq 1$ $1/u + 1/v \leq 1$, if $\mathbb{E}|X_s|^u$ and $\mathbb{E}|X_t|^v$ are finite (Davydov, [6]). There are similar upper bounds expressed through φ and ψ which we will omit. One should note that the inequalities

$$\alpha(s,t) \leq \varphi(s,t) \leq \psi(s,t)$$

make possible the transition from bounds in terms of $\alpha(s,t)$ to $\varphi(s,t)$ and from $\varphi(s,t)$ to $\psi(s,t)$ by means of direct change of mixing functions (see Iosifescu, [15]). Let us now generalize these bounds for $\mathbb{E}_c X_{t_1} \dots X_{t_k}$.

Theorem 5.1 *If $|X_{t_j}| \leq C$ with probability 1, $j = 1, \dots, k$, $k \geq 2$, then for all $i = 1, \dots, k-1$*

$$|\mathbb{E}_c X_{t_1} \dots X_{t_k}| \leq 2^k C^k \alpha(t_i, t_{i+1}).$$

Proof. The upper bound for the centered moments will be proved after introducing new random variables, which will make the notation of the proof easier.
Associate random variables $Y_{t_1}, \dots, Y_{t_k}$ with the random variables $X_{t_1}, \dots, X_{t_k}$ by the relations

$$\begin{aligned} Y_{t_j} &= X_{t_j}(Y_{t_{j+1}})_c, \quad 1 \leq j < k, \\ Y_{t_k} &= X_{t_k}, \end{aligned} \tag{5.2}$$

where the symbol "c" as subscript of a random variable denotes the operation of centering of a random variable by its mean, as we have already used it. Obviously, for all i, j, $1 \leq j \leq i < k$,

$$Y_{t_j} = X_{t_j}(X_{t_{j+1}} \dots (X_{t_i}(Y_{t_{i+1}})_c)_c)_c. \tag{5.3}$$

In particular,

$$Y_{t_j} = X_{t_j}(X_{t_{j+1}} \dots (X_{t_{k-1}}(X_{t_k})_c)_c)_c, \tag{5.4}$$

$$\mathbb{E}Y_{t_1} = \mathbb{E}_c X_{t_1} \dots X_{t_k}. \tag{5.5}$$

Due to measurability of $X_{t_j}, \dots, X_{t_i}$ with respect to $\mathcal{F}_1^{t_i}$ we obtain from (5.3)

$$\mathbb{E}(Y_{t_j}|\mathcal{F}_1^{t_i}) = X_{t_j}(X_{t_{j+1}} \dots (X_{t_{k-1}}(\mathbb{E}(Y_{t_{i+1}}|\mathcal{F}_1^{t_i}))_c)_c)_c,$$

$$\mathbb{E}Y_{t_1} = \mathbb{E}X_{t_1}(X_{t_2} \dots (X_{t_i}(\mathbb{E}(Y_{t_{i+1}}|\mathcal{F}_1^{t_i}))_c)_c)_c. \tag{5.6}$$

A method for finding upper bounds for the centered moments is based on successive application of the Hölder and Minkowski inequalities to equality (5.6) as well as on relation (5.5). We will illustrate this with an example of three random variables. The first identity in the sequel comes from (5.1).

$$\begin{aligned}
&|\mathbb{E}_c X_1 X_2 X_3| \\
&\quad = |\mathbb{E}X_1X_2X_3 - \mathbb{E}X_1\mathbb{E}X_2X_3 - \mathbb{E}X_1X_2\mathbb{E}X_3 + \mathbb{E}X_1\mathbb{E}X_2\mathbb{E}X_3| \\
&\quad \le 4|\mathbb{E}X_1X_2X_3| \le 4\mathbb{E}^{1/u_1}|X_1|^{u_1}\mathbb{E}^{1/v_1}|X_2X_3|^{v_1} \\
&\quad \le 4\mathbb{E}^{1/u_1}|X_1|^{u_1}\mathbb{E}^{1/v_1u_2}|X_2|^{v_1u_2}\mathbb{E}^{1/v_1v_2}|X_3|^{v_1v_2},
\end{aligned}$$

where the first inequality is due to Minkowski, and the last two due to Hölder's inequality

$$\|fg\|_1 \le \|f\|_p \|g\|_g$$

for $1 \le p, q \le \infty$ and $1/p + 1/q = 1$. The p-norm is defined in $L_2(\Omega, \mathcal{F}, \mathbb{P})$ as follows

$$\|f\|_p := \left(\int_\Omega |f|^p d\mathbb{P}\right)^{\frac{1}{p}} = \mathbb{E}^{\frac{1}{p}}|f|^p.$$

Since we have

$$\begin{aligned}
|(X_{t_i}(\mathbb{E}Y_{t_{i+1}}|\mathcal{F}_1^{t_i})_c| &= |X_{t_i}(\mathbb{E}Y_{t_{i+1}}|\mathcal{F}_1^{t_i})_c - \mathbb{E}[X_{t_i}(\mathbb{E}Y_{t_{i+1}}|\mathcal{F}_1^{t_i})_c]| \\
&\le 2|X_{t_i}(\mathbb{E}Y_{t_{i+1}}|\mathcal{F}_1^{t_i})_c|,
\end{aligned}$$

which follows from Jensen's inequality, we can conclude

$$\begin{aligned}
&|\mathbb{E}_c X_{t_1} \dots X_{t_k}| \le 2^{i-1}\mathbb{E}^{1/u_1}|X_{t_1}|^{u_1}\mathbb{E}^{1/(v_1u_2)}|X_{t_2}|^{v_1u_2}\cdot \\
&\quad \cdot \mathbb{E}^{1/(v_1\cdots v_{i-2}u_{i-1})}|X_{t_{i-1}}|^{v_1\cdots v_{i-2}u_{i-1}}\mathbb{E}^{v_1\cdots v_{i-2}v_{i-1}}|X_{t_i}(\mathbb{E}[Y_{t_{i+1}}|\mathcal{F}_1^{t_i}])_c|^{v_1\cdots v_{i-2}v_{i-1}},
\end{aligned} \tag{5.7}$$

where $1/u_j + 1/v_j = 1, \quad u_j,\ v_j \ge 1,\ j = 1, \dots, i,\ i = 1, \dots, k-1$. Since the r.v. X_t, $t = t_1, \dots, t_k$ are bounded by C with probability 1, we obtain

$$|\mathbb{E}_c X_{t_1} \dots X_{t_k}| \le 2^{i-1}C^{i-1}|X_{t_i}(\mathbb{E}[Y_{t_{i+1}}|\mathcal{F}_1^{t_i}] - \mathbb{E}Y_{t_{i+1}})|,$$

where we have used the tower property of the conditional expectation to obtain $\mathbb{E}[\mathbb{E}[Y_{t_{i+1}}|\mathcal{F}_1^{t_i}]|\mathcal{F}_0] = \mathbb{E}Y_{t_{i+1}}$, where $\mathcal{F}_0 = \{\{\emptyset\}, \Omega\}$.
From (5.4) we conclude, under the assumption of the theorem, that $|Y_{t_{i+1}}| \leq 2^{k-i-1}C^{k-1}$ with probability 1. The proof is completed by applying the inequality (A).

The assumption in the last theorem is quite a stringent one, not even a Gaussian stochastic process fulfills this restriction. Above we also cited another upper bound found by Davydov, which does not use this strong assumption that $|X_{t_j}| \leq C$ with probability 1. Now we will present another theorem which gives us an upper bound for the k-th order centered moment of a random process with mixing. It only requires the existence of the p_j-th absolute moment for some collection $p_j \geq 1$ such that the sum of the inverses of the p_j is less or equal to one. The following theorem and its proof is also from [24].

Theorem 5.2 *If for some collection $p_j \geq 1$, $j = 1, \ldots, k$, such that*

$$\sum_{j=1}^{k} \frac{1}{p_j} \leq 1, \quad k = 2, 3, \ldots,$$

there exist $\mathbb{E}|X_{t_j}|^{p_j}$, $j = 1, \ldots, k$, then for all $i = 1, \ldots, k-1$

$$|\mathbb{E}_c X_{t_1} \ldots X_{t_k}| \leq 3 \cdot 2^{k-1} \alpha^{1-\sum_{j=1}^{k} \frac{1}{p_j}}(t_i, t_{i+1}) \prod_{j=1}^{k} \mathbb{E}^{\frac{1}{p_j}} |X_{t_j}|^{p_j}.$$

For the last theorem there exist also upper bounds expressed through φ and ψ with a slight modification.

Proof. To prove this theorem we will cite a lemma from [21] without proof.

Lemma 5.3 *If a r.v. Y is $\mathcal{F}_t^\infty$-measurable, then for any u and v, $1 \leq u \leq v$,*

$$\mathbb{E}^{1/u}|\mathbb{E}(Y|\mathcal{F}_1^s) - \mathbb{E}Y|^u \leq 2(1 + 2^{1/u})(\alpha(s,t))^{1/u-1/v}\mathbb{E}^{1/v}|Y|^v \tag{5.8}$$

and

$$\mathbb{E}^{1/u}|\mathbb{E}(Y|\mathcal{F}_1^s) - \mathbb{E}Y|^u \leq 2(\varphi(s,t))^{1-1/v}\mathbb{E}^{1/v}|Y|^v \tag{5.9}$$

hold.

After applying (5.8) to the inequality (5.7), we obtain

$$\begin{aligned}|\mathbb{E}_c X_{t_1}\dots X_{t_k}| \le 2^{i-1}\mathbb{E}^{1/u_1}|X_{t_1}|^{u_1}\mathbb{E}^{1/(v_1u_2)}|X_{t_2}|^{v_1u_2}\cdot\\ \cdot\mathbb{E}^{1/(v_1\cdots v_{i-1}u_i)}|X_{t_i}|^{v_1\cdots v_{i-1}u_i}\cdot 2(1+2^{1/(v_1\cdots v_i)})\cdot\\ \cdot\alpha^{1/(v_1\cdots v_i)-1/(v_1\cdots v_i(1+\varepsilon))}(t_i,t_{i+1})\cdot\\ \cdot\mathbb{E}^{1/((1+\varepsilon)v_1\cdots v_iu_{i+1})}|X_{t_{i+1}}|^{(1+\varepsilon)v_1\cdots v_iu_{i+1}}\dots\\ \cdot\mathbb{E}^{1/((1+\varepsilon)v_1\cdots v_{k-2}u_{k-1})}|X_{t_{k-1}}|^{(1+\varepsilon)v_1\cdots v_{k-2}u_{k-1}}\cdot\\ \cdot\mathbb{E}^{1/((1+\varepsilon)v_1\cdots v_{k-1})}|X_{t_k}|^{(1+\varepsilon)v_1\cdots v_{k-1}},\end{aligned} \tag{5.10}$$

where $u_j, v_j \ge 1$ and $(1/u_j)+(1/v_j)=1$ for $j=1,\dots,k-1$ and $\varepsilon\ge 0$. Now put

$$\begin{aligned}p_1 &= u_1,\\ p_2 &= v_1u_2,\\ &\dots\\ p_i &= v_1\cdots v_{i-1}u_i,\\ &\dots\\ p_{k-1} &= (1+\varepsilon)v_1\cdots v_{k-2}u_{k-1},\\ p_k &= (1+\varepsilon)v_1\cdots v_{k-1}.\end{aligned}$$

Since we now have

$$\frac{1}{v_1\cdots v_i} = 1-\sum_{j=1}^{i}\frac{1}{p_j},$$

and

$$\frac{1}{(1+\varepsilon)v_1\cdots v_i} = \sum_{j=i+1}^{k}\frac{1}{p_j},$$

we conclude

$$\frac{1}{v_1\cdots v_i}-\frac{1}{(1+\varepsilon)v_1\cdots v_i} = 1-\sum_{j=1}^{k}\frac{1}{p_j}$$

and thus the validity of the theorem.

Let us consider a case when the variables X_t are related to a Markov chain ξ_t. We call a random variable X_t related to a Markov chain if X_t can be written as $X_t = g_t(\xi_t)$, where $g_t(x)$ is a measurable function for each t. This concept a random variable being related to a Markov chain is also called a Hidden Markov Model (HMM). The reason for this nomenclature is somehow obvious. The Hidden Markov Model is a finite set of states, each of which

is associated with a (generally multidimensional) probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are "hidden" to the outside; hence the name Hidden Markov Model. The Markov property says that

$$\mathbb{P}_{t+1,t}(x, A) = \mathbb{P}(\xi_{t+1} \in A \big| \xi_t = x), \quad \mathbb{P}_t(A) = \mathbb{P}(\xi_t \in A)$$

must hold. In other words, the outcome of the random variable ξ_{t+1} depends only on its value (state) at the time point t, i.e.

$$\mathbb{P}(\xi_{t+1} \in A | \mathcal{F}_1^t) = \mathbb{P}(\xi_{t+1} \in A | \mathcal{F}_t^t),$$

where $\mathcal{F}_s^t = \sigma\{\xi_u,\ s \le u \le t\}$. Then

$$\varphi(s,t) = \sup_{x, A \in \mathcal{F}_s^t} |\mathbb{P}_{s,t}(x, A) - \mathbb{P}_t(A)| \le 1 - \alpha_{s,t},$$

where $\alpha_{s,t}$ is the ergodicity coefficient

$$\alpha_{s,t} = 1 - \sup_{x, y, A \in \mathcal{F}_s^t} |\mathbb{P}_{s,t}(x, A) - \mathbb{P}_{s,t}(y, A)|$$

(see Dobrushin [7],[8]) Let $\alpha^{(n)} = \min_{1 \le s < n} \alpha_{s,s+1}$ be the ergodicity coefficient of the chain. It is known that $1 - \alpha_{s,t} \le (1 - \alpha^{(n)})^{t-s} \le \exp\{-\alpha^{(n)}(t-s)\}$ for all $1 \le s \le t \le n$.
The next theorems and corollaries concerning a stochastic process X_t related to a Markov chain ξ_t will be stated without proof due to their length and since they do not contain any new ideas. The proofs are presented in [24], chapter 4. We introduce time indices l_j, $j = 1, \ldots, r$, where $r \le k$, as being those t_i in the sequence of time indices $t_1 \le t_2 \le \cdots \le t_k$ for which the strict inequality holds. They bound the centered moment of $X_{t_1}, \ldots, X_{t_k}$ with the mixing functions depending only on those t_j that differ from t_{j+1} and t_{j-1}.

Theorem 5.4 *Let X_t be related to a Markov chain ξ_t. If $|X_{l_j}| \le C$ with probability 1, $j = 1, \ldots, r$, $r = 2, 3, \ldots$, then*

1)

$$|\mathbb{E}_c X_{t_1} \cdots X_{t_k}| \le 2^{k-1}\ C^k \prod_{j=1}^{r-1} \varphi(l_j, l_{j+1}),$$

2)

$$|\mathbb{E}_c X_{t_1} \cdots X_{t_k}| \le 2^{k-1}\ C^k \prod_{j=1}^{r-1} \psi(l_j, l_{j+1}).$$

Theorem 5.5 *Let X_t be related to a Markov chain ξ_t. If for some collection $q_j \ge 1$, $j = 1, \ldots, r$, $r = 2, 3, \ldots$ such that $\sum_{j=1}^{r} \frac{1}{q_j} = 1$, and if there exist $\mathbb{E}|X_{l_j}|^{m_j q_j}$, $j = 1, \ldots, r$, then*

$$|\mathbb{E}_c X_{t_1} \cdots X_{t_k}| \le 2^{k-1} \prod_{j=1}^{r-1} \varphi^{\sum_{i=1}^{j} \frac{1}{q_j}}(l_j, l_{j+1}) \prod_{j=1}^{r} \mathbb{E}^{\frac{1}{q_j}} |X_{l_j}|^{m_j q_j}.$$

Corollary 5.6 *Let X_t be related to a Markov chain ξ_t. If for some $\gamma_1 \ge 0$, $H_1 > 0$*

$$\mathbb{E}|X_{l_j}|^k \le (k!)^{1+\gamma_1} H_1^k, \quad j = 1, \ldots, r,\ r = 2, 3, \ldots,\ k = 2, 3, \ldots,$$

then for any $\delta \ge 0$

$$|\mathbb{E}_c X_{t_1}, \ldots, X_{t_k}| \le 2^{k-1} (k!)^{1+\gamma_1} \widehat{(1+\delta)}^{(1+\gamma_1)k} H_1^k \prod_{j=1}^{r-1} \varphi^{\frac{\delta}{1+\delta}}(l_j, l_{j+1}),$$

where $\hat{u} = \min\{v \ge u | v \text{ is integer}\}$. Analogously a similar result can be obtained for $\psi(l_j, l_{j+1})$.

5.2 Bounds of mixed cumulants of random processes with mixing

After being able to bound $\mathbb{E}_c X_{t_1} \ldots X_{t_k}$ from above and having available Lemma 2.3 as well as taking into account the behaviour of $N_\nu(I_1, \ldots, I_\nu)$, we obtain the bounds for the mixed cumulants $\Gamma(X_{t_1}, \ldots, X_{t_k})$.

Theorem 5.7 *If $|X_{t_j}| \le C$ a.s., $j = 1, \ldots, k$, $k = 2, 3, \ldots$, then for all $i = 1, 2, \ldots, k-1$*

1)

$$|\Gamma(X_{t_1}, \ldots, X_{t_k})| \le (k-1)!\ 2^k\ C^k\ \alpha(t_i, t_{i+1}),$$

2)

$$|\Gamma(X_{t_1}, \ldots, X_{t_k})| \le (k-1)!\ 2^{k-1} C^k \varphi(t_i, t_{i+1}),$$

3)

$$|\Gamma(X_{t_1}, \ldots, X_{t_k})| \le (k-1)!\ 2^{k-2} C^k \psi(t_i, t_{i+1}).$$

As we have already mentioned, the knowledge of the structure of $N_\nu(I_1,\dots,I_\nu)$ will be needed. In the proof of this theorem we will need the results of Lemma 2.3, namely (2.10), (2.11) and the inequality $0 \le N_\nu(I_1,\dots,I_\nu) \le (\nu-1)!$. Let us define the operation $[\mathcal{A}]_I$ for an arbitrary $\mathcal{A}$ and the set I:

$$[\mathcal{A}] := [\mathcal{A}]_I = \{t \in I | a_1 \le t \le a_s\} = [a_1, a_s] \cap I.$$

Now let $\{\mathcal{A}_1,\dots,\mathcal{A}_\mu\}$ be a system of subsets of the set I. We say that the system $\{\mathcal{A}_1,\dots,\mathcal{A}_\mu\}$ essentially covers the point $t \in I$, if

$$\{q | t \in [\mathcal{A}_q \backslash \{t\}],\ 1 \le q \le \mu\} \ne \emptyset.$$

In other words, there must exist $\mathcal{A}_p \in \{\mathcal{A}_1,\dots,\mathcal{A}_\mu\}$ such that $t \in [\mathcal{A}_p \backslash \{t\}]$. This is the right place to bring an example to clarify what we mean. Let $\mu = 8$ and $\{\mathcal{A}_1,\dots,\mathcal{A}_\mu\}$ be the partition $\{I_1,\dots,I_8\}$ of the set $\{t_1,\dots,t_{14}\}$ with $I_1 = \{t_1, t_5\}$, $I_2 = \{t_2, t_9\}$, $I_3 = \{t_3, t_6\}$, $I_4 = \{t_4\}$, $I_5 = \{t_7\}$, $I_6 = \{t_8, t_{13}\}$, $I_7 = \{t_{10}, t_{14}\}$ and $I_8 = \{t_{11}, t_{12}\}$. If we choose $t = t_{11}$, then

$$\{q | t \in [I_q \backslash \{t\}],\ 1 \le q \le 8\} = \{6, 7\},$$

since

$$t_{11} \in [I_6 \backslash \{t_{11}\}] = \{t_8, t_9, t_{10}, t_{11}, t_{12}, t_{13}\},$$
$$t_{11} \in [I_7 \backslash \{t_{11}\}] = \{t_{10}, t_{11}, t_{12}, t_{13}, t_{14}\},$$

but not in I_8, since $[I_8 \backslash \{t_{11}\}] = [t_{12}] = \{t_{12}\}$.
The number

$$n_t(\mathcal{A}_1,\dots,\mathcal{A}_\mu) = |\{q | t \in [\mathcal{A}_q \backslash \{t\}],\ 1 \le q \le \mu\}| \tag{5.11}$$

will be called the number of maximal covering of a point $t \in I$ by the system $\{\mathcal{A}_1,\dots,\mathcal{A}_\mu\}$. In our example

$$n_{t_{11}}(I_1,\dots,I_8) = |\{6, 7\}| = 2.$$

It turns out the numbers $N_\nu(I_1,\dots,I_\nu)$ emerging in formula (2.10) can be expressed by

$$\begin{aligned} &N_1(I) = 1, \\ &N_\nu(I_1,\dots,I_\nu) = \prod_{j=2}^{\nu} n_{t_1^{(p)}}(I_1,\dots,I_\nu), \end{aligned} \tag{5.12}$$

where for a partition $\{I_1,\dots,I_\nu\}$ of a set I of cardinality k, $I_p = \{t_1^{(p)},\dots,t_{k_p}^{(p)}\}$ is the set of the p-th partition with $t_1^{(p)} \le \dots \le t_{k_p}^{(p)}$, $1 \le p \le \nu$ and

$k_1 + \cdots + k_\nu = k$. The last expression (5.12) is not yet legible enough. Thus, we will illustrate the structure of the numbers $N_\nu(I_1, \ldots, I_\nu)$ by the means of graph theory. Therefore, we will use the above mentioned example. $I_1, \ldots, I_8$ were deliberately arranged according to the increasing of the first (leftmost) elements. Assign two graphs $G^{(1)}_{\{I_1,\ldots,I_\nu\}}$ and $G^{(2)}_{\{I_1,\ldots,I_\nu\}}$ to each partition $\{I_1, \ldots, I_\nu\}$.
Graph $G^{(1)}_{\{I_1,\ldots,I_\nu\}}$ is constructed as follows: its vertices are the points of the set I, arranged according to increase, i.e. vertex 1 of the graph correspond to the point t_1, vertex 2 to the point t_2, etc. The vertices in some block of the partition are connected in pairs by arcs (oriented edges) according to increase of their numbers. If in a certain block there is only one point, then the vertex, corresponding to it, is connected by loop.
Graph $G^{(2)}_{\{I_1,\ldots,I_\nu\}}$ consists of the vertices, corresponding to the blocks of partition and numbered in order of increase of the leftmost points.The vertices of the graph i and j, for which $[I_i] \cap [I_j] \neq \emptyset$, are connected by links (nonoriented edges). Thus, we have $n_{t_1^{(2)}} = n_{t_1^{(5)}} = n_{t_1^{(6)}} = n_{t_1^{(7)}} = 1$, $n_{t_1^{(3)}} = n_{t_1^{(8)}} = 2$ and $n_{t_1^{(3)}} = 3$. Obviously

$$n_{t_1^{(p)}}(I_1, \ldots, I_\nu) \leq p - 1, \quad 2 \leq p \leq \nu,$$

due to construction. The expression (2.12), namely

$$0 \leq N_\nu(I_1, \ldots, I_\nu) \leq (\nu - 1)!$$

is a simple consequence of the last inequality, bearing in mind that (5.12) holds. Even a stronger inequality holds, namely

$$N_\nu(I_1, \ldots, I_\nu) \leq \min\{(\nu - 1)!, \lfloor k/2 \rfloor !\},$$

which can be seen from the structure of graph $G^{(1)}_{\{I_1,\ldots,I_\nu\}}$. Recall that k denotes the cardinality of I. Concerning $G^{(2)}_{\{I_1,\ldots,I_\nu\}}$, it turns out that the numbers $N_\nu(I_1, \ldots, I_\nu)$ are zero on those partitions $(I_1, \ldots, I_\nu)$ of the set I for which $G^{(2)}_{\{I_1,\ldots,I_\nu\}}$ is disconnected. Conversely, for the partitions corresponding to the connected graph $G^{(2)}_{\{I_1,\ldots,I_\nu\}}$, the numbers N_ν are strictly positive.
To prove Theorem 5.7 we will yet need some lemmas which use the notion of the Stirling number of the second kind, which we will also introduce before stating the subsequent lemmas.
Again, $\sum_{\bigcup_{p=1}^{\nu} I_p = I}$ will denote the sum over all $\nu-$ block partitions $\{I_1, \ldots, I_\nu\}$ of the set I. Any finite sequence of positive integers $k_1, \ldots, k_\nu$ will be called a decomposition of a positive integer k, if $\sum_{p=1}^{\nu} k_p = k$.

By $c(k,\nu)$ we denote the number of all such decompositions of the number k in ν components and by $\sum_{k_1+\ldots+k_\nu=k}$ the sum over all such decompositions. Then

$$c(k,\nu) = \binom{k-1}{\nu-1} = \frac{(k-1)!}{(\nu-1)!(k-\nu)!},$$

$$c(k,\nu) = \sum_{k_1+\ldots+k_\nu=k} 1.$$

To see the first equation, simply think of k unnumbered balls arranged on a line. Every ball represents the number one. Now you have $\nu-1$ bars to set somewhere between the balls to obtain a partition of ν blocks. Since, there is no need to set a bar before the first nor after the last ball, you can choose between $k-1$ positions. Thus, you have $\binom{k-1}{\nu-1}$ possibilities to do so.
The number $s(k,\nu)$ of ways of partitioning a k-element set into ν nonempty subsets is called the Stirling number of the second kind. They can be represented in various kinds:

$$s(k,\nu) = |\{I_1,\ldots,I_\nu\}|,$$

$$s(k,\nu) = \sum_{\bigcup_{p=1}^{\nu} I_p = I} 1,$$

$$s(k,\nu) = \sum_{k_1+\ldots+k_\nu=k} \frac{k!}{k_1!\ldots k_\nu!\nu!}.$$

There is a connection between the Stirling numbers of the second kind $s(k,\nu)$ and the coefficients in the expansion of x^k in the basis $(x)_1$, $(x)_2,\ldots,(x)_k$, where $(x)_k = x(x-1)\cdots(x-k+1)$, namely they are the same, i.e.

$$x^k = \sum_{\nu=0}^{k} s(k,\nu)(x)_\nu.$$

Now we can state the lemma.

Lemma 5.8

$$\sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1,\ldots,I_\nu) = \sum_{j=0}^{\nu-1} (-1)^j \binom{k-\nu+j}{k-\nu} (\nu-j-1)!\; s(k,\nu-j) \tag{5.13}$$

Proof. The proof is elementary, since it uses only basic methods of combinatorics, see [24].

Denote

$$N(k,\nu) = \sum_{\substack{\nu \\ \bigcup_{p=1} I_p = I}} N_\nu(I_1,\dots,I_\nu).$$

In [26] Statulevicius and Jakimavicius proved that from (5.13) it follows that

$$\sum_{\nu=1}^{k} N(k,\nu) = (k-1)!. \tag{5.14}$$

Besides, (5.14) expresses also the well-known fact that the cardinality of the set of permutations of the set $\{1,\dots,k-1\}$ is $(k-1)!$ After stating the next lemma, the assertion of Theorem 5.7 should be clear.

Lemma 5.9

$$|\Gamma(X_{t_1},\dots,X_{t_k})| \le (k-1)! \max_{1\le\nu\le k} \prod_{p=1}^{\nu} |\mathbb{E}_c X_{I_p}|. \tag{5.15}$$

Proof. The first inequality in the sequel is simply the triangle inequality applied upon (2.10).

$$|\Gamma(X_{t_1},\dots,X_{t_k})| \le \sum_{\nu=1}^{k} \sum_{\substack{\nu \\ \bigcup_{p=1} I_p = I}} N_\nu(I_1,\dots,I_\nu) \prod_{p=1}^{\nu} |\mathbb{E}_c X_{I_p}| \le$$

$$\le \max_{1\le\nu\le k} \prod_{p=1}^{\nu} |\mathbb{E}_c X_{I_p}| \sum_{\nu=1}^{k} \sum_{\substack{\nu \\ \bigcup_{p=1} I_p = I}} N_\nu(I_1,\dots,I_\nu) \le (k-1)! \max_{1\le\nu\le k} \prod_{p=1}^{\nu} |\mathbb{E}_c X_{I_p}|.$$

Proof of Theorem 5.7 The proof of Theorem 5.7 follows directly from Theorem 5.1 and Lemma 5.9.

5.3 Bounds of cumulants of sums of dependent random variables

As before, $S_n = \sum_{t=1}^{n} X_t$ and $\Gamma_k(S_n)$ denote the k-th order cumulant of the sum S_n and

$$\Lambda_n(f,u) := \max\{1, \max_{1\le s\le n} \sum_{t=s}^{n} f^{1/u}(s,t)\},$$

where $f(s,t)$ is one of the mixing functions α, φ or ψ and $u > 0$. We need the next theorems for proving the theorems and inequalities of large deviations

for sums of dependent random variables. We omit the proofs due to the length of their proofs and due to the lack of new ideas, which again can be found in [24]. Thus, the proofs are essentially based on the ideas that we have already presented in this chapter.

Theorem 5.10 *If* $|X_t| \le C$ *with probability* 1, $t = 1, 2, \ldots, n$, *then for all* $k \ge 3$, $\beta > 0$, $\delta > 0$

1)

$$|\Gamma_k(S_n)| \le 2k!\ 8^{k-1}\ C^k \Lambda_n^{k-1}(\alpha, (k-1)n),$$

2)

$$|\Gamma_k(S_n)| \le k!\ 8^{k-1}\ C^{k-2}\Lambda_n^{k-2}(\varphi, (1+\beta)(1+1/\delta)(k-2))\cdot \\ \cdot \sum_{1\le s\le t\le n} \varphi^{\frac{\beta\delta}{(1+\beta)(1+\delta)}}(s,t)\mathbb{E}^{\frac{\delta}{1+\delta}}|X_s|^{1+\frac{1}{\delta}}\mathbb{E}^{\frac{1}{1+\delta}}|X_t|^{1+\delta}.$$

Theorem 5.11 *If for some* $k \in \{2, 3, \ldots\}$ *and* $\delta > 0$ *there exist* $\mathbb{E}|X_t|^{(1+\delta)k}$, $t = 1, 2, \ldots, n$, *then for all* $\beta > 0$

1)

$$|\Gamma_k(S_n)| \le 2k!\ 12^{k-1}\Lambda_n^{k-1}(\alpha, (1+1/\delta)(k-1)) \max_{1\le t\le n} \mathbb{E}^{\frac{1}{1+\delta}}|X_t|^{(1+\delta)k} \cdot n,$$

2)

$$|\Gamma_k(S_n)| \le k!\ 8^{k-1}\Lambda_n^{k-2}(\varphi, (1+\beta)(1+1/\delta)(k-2)) \max_{1\le t\le n} \mathbb{E}^{\frac{k-2}{(1+\delta)k}}|X_t|^{(1+\delta)k}\cdot \\ \cdot \sum_{1\le s\le t\le n} \varphi^{\frac{\beta\delta}{(1+\beta)(1+\delta)}}(s,t)\mathbb{E}^{\frac{1}{(1+\delta)k}}|X_s|^{(1+\delta)k}\mathbb{E}^{\frac{1}{(1+\delta)k}}|X_t|^{(1+\delta)k}.$$

Theorem 5.12 *If for some* $\gamma_2 \ge 0$, $H_2 > 0$

$$|\mathbb{E}(X_t^k|\mathcal{F}_1^{t-1})| \le (k!)^{1+\gamma_2}H_2^k \quad \textit{with probability } 1$$

for $t = 1, \ldots, n$ *and* $k \ge 2$ *, then*

$$|\Gamma_k(S_n)| \le 2\ (k!)^{1+\gamma_2}16^{k-1}H_2^k\Lambda_n^{k-1}(\alpha, k-1)\cdot n.$$

Theorem 5.13 *Let* X_t *be related to a Markov chain* ξ_t. *If* $|X_t| \le C$ *with probability 1,* $t = 1, \ldots, n$, *then for all* $k = 2, 3, \ldots$, $\delta > 0$

1)

$$|\Gamma_k(S_n)| \le k!\ 8^{k-1}\ C^k \Lambda_n^{k-1}(\varphi,1)n,$$

2)

$$|\Gamma_k(S_n)| \le k!\ 8^{k-1}\Lambda_n^{k-2}(\varphi,1+1/\delta)\cdot \\ \cdot \sum_{1\le s\le t\le n} \varphi^{\frac{\delta}{(1+\delta)}}(s,t)\mathbb{E}^{\frac{\delta}{1+\delta}}|X_s|^{1+\frac{1}{\delta}}\mathbb{E}^{\frac{1}{(1+\delta)}}|X_t|^{1+\delta}.$$

Theorem 5.14 *Let X_t be related to a Markov chain ξ_t. If for some $k \in \{2,3,\ldots\}$ and $\delta > 0$ there exist $\mathbb{E}|X_t|^{(1+\delta)k}$, $t = 1,2,\ldots,n$, then*

1)

$$|\Gamma_k(S_n)| \le k!\ 8^{k-1}\Lambda_n^{k-1}(\varphi,1+1/\delta)\max_{1\le t\le n}\mathbb{E}^{\frac{1}{1+\delta}}|X_t|^{(1+\delta)k}\cdot n,$$

2)

$$|\Gamma_k(S_n)| \le k!\ 8^{k-1}\Lambda_n^{k-2}(\varphi,1+1/\delta)\max_{1\le t\le n}\mathbb{E}^{\frac{k-2}{(1+\delta)k}}|X_t|^{(1+\delta)k}\cdot \\ \cdot \sum_{1\le s\le t\le n} \varphi^{\frac{\delta}{(1+\delta)}}(s,t)\mathbb{E}^{\frac{1}{(1+\delta)k}}|X_s|^{(1+\delta)k}\mathbb{E}^{\frac{1}{(1+\delta)k}}|X_t|^{(1+\delta)k}.$$

Theorem 5.15 *Let X_t be related to a Markov chain ξ_t. If for some $\gamma_2 \ge 0$, $H_2 > 0$*

$$|\mathbb{E}(X_t^k|\mathcal{F}_1^{t-1})| \le (k!)^{1+\gamma_2}H_2^k \quad \textit{with probability } 1$$

for $t = 1,\ldots,n$ and $k \ge 2$, then

$$|\Gamma_k(S_n)| \le 2\ (k!)^{1+\gamma_2}16^{k-1}H_2^k\Lambda_n^{k-1}(\varphi,1)n.$$

5.4 Theorems and inequalities of large deviations for sums of dependent random variables

The bounds, we stated in chapter 5.3 and the Lemma 4.2 opens us the way to state theorems and inequalities of large deviations for the distribution $\mathbb{P}(Z_n \ge x)$ of the normed sum $Z_n = S_n/B_n$, $B_n^2 = \mathbb{E}S_n^2$ (everywhere $\mathbb{E}X_t = 0$, $t = 1,\ldots,n$). The theorems in the sequel will be stated only in the case of a stationary process X_t, $t = 1,2,\ldots$ out of practical reasons, namely to avoid cumbersome expressions in the proofs. In the case of a general non

stationary sequence in theorems in chapter 5.3 it is better to bound the k-th cumulant of the sum S_n with the help of $\Lambda_n^{k-2} L_{k,n}$, instead of

$$\Lambda_n^{k-2} \max_{1 \le t \le n} \mathbb{E}|X_t|^k / B_n^k,$$

where

$$L_{k,n} = \sum_{t=1}^{n} \mathbb{E}|X_t|^k / B_n^k.$$

Before stating the theorems we will introduce the notion of m-dependence, which will be needed in the last two theorems. The definition of m-dependent strictly stationary processes is taken from [5], chapter six on asymptotic theory.

m-Dependence: A strictly stationary sequence (X_t) is said to be m-dependent (where m is a non-negative integer) if for each t the random vectors $(X_j,\ j \le t)$ and $(X_j,\ j \ge t+m+1)$ are independent.

Remark Since for a strictly stationary process $(X_t,\ t = 0, \pm 1, \pm 2, \ldots)$ the two infinite random vectors $(X_j,\ j \le 0)$ and $(X_j,\ j \ge m+1)$ have the same joint distribution as the random vectors $(X_j,\ j \le t)$ and $(X_j,\ j \ge t+m+1)$. In checking for m-independence, it is sufficient, thus, to check the independence of the former two.

Remark The property of m-dependence generalizes the notion of independence in a natural way. Observations of an m- dependent process are independent, if there is enough distance in time, namely more than m time units. For the special case of $m = 0$ m-dependence reduces to independence. MA(q) processes are m-dependent with $m = q$.

In the following theorems we consider a stationary process X_t with $\mathbb{E}X_1 = 0$, $\mathbb{E}X_1^2 = 1$and let there exist a $\sigma_0 > 0$ such that $B_n^2 = \mathbb{E}S_n^2 \ge \sigma_0^2 n$. The last condition requires that the variance of S_n grows at least with n.

Theorem 5.16 *If $|X_1| \le C$ with probability 1 and if*

$$\alpha(s,t) \le K_1 \exp\{-b_1(t-s)\}, \quad K_1 > 0, \quad b_1 > 0,$$

then

$$|\Gamma_k(Z_n)| \le (k!)^2 B_1 \left(\frac{8Ce}{b_1 B_n} \right)^{k-2},$$

$k \ge 2$, $B_1 = 8C^2 K \exp\{1+b_1\}/(b_1\sigma_0^2)$, $K = \max\{1, K_1\}$, and for $\xi = Z_n$ the relation of large deviations (4.4) *holds with*

$$\gamma = 1, \quad \bar{\Delta} = \frac{b_1 B_n}{8eC} \quad \textit{and} \quad H = 4B_1.$$

Theorem 5.17 *If for some $\gamma_2 \geq 0$, $H_2 > 0$*

$$|\mathbb{E}(X_t^k|\mathcal{F}_1^{t-1})| \leq (k!)^{1+\gamma_2} H_2^k \quad \textit{with probability } 1$$

for $t = 1, \ldots, n$ and $k \geq 2$, and if $\alpha(s,t) \leq K_1 \exp\{-b_1(t-s)\}$, for $K_1 > 0$ and $b_1 > 0$, then

$$|\Gamma_k(Z_n)| \leq (k!)^{2+\gamma_2} B_2 \left(\frac{16 H_2 e}{b_1 B_n}\right)^{k-2},$$

$k \geq 2$, $B_2 = 16 H_2^2 K \exp\{1+b_1\}/(b_1\sigma_0^2)$ and for $\xi = Z_n$ the relation of large deviations (4.4) *holds with*

$$\gamma = 1 + \gamma_2, \quad \bar{\Delta} = \frac{b_1 B_n}{16 e H_2} \quad \textit{and} \quad H = 2^{2+\gamma_2} B_2.$$

Theorem 5.18 *Let random variables X_t be related to a Markov chain ξ_t. If $|X_1| \leq C$ with probability 1, and if $\varphi(s,t) \leq \exp\{-b_2(t-s)\}$ for $b_2 > 0$, then*

$$|\Gamma_k(Z_n)| \leq k!\ B_3 \left(\frac{8(1+b_2)C}{b_2 B_n}\right)^{k-2},$$

$k \geq 2$, $B_3 = 8C^2(1+b_2)/(b_2\sigma_0^2)$ and for $\xi = Z_n$ the relation of large deviations (4.4) *holds with*

$$\gamma = 0, \quad \bar{\Delta} = \frac{b_2 B_n}{8(1+b_2)C} \quad \textit{and} \quad H = 2B_3.$$

Theorem 5.19 *Let random variables X_t be related to a Markov chain ξ_t. If for some $\gamma_2 \geq 0$, $H_2 > 0$*

$$|\mathbb{E}(X_t^k|\mathcal{F}_1^{t-1})| \leq (k!)^{1+\gamma_2} H_2^k \quad \textit{with probability } 1$$

for $t = 1, \ldots, n$ and $k \geq 2$, and if $\varphi(s,t) \leq \exp\{-b_2(t-s)\}$, for $K_1 > 0$ and $b_1 > 0$, then

$$|\Gamma_k(Z_n)| \leq (k!)^{1+\gamma_2} B_4 \left(\frac{16(1+b_2)H_2}{b_2 B_n}\right)^{k-2},$$

$k \geq 2$, $B_4 = 16 H_2^2 (1+b_2)/(b_2\sigma_0^2)$ and for $\xi = Z_n$ the relation of large deviations (4.4) *holds with*

$$\gamma = \gamma_2, \quad \bar{\Delta} = \frac{b_2 B_n}{16(1+b_2)H_2} \quad \textit{and} \quad H = 2^{1+\gamma_2} B_4.$$

Theorem 5.20 *Let random variables X_t be m-dependent. If $|X_1| \le C$ with probability 1, then*

$$|\Gamma_k(Z_n)| \le k!\ B_5 \left(\frac{8(1+m)C}{B_n}\right)^{k-2}, \quad k \ge 2,$$

$B_5 = 16C^2(1+m)/\sigma_0^2$ and for $\xi = Z_n$ the relation of large deviations (4.4) is valid with

$$\gamma = 0, \quad \bar{\Delta} = \frac{B_n}{8(1+m)C} \quad \textit{and} \quad H = 2B_5.$$

Theorem 5.21 *Let random variables X_t be m-dependent. If for some $\gamma_2 \ge 0$, $H_2 > 0$*

$$|\mathbb{E}(X_t^k | \mathcal{F}_1^{t-1})| \le (k!)^{1+\gamma_2} H_2^k \quad \textit{with probability } 1$$

for $t = 1, \dots, n$ and $k \ge 2$, and if $\varphi(s,t) \le \exp\{-b_2(t-s)\}$, for $K_1 > 0$ and $b_1 > 0$, then

$$|\Gamma_k(Z_n)| \le (k!)^{1+\gamma_2} B_6 \left(\frac{16(1+m)H_2}{B_n}\right)^{k-2}, \quad k \ge 2,$$

$B_6 = 32H_2^2(1+m)/\sigma_0^2$ and for $\xi = Z_n$ the relation of large deviations (4.4) is valid with

$$\gamma = \gamma_2, \quad \bar{\Delta} = \frac{B_n}{16(1+m)2^{\gamma_2}H_2} \quad \textit{and} \quad H = 2^{1+\gamma_2}B_6.$$

Proof of Theorems 5.16-5.21 Theorems 5.16-5.21 are proved by direct calculating γ, $\bar{\Delta}$ and H and applying, as we have already mentioned, the results of Theorems 5.10-5.15 in Lemma 4.2. Let us notice some important consequences of the assumption $f(s,t) \le K\exp\{-b_2(t-s)\}$ for $K \ge 1$, namely

$$\begin{aligned}\Lambda_n(f,1) &\le (1 + \exp\{-b\} + \dots + \exp\{-b(t-s)\}) \le \\ &\le K/(1-\exp\{-b\}) = K(1 + 1/(\exp\{b\}-1)) \le K(1+1/b),\end{aligned}$$

$$\Lambda_n(f, k-1) \le K^{\frac{1}{k-1}}(1 + (k-1)/b),$$

$$\Lambda_n(f, 1+1/\delta) \le K^{\frac{\delta}{1+\delta}}(1 + (1+\delta)/b\delta), \quad \delta > 0,$$

$$\Lambda_n(f, (1+1/\delta)(k-1)) \le K^{\frac{\delta}{(1+\delta)(k-1)}}(1 + (1+\delta)(k-1)/b\delta), \quad \delta > 0.$$

Applying the inequality $k^k \leq k! \exp\{k\}$ to the last three inequalities and to the condition $k \geq 2$, we obtain

$$\begin{aligned}\Lambda_n^{k-1}(f,k-1) &\leq K(1+(k-1)/b)^{k-1} = \\ &= K((k-1)/b)^{k-1}(1+b/(k-1))^{k-1} \leq K(e/b)^{k-1}(k-1)!\, e^b \leq \\ &\leq k!\ (K/2b)\exp\{1+b\}(e/b)^{k-2},\end{aligned}$$

$$\begin{aligned}&\Lambda_n^{k-1}(f,(1+1/\delta)(k-1)) \leq K^{\frac{\delta}{(1+\delta)}}(1+(1+\delta)(k-1)/b\delta)^{k-1} = \\ &= K^{\frac{\delta}{(1+\delta)}}((1+\delta)(k-1)/b\delta)^{k-1}\Big(1+b\delta/\big((1+\delta)(k-1)\big)\Big)^{k-1} \leq \\ &\leq K^{\frac{\delta}{(1+\delta)}}((1+\delta)(k-1)/b\delta)^{k-1}\exp\{b\delta/(1+\delta)\} \leq \\ &\leq k!\ K^{\frac{\delta}{(1+\delta)}}((1+\delta)/2b\delta)\exp\{1+b\delta/(1+\delta)\}((1+\delta)e/b\delta)^{k-2}.\end{aligned}$$

For Theorems 5.20 and 5.21 we have to note that in the case of m-dependent random variables the inequality

$$\Lambda_n\big(\bar{m},(1+1/\delta)(k-1)\big) \leq m+1, \quad \delta > 0,$$

holds, where $\bar{m}(s,t)$ is the function of m-dependence.

5.5 A questionable generalization of Theorem 4.1

The theorems 5.16-5.21 are the fundamentals to develop further Theorem 4.1 and generalize it for stationary stochastic processes. I haven't succeeded in this task, and thus this question will stay open. The problem is to find the cumulants for the sum of random variables $V_i^2 - 1$, but where the V_i are dependent and not necessarily chi-squared distributed. This is a topic for further research.

6 Conclusion

The book [24] written by L. Saulis and V.A. Statulevicius was published in 1991, whereupon the fundamentals for limit theorems for large deviations were set even much earlier in works like [26]. Though it appears that the general theory of large deviations has become an important part of probability theory, especially in the field of finance and insurance mathematics (see [9]), where they are used to model extremal events. To mention a concrete example: they find immediate applications for the valuation of certain quantities which are closely related to reinsurance problems, see [9].
We have successfully used the theory of large deviations to show how regularized estimator of large covariance matrices converge to the population covariance matrix of multivariate normal i.i.d. stochastic processes, if the matrices are well-conditioned as long as long as $\frac{\log p}{n} \to 0$. Based on the article of Bickel and Levina, [2], we have not managed to establish a convergence result using theorems for large deviations from [24] for stationary processes and we leave this question unanswered in the hope to be solved in the future by someone. The flaws of the sample covariance matrix in case when the dimension of the random vector is bigger than the sample size are well documented in the literature and we have also illustrated the drawbacks of it. Since, we have showed that the banded estimator converges to the population covariance matrix and the Cholesky factor converges to the inverse of the population covariance matrix under certain conditions, no one should use the sample covariance matrix anymore in the case of $p > n$. The just mentioned results can bring a significant improvement in the finance industry, where it is necessary to have a reliable estimator of the population covariance matrix, especially in the field of portfolio optimization, where most often the number of assets is much larger than the number of observations. This assertion is also amplified by the numerical results conducted on simulations and real life data by Bickel and Levina in [2].

References

[1] R. Bentkus and R. Rudzkis. On exponential estimates of the distribution of random variables. *Lithuanian Mathematical Journal*, 20:15–30, 1980.

[2] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[3] Torben Maak Bisgaard and Zoltan Sasvari. *Characteristic Functions and Moment Sequences: Positive Definiteness in Probability*. Nova Science Publishers, Inc, 2000.

[4] J.R. Blum, D.L. Hanson, and L.H. Koopmans. On the strong law of large numbers for a class of stochastic processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 2:1–11, 1963.

[5] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer-Verlag, 2002.

[6] Yu.A. Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory of Probability and its Applications*, 13(4):691–696, 1968.

[7] R.L. Dobrushin. Limit laws for Markov chains. *Izvestija AN SSSR. Ser. Math.*, 17:291–330, 1953.

[8] R.L. Dobrushin. Central limit theorems for non-stationary Markov chains I,II. *Theory of Probability and its Application*, 1:72–89, 365–425, 1956.

[9] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events*. Springer-Verlag, 1997.

[10] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *J. Geophys. Res.*, 99:10143–10162, 1994.

[11] Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

[12] G.H. Golub and C.F. Van Loan. *Matrix Computations*. John Hopkins Univ.Press, 1989.

[13] J. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93:85–98, 2006.

[14] I.A. Ibragimov. Some limit theorems for strongly stationary random processes. *Sov. Math. Dokl.*, 125:711–714, 1959.

[15] M. Iosifescu. Recent advances in mixing sequences of random variables. *Proceedings of Third International Summer School on Prob.Theory and Math.Stat.*, pages 111–138, 1980.

[16] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:34–45, 1960.

[17] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

[18] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[19] V.P. Leonov and A.N. Shiryaev. On a method of calculating semi-invariants. *Theory of Probability and its Applications*, 4:319–329, 1959.

[20] V.A. Marcenko and L.A. Pastur. Distributions of eigenvalues of some sets of random matrices. *Math. USSR-Sb.*, 1:507–536, 1967.

[21] D.L. McLeish. Invariance principles for dependent variables. *Zur Wahrscheinlichkeit verwandte Gebiete*, 32:165–178, 1975.

[22] Yu.V. Prohorov and Yu. A. Rozanov. *Probability Theory.* Springer-Verlag, 1969.

[23] M.. Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings Nat. Acad. Sci. U.S.A.*, 42:43–47, 1956.

[24] L. Saulis and V.A. Statulevicius. *Limit Theorems for Large Deviations.* Kluwer Academic Publishers, 1991.

[25] J. Schur. Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, 140:1–28, 1911.

[26] V. Statulevicius and D. Jakimavicius. Estimates of cumulants and centered moments of mixing random processes I, II. *Lithuanian Mathematical Journal*, 28:112–129, 360–375, 1988.

[27] V.A. Volkonskii and Yu.A. Rozanov. Some limit theorems for random functions. i. *Theory of Probability and its Applications*, 4(2):178–197, 1959.

[28] W.B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90:831–844, 2003.