Ridge-Penalty Regularization for Kernel-CCA

Michael Reiter
Vienna University of Technology
Institute of Computer Aided Automation
PRIP
reiter@prip.tuwien.ac.at

Thomas Melzer
Vienna University of Technology
Institute of Photogrammetry and
Remote Sensing
tm@ipf.tuwien.ac.at

Abstract:
CCA and Kernel-CCA are powerful statistical tools that have been successfully employed for feature extraction. However, when working in high-dimensional signal spaces, care has to be taken to avoid overfitting. This paper discusses the influence of ridge penalty regularization on kernel-CCA by relating it to multivariate linear regression*(MLR) and* partial least squares*(PLS). Experimental results of a pose estimation task will be given.*

1 Introduction

Canonical correlation analysis (CCA) and its kernel-based generalization, *kernel-CCA*, are very powerful and versatile tools especially well suited for relating two sets of measurements. They have also been successfully used for feature extraction [6]. Like *principal components analysis* (PCA), CCA also reduces the dimensionality of the original signals, since only a few factor-pairs are normally needed to represent the relevant information; unlike PCA, however, CCA and kernel-CCA take into account the relationship between the two signals (in the correlation sense), which makes them better suited for regression tasks than PCA.

Partial least squares(PLS), *multivariate linear regression*(MLR) and *principal component regression*(PLR) are among other linear regression methods, that are related to CCA. Only MLR gives a direct solution to the linear regression problem. PCR operates via (by regressing on) the principal components of the input signal. PLS and CCA will find pairs of directions that yield maximum covariance resp. maximum correlation between the two signals. CCA, in particular has some very attractive properties (for example, unlike PLS, MLR and PCR it is invariant w.r.t. affine transformations-and thus scaling-of the input variables) and can not only be used for regression purposes but whenever we need to establish a relation between two sets of measurements.

When the number of training data N is small compared to dimensionalities p and q of the signal spaces, CCA may obtain solutions (i.e., factor pairs) with maximum correlation, whose direction is determined mainly by the noise in the training data and not by the "true" un-

derlying linear relationship between $\mathbf{x}$ and $\mathbf{y}$. This problem is known as overfitting. In object recognition applications, where the N is typically much smaller than $p+q$ the situation is even worse, since the setting becomes ill-posed and there are exits an at least $(p+q)-N$ dimensional space of canonical vectors with corresponding correlation $\rho = 1$.

This is also true for kernel-CCA, although the implicit nature of the feature mappings induced by the kernels makes it far more difficult to obtain an estimate for the capacity $(p+q)$. (In general, the dimensionality of the feature space will be considerably higher than the dimensionality of the original signal space; however, the kernel can also introduce a smoothing - i.e., regularization - effect [2].)

In [5] we have shown how to implement capacity control using *ridge*-style regularization in order to avoid overfitting. The aim of this paper is to provide a deeper understanding of the regularization effects on the CCA solution. After a brief review of kernel-CCA in section 2 we will compare in section 3 the definition of CCA with PLS and MLR and show that by ridge-regularization we can transform CCA into MLR and PLS. In section 4 we will further show experimentally that regularized kernel-CCA consistently outperforms PCA. Conclusions will be drawn in section 5.

2 Kernel-CCA

Given N pairs of mean-normalized observations $(\mathbf{x}_i^T, \mathbf{y}_i^T)^T \in \mathbb{R}^{p+q}$, and data matrices $\mathbf{X} = (\mathbf{x}_1..\mathbf{x}_N) \in \mathbb{R}^{p\times N}$, $\mathbf{Y} = (\mathbf{y}_1..\mathbf{y}_N) \in \mathbb{R}^{q\times N}$, the canonical factors are obtained as $\mathbf{w}_x^* = \mathbf{X}\mathbf{f}, \mathbf{w}_y^* = \mathbf{Y}\mathbf{g}$, whereby $\mathbf{f}$ and $\mathbf{g}$ are the stationary points of the *Rayleigh quotient*

$$\rho = \frac{\begin{pmatrix} \mathbf{f}^T & \mathbf{g}^T \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{KL} \\ \mathbf{LK} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}}{\begin{pmatrix} \mathbf{f}^T & \mathbf{g}^T \end{pmatrix} \begin{pmatrix} \mathbf{K}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{L}^2 \end{pmatrix} \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}}, \tag{1}$$

and $\mathbf{K}, \mathbf{L}$ are Gram matrices defined by $\mathbf{K}_{ij} = \mathbf{x}_i^T\mathbf{x}_j$ and $\mathbf{L}_{ij} = \mathbf{y}_i^T\mathbf{y}_j$, $\mathbf{K}, \mathbf{L} \in \mathbb{R}^{N\times N}$. Eq. 1 can be regarded as the *dual formulation* of CCA, which yields the coefficients of the linear expansions, which give us the solutions to the primal problem. ρ is known as *canonical correlation* and measures the correlation of the projections $\mathbf{w}_x^T\mathbf{X}, \mathbf{w}_y^T\mathbf{Y}$ onto the canonical factors.

The dual formulation makes it possible to compute CCA on non-linearly mapped data without actually having to compute the mapping itself. This can be done by substituting the Gram matrices by kernel matrices (see [6]).

To see the effect of the implicit non-linear mapping induced by kernel matrices, consider the example shown in figure 1. For this experiment, an image set of a toy figure (see figure 2(a))

was acquired with two varying pose parameters (pan and tilt). $\mathbf{X}$ consists of the first three eigenspace coefficients (obtained by PCA) of all images. $\mathbf{Y}$ is the set of corresponding pose parameters. The visualization of the parametric manifold [8] was obtained by plotting the projections $\mathbf{w}_x^T\mathbf{X}, \mathbf{w}_y^T\mathbf{Y}$ on the canonical vectors obtained by CCA of $\mathbf{X}$ and $\mathbf{Y}$, whereby neighboring (w.r.t. pose parameters) projections are connected. For kernel-CCA the projections of the non-linearly transformed $\mathbf{Y}$ was computed by evaluation of the same kernel function that was used for computing $\mathbf{L}$ (for details see [6]).

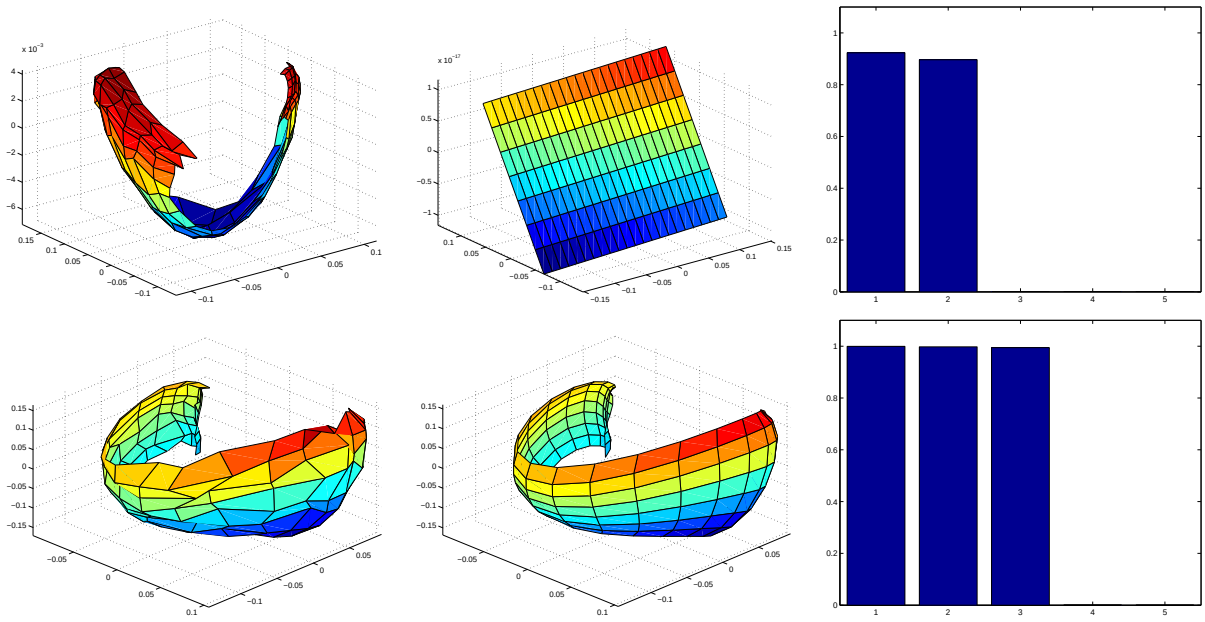

Figure 1: Projections on the canonical vectors of X (left) and Y (middle) and canonical correlations (right). The upper row shows plots for standard CCA and the lower row shows plots for kernel-CCA using an *rbf*-kernel in the output space.

3 Ridge-Penalty Regularization

Direct solutions obtained by standard CCA will explain the training data with low empirical error but with poor prediction accuracy on unseen data. As mentioned in the introduction, for standard CCA, a small ratio of $\frac{N}{p+q}$ will typically lead to solutions (factors) that exploit small deviations in the training data orthogonal to the signal variance (which are likely caused by noise). This phenomenon which arises when we try to fit an overly complex model to finite data is called overfitting. Besides this issue we also have to consider numerical problems: Unless the matrices $\mathbf{K}^2$ and $\mathbf{L}^2$ are non-singular computation of CCA becomes an ill-posed problem (this is the case for $N < p + q$). One remedy is to calculate CCA using the SVD-generalized-inverse.

An approach to dealing with singular gram matrices and to controlling complexity is to add a multiple of the identity matrix $\lambda\mathbf{I}, \lambda > 0$ to $\mathbf{K}$ and $\mathbf{L}$; this operation simply shifts the

eigenvalues by λ, and, thus, if λ is chosen large enough, will render both matrices positive definite. Also, by adding $\lambda\mathbf{I}$ the influence of directions with small eigenvalues is reduced [3].

The ridge-regularization technique, which was originally introduced to deal with singular gram matrices in the context of multivariate regression, has recently received much attention in the field of kernel-methods [4, 7]. It can be shown that ridge-penalties constrain the length of the solutions $\mathbf{w}_x$ and $\mathbf{w}_y$ whereby one can achieve complexity control. To gain a better understanding of the effect of the regularization term, we consider the standard (primal) definition of CCA

$$\rho_{CCA} = \frac{\mathbf{w}_x^T \hat{\mathbf{C}}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \hat{\mathbf{C}}_{xx} \mathbf{w}_x \mathbf{w}_y^T \hat{\mathbf{C}}_{yy} \mathbf{w}_y}}, \tag{2}$$

where $\hat{\mathbf{C}}_{xy}$ is the estimated between-set covariance matrix and $\hat{\mathbf{C}}_{xx}$, $\hat{\mathbf{C}}_{yy}$ are estimated within-set covariance matrices.

We compare Eq. 2 with the defining equations for partial least squares (PLS) and multivariate linear regression (MLR) [1]. PLS, which maximizes the covariance between $\mathbf{x}$ and $\mathbf{y}$, replaces both $\hat{\mathbf{C}}_{xx}$ and $\hat{\mathbf{C}}_{yy}$ in the denominator by the unit matrix,

$$\rho_{PLS} = \frac{\mathbf{w}_x^T \hat{\mathbf{C}}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y}}, \tag{3}$$

while MLR, which performs a least squares regression onto $\mathbf{y}$, retains the normalization by the variance of the predictor variable $\mathbf{x}$, but discards the variance-normalization w.r.t. $\mathbf{y}$ (where the square error is defined), i.e.,

$$\rho_{MLR} = \frac{\mathbf{w}_x^T \hat{\mathbf{C}}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \hat{\mathbf{C}}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{w}_y}}. \tag{4}$$

Thus, as also pointed out in [1], all three approaches effectively solve the same problem, namely maximization of the covariance, but are subject to different scalings of the variables.

As mentioned above, the regularization term $\lambda\mathbf{I}$ can be used to render singular covariance matrices positive definite. If λ is increased even further, the matrices will eventually become isotropic. Hence, for sufficiently large λ, regularized CCA becomes equivalent to PLS in the sense that both approaches will yield the same extremum points (the extremum values, however, will differ approximately by a factor $\frac{1}{\lambda}$). By the same argument, we can transform CCA into MLR; if we use different regularization parameters λ_x and λ_y for $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$, respectively, their relative magnitude determines whether (or, more precisely: to which extent) we perform a regression onto $\mathbf{x}$ or onto $\mathbf{y}$. As mentioned above solutions orthogonal to the signal variance are not always desirable; in such cases the regularization parameter λ can be used to adjust the influence of the signal variance on the solutions $\mathbf{w}_x, \mathbf{w}_y$ [3].

4 Experiments

Experiments for appearance based pose estimation were conducted on 8 test objects where the image sets were acquired with two controlled DOFs (pan and tilt). For each object, the pose parameters were in range 0 to 90 degrees (pan) and 15 to 43 degrees (tilt), resulting in 690 images per object; the images were of size 64×64.

For kernel-CCA we used an RBF-kernel with kernel width $\sigma = 54$ only for the output (pose) space, but did not kernelize the input space, i.e., evaluation of the kernel was required only for feature extraction, but not for computation of the actual features. To calculate the predicted pose values from the feature projections a pose model was constructed using a bicubic spline interpolation/resampling approach (see [8]).

A comparison of pose estimation performance of regularized kernel-CCA and PCA is given in table 1 for increasingly smaller training sets. The size of the training set (s) is given as approximate proportion of the whole data set. The number of factors (features) used was $d = 3$ for both approaches. The optimal parameters for RBF-kernel-width and regularization were determined by 2-fold cross-validation.

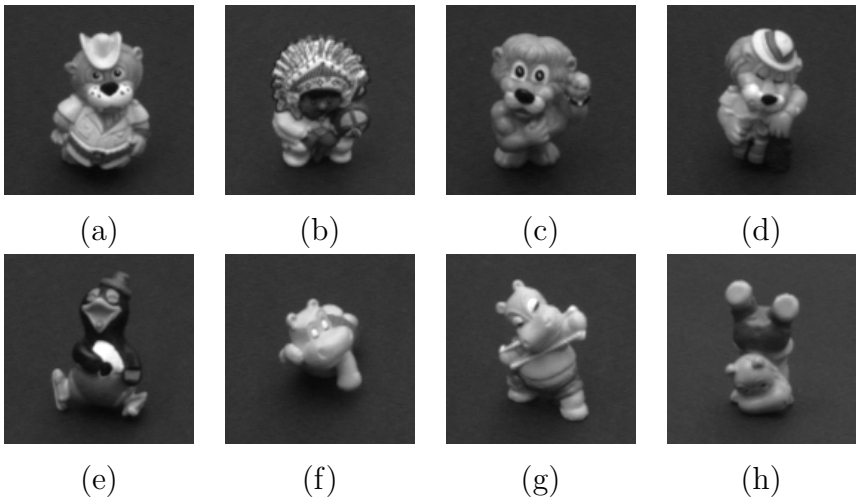

Figure 2: Example images of the objects used in the pose estimation experiments

5 Conclusion

In this paper we have used ridge-style regularization to adjust the influence of signal-variance on the solution of CCA implicitly assuming that the directions of low within-set variance are the "less informative" directions. Although the results with kernel-CCA are very promising, the influence of ridge-regularization for the non-linear case will have to be investigated more thoroughly. There are several other "shrinkage" methods that might prove useful for applying to CCA. This will be subject of future work.

object	algorithm	$s = 25\%$		$s = 11\%$		$s = 4\%$		$s = 2\%$		$s = 1\%$	
		avg	std	avg	std	avg	std	avg	std	avg	std
(a)	kCCA	0.32	0.61	0.96	1.17	2.64	2.75	3.49	3.51	8.61	7.61
	PCA	3.54	4.76	5.48	6.60	6.79	8.41	4.59	4.89	10.66	9.72
(b)	kCCA	0.22	0.60	0.65	1.02	1.91	2	2.40	1.77	3.09	2.63
	PCA	1.48	1.59	2.28	2.42	2.58	2.84	2.55	1.90	3.48	3.67
(c)	kCCA	0.32	0.57	0.83	1.02	2.68	3.36	2.08	1.74	4.48	4.31
	PCA	2.49	4.27	2.91	3.98	3.74	4.68	4.96	7	7.35	8.46
(d)	kCCA	0.26	0.45	1.36	2.28	3.82	4.01	3.99	4.09	5.62	6.64
	PCA	4.11	5.04	4.59	4.74	7.28	7.65	7	6.02	8.09	9.15
(e)	kCCA	0.14	0.28	0.70	0.66	1.58	1.25	2.05	1.32	2.79	2.16
	PCA	2.43	1.66	2.92	2.43	2.67	2.16	3.12	3.35	5.37	4.36
(f)	kCCA	0.22	0.61	0.65	1.02	1.91	2	2.40	1.77	3.09	2.63
	PCA	2.15	2.15	2.28	2.42	2.58	2.84	2.55	1.90	3.48	3.67
(g)	kCCA	0.02	0.13	0.22	0.44	1.41	1.47	1	1.20	2.56	2.10
	PCA	1.23	1.09	1.63	1.68	2.08	2.01	1.62	1.39	2.68	2.19
(h)	kCCA	0.09	0.23	0.42	0.81	1.96	2.24	1.61	1.19	2.91	3.92
	PCA	1.50	1.14	1.98	3.04	2.81	4.01	1.97	1.52	3.12	3.92

Table 1: Mean (avg) and standard deviation (std) of the pose estimation error distribution for 8 test objects using training sets of different size.

References

[1] Magnus Borga. *Learning Multidimensional Signal Processing.* Linköping Studies in Science and Technology, Dissertations, No. 531. Department of Electrical Engineering, Linköping University, Linköping, Sweden, 1998.

[2] Frederico Girosi, Micheal Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, New York, 2001.

[4] P. L. Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377, 2000.

[5] Thomas Melzer and Michael Reiter. Regularized kernel-CCA. In *Proc. of Workshop of the Austrian Association for Pattern Recognition*, pages 119–124. Österreichische Computer Gesellschaft, 2002.

[6] Thomas Melzer, Michael Reiter, and Horst Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 36:1961–1971, 2003.

[7] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing*, volume 9, pages 41–48. IEEE, 1999.

[8] H. Murase and S.K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, January 1995.