Partial Least Squares for Serially Dependent Data

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

– Doctor rerum naturalium –

der Georg-August-Universiät Göttingen

im Promotionsprogramm “Mathematik”

der Georg-August University School of Science (GAUSS)

vorgelegt von

Marco Singer

aus Braunschweig, Deutschland

Göttingen, 2016

Betreuungsausschuss

Prof. Dr. Tatyana Krivobokova, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Prof. Dr. Axel Munk, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Mitglieder der Prüfungskommission

Referentin: Prof. Dr. Tatyana Krivobokova, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Korreferent: Prof. Dr. Axel Munk, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Dominic Schuhmacher, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Prof. Dr. Gerlind Plonka-Hoch, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen

Prof. Dr. Thorsten Hohage, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen

Prof. Dr. Bert L. de Groot, Max-Planck Institut für biophysikalische Chemie, Göttingen

Tag der mündlichen Prüfung: 04.08.2016

Acknowledgements

I am grateful to my main advisor Prof. Dr. Tatyana Krivobokova who always took time to discuss problems with me and whose suggestions and ideas helped greatly in my research. I thank my second advisor Prof. Dr. Axel Munk for his help in solving some difficult problems, his advising on our papers and the engaging discussions we had.

I would also like to mention the support of my friends and colleagues Luis Haupt, Francisco Rosales, Laura Termeer, Martin Wehmeier and Jisu Yoon, some of which also helped me greatly with the proofreading of this thesis.

I would like to acknowledge the German-Swiss research group FOR916 (statistical regularization and qualitative constraints) of which I was a part of and the Deutsche Forschungsgemeinschaft (DFG) for their financial support.

Last but not least I want to thank my parents, Regine and Konrad Singer. Without them my academic career would not have been possible.

Contents

Chapter 1

Introduction

1.1 Overview

1.1.1 Motivating example

Proteins are macromolecules consisting of long chains of amino acids that occur in all types of cells in living organisms. They are responsible for many different functions, among others the transport of substances and enabling important chemical reactions.

Examples of proteins are hemoglobin, which transports oxygen in the blood stream, aquaporins that are essentially channels that control the flow of water into and out of cells and lysozymes, which are responsible for the splitting of chemical compounds with water, see, e.g., Branden and Tooze (1998) for more details on proteins and their structure.

Because proteins perform these crucial tasks the study of their biological functions is important. One approach, that we will be dealing with in this work, is the analysis of the function-dynamic relationship. It is well known that the collective motions of the atoms of a protein are important for its biological function, see Henzler-Wildman and Kern (2007). Among experimental methods like nuclear magnetic resonance spectroscopy (Mittermaier and Kay, 2006) or X-ray crystallography (Bourgeois and Royant, 2005) computational methods like molecular dynamics simulations have become crucial tools for the analysis of this relationship (Berendsen and Hayward, 2000).

We concentrate on data acquired by the latter method, i.e., we deal with the simulated dynamics of proteins by the Gromacs software (Abraham et al., 2014). Typical functions of proteins like the opening of channels or the changing geometry of binding sites where chemical reactions take place happen over small time frames of femto- or picoseconds (Tuckerman et al., 1991). This necessitates the gathering of a large number of observations $n \in \mathbb{N}$ over very small intervals of time, making molecular dynamics simulations particularly useful.

The atoms of the backbone, i.e., the longest chain of amino acids the protein consists of, are often used for this analysis. If the backbone consists of $p \in \mathbb{N}$ atoms in Cartesian coordinates $A_{t,1}, \ldots, A_{t,p} \in \mathbb{R}^3$ observed at times $t = 1, \ldots, n$ the protein dynamics are encoded in the design matrix $X = (X_1, \ldots, X_n)^{\mathrm{T}} \in \mathbb{R}^{n \times (3p)}$ that consists of $X_t = (A_{t,1}^{\mathrm{T}}, \ldots, A_{t,p}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{3p}$, $t = 1, \ldots, n$ (Brooks and Karplus, 1983).

The biological function of a protein will usually be measured by the opening area of a binding site or the distance between two (or several) groups of atoms or residues that are assumed to be responsible for its performance. These measurements at time $t = 1, \ldots, n$ will be denoted by y_t and gathered in the vector $y = (y_1, \ldots, y_n)^{\mathrm{T}}$.

Hence the question of the function-dynamic relationship can be recast into asking how X and y are related. Although the dynamics and function are explicitly available the high number of atoms and observations can be cumbersome for regression analysis. As the motions of the atoms are highly dependent in space it is clear that we encounter a collinearity problem in X, see Hub and de Groot (2009).

Furthermore not all atom motions might be important for the function of interest. If the protein has a binding site the movements of atoms surrounding this site might be more important than atoms that are farther away. One approach would be to incorporate this information into the model building process. Here we will focus on regression methods that automatically find the important information in a subspace of the column space of X.

This problem is well known in biophysics and several different approaches were used to identify the important motions. Some of the most popular techniques are principal component analysis, normal mode analysis and functional mode analysis, (Kitao and Gō, 1999; Gō and Noguti, 1983; Hub and de Groot, 2009). The first two methods find motions that have a large variance in X or that occur with a low frequency, but neglect the information present in y. Functional mode analysis seeks to find the collective motions that are highly correlated with y. Recently, Krivobokova et al. (2012) proposed the use of the partial least squares algorithm to uncover the responsible collective motions. It was seen that this method is related to functional mode analysis and partial least squares was successfully used to uncover several function-dynamic relationships. This is the main motivation for the topic of this thesis and the analysis of the partial least squares algorithm.

The other motivation is the fact that $\{X_t\}_{t=1}^n$ are, as motions of atoms over time, inherently highly dependent and the need for methods that can deal with these types of data arises. It is well known that the dynamics of proteins have long autocorrelations that decay slowly (Nadler et al., 1987). In Alakent et al. (2004) autoregressive integrated moving average time series were used to model and analyze them. This lead to the interest in studying how the partial least squares algorithm performs when the data are time series with (possibly) long autocorrelations.

1.1.2 Regularized regression

Regularized regression is an important topic in modern statistics. For illustration purposes we consider the fixed design regression problem

$$y = X\beta + \varepsilon, \tag{1.1}$$

with $X \in \mathbb{R}^{n\times d}$, $\beta \in \mathbb{R}^d$ and ε is an n-dimensional random vector with independent and identically distributed components. We assume throughout this chapter that $n \geq d$, i.e., we have more observations than variables. Assuming that the columns of X have mean zero and that y is centred we denote the sample covariance matrix with $A = n^{-1}X^{\mathrm{T}}X$ and the cross covariance with $b = n^{-1}X^{\mathrm{T}}y$. The ordinary least squares estimator $\widehat{\beta}_{OLS} = A^{-1}b$ is the minimizer (in β) of the squared Euclidean distance between y and $X\beta$. This estimator is unbiased and has several other important properties, see, e.g., Rao and Toutenburg (1999), chapter 3.

On the other hand it is obvious that the variance of $\widehat{\beta}_{OLS}$ is high when A is ill-conditioned. This problem is closely related to high collinearity in the columns of X and thus A will have small eigenvalues. As was mentioned in the previous section this occurs in the modelling of protein dynamics.

This complication can lead to unstable estimates of the coefficients of β and, although the data used for model building can be estimated exactly, can lead to a poor generalization error also known as model overfitting (Hawkins, 2004).

When the quality of an estimator $\widehat{\beta}$ is measured via the mean squared error the well known bias-variance decomposition can be used to analyze its behaviour. A biased estimator can improve upon $\widehat{\beta}_{OLS}$ in this sense if the variance is significantly lowered and at the same time the bias increases only slightly.

We consider estimators of the form $\widehat{\beta}_{f_\theta} = f_\theta(A)b$ for a function $f_\theta : [0, \infty) \to \mathbb{R}$ that depends on a parameter $\theta \in \Theta \subset \mathbb{R}$. Usually f_θ is chosen such that $f_\theta(A)$ is better conditioned than A^{-1}. Here $f_\theta(A)$ is to be understood as the functional calculus of A, i.e., applying f_θ to the eigenvalues of A. Of course for $\widehat{\beta}_{OLS}$ we have $f_\theta(x) = x^{-1}$, $x > 0$. Typically f_θ has the role of a function that regularizes x^{-1} and the degree of regularization depends on the regularization parameter θ.

We will first consider linear methods, that is, f_θ does not depend on y. Among this class of methods are two of the most well known regression techniques, ridge regression and principal component regression. Partial least squares is a nonlinear regression technique and will be the focus of the next section.

Ridge regression (Hoerl and Kennard, 1970) is a biased method that is frequently used by statisticians when the regressor matrix is ill conditioned. It is also known in the literature of ill-posed problems as Tikhonov regularization (Tikhonov and Arsenin, 1977).

The regularization function is $f_\theta(x) = (x+\theta)^{-1}$, $x \geq 0$, for a parameter $\theta > 0$. For any $\theta > 0$ the matrix $A + \theta I_d$, I_d being the $d \times d$ identity matrix, is invertible. Furthermore for small θ the perturbation of the original problem might be small enough that $\widehat{\beta}_\theta^{RR} = f_\theta(A)b$ is a good estimator for β with low variance.

It can be shown that the ridge estimator is the solution to the optimization problem $\min_{v \in \mathbb{R}^d} \|Xv - y\|^2 + \theta \|v\|^2$ and thus large choices of θ shrink the coefficients towards zero. This hinders the regression estimates from blowing up like they can in the ordinary least squares estimator.

The simple description makes the theoretical analysis of ridge regression attractive. The optimality under a rotational invariant prior distribution on the coeffients β in Bayesian statistics (Frank and Friedman, 1993), make ridge regression a strong regularized regression technique when there is no prior belief on the size of the coefficients β. A major disadvantage is the need for the inversion of a $d \times d$ matrix that can be quite cumbersome if d is large. The choice of θ is crucial, see Khalaf and Shukur (2005) for an overview of approaches.

Principal component regression is a technique that is based on principal component analysis (Pearson, 1901). Let us denote the eigenvalues of A with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ and the corresponding eigenvectors $v_1, \ldots, v_d \in \mathbb{R}^d$. Denote with I the indicator function. For principal component regression the function $f_a(x) = x^{-1}I(x \geq \lambda_a)$, $x > 0$, is used with regularization parameter $\theta = a \in \{1, \dots, d\}$, i.e., all eigenvalues that are smaller than λ_a are ignored for the inversion of A. This leads to the estimator $\widehat{\beta}_a^{PCR} = f_a(A)b$, $a = 1, \dots, d$, that avoids the collinearity problem if a is not chosen too large.

The principal component regression estimators can also be written as $\widehat{\beta}_a^{PCR} = W_a(W_a^{\mathrm{T}}AW_a)^{-1}W_a^{\mathrm{T}}b$. The matrix $W_a = (w_1, \dots, w_a) \in \mathbb{R}^{d \times a}$ is calculated as follows. In the first step the aim is to find a vector $w_1 \in \mathbb{R}^d$ that maximizes the empirical variance of Xv_1 and has unit norm, yielding $w_1 = v_1$. Subsequent principal component vectors are calculated in the same way under the additional constraint that they are orthogonal to $w_1, \dots, w_{i-1}$. This gives $w_i = v_i$. See Jolliffe (2002) for details on the method.

Thus principal component regression also solves the problem of dimensionality reduction as we restrict our estimator to the space spanned by the first a eigenvectors. These eigenvectors are the ones that contribute most to the variance in X. For proteins this corresponds to the largest collective motions. To compute the principal component estimator it is necessary to calculate the first a eigenvectors of the matrix A, which, similar to the inversion in ridge regression, can be time intensive for large matrices. The number of used eigenvalues is crucial for the regularization properties of principal component regression and there are several ways to choose them, e.g., cross validation or the explained variance in the model.

We will mention some other methods, that are not necessarily linear in y, only shortly: the least absolute shrinkage and selection operator (Tibshirani, 1996), factor analysis (Gorsuch, 1983), least angle regression (Efron et al., 2004) and variable subset selection (Guyon and Elisseeff, 2003), to name only a few. We refer to Hastie et al. (2009) for an overview of the mentioned as well as other regularized regression methods.

1.1.3 Partial least squares

Partial least squares has developed into a wide array of methods that deal with finding relationships between a response and a regressor. Usually latent variable models are considered where not all the information of the regressor X is useful for the prediction of y. Instead there is a subset of vectors in the column space of X that contains all the information. More precisely, the considered (multivariate) models are of the form

$$X = NP^{\mathrm{T}} + X_0, \quad Y = NQ^T + Y_0, \tag{1.2}$$

with the latent variables $N \in \mathbb{R}^{n \times l}$ connecting X and Y, the X-loadings $P \in \mathbb{R}^{d \times l}$, the Y-loadings $Q \in \mathbb{R}^{e \times l}$ and $X_0 \in \mathbb{R}^{n \times d}$, $Y_0 \in \mathbb{R}^{n \times e}$. We take $l \leq d$ the number of latent variables and $e \in \mathbb{N}$ the number of response variables. The matrices X_0 and Y_0 are typically considered as residuals that have no meaningful information on the relationship between X and Y that is contained in N. This is enforced, e.g., by assuming that X_0 is uncorrelated to Y and Y_0 is uncorrelated to X.

In order to find these latent relationship, partial least squares regression was suggested by Wold et al. (1984) for applications in chemometrics and is based on the work of Wold (1966). Some other methods that take the basis concept of this algorithm and use it for the discovery of latent relationships are partial least squares path modelling or multi-block partial least squares, see, e.g., Vinzi et al. (2010), for an overview.

Here we will focus on the (univariate) partial least squares regression as presented in Helland (1988) with $e = 1$ in model (1.2). The main idea of partial least squares is similar to the one presented in Section 1.1.2 for principal component regression. In the first step the vector

$w_1 \in \mathbb{R}^d$ is calculated such that it maximizes the empirical covariance between Xw_1 and y and is of unit norm. Subsequent components w_i, $i = 2, \ldots, d$, are calculated in the same way with the additional restriction that they are orthogonal to $w_1, \ldots, w_{i-1}$. Thus, whereas principal component regression only considers the variance of X to construct the model, partial least squares inherently takes the response into account. The algorithm can be formulated as a two step recursion

$$\begin{aligned} w_{i+1} &= b - A\widehat{\beta}_i, \quad \widehat{\beta}_0 = 0, \\ \widehat{\beta}_i &= W_i(W_i^{\mathrm{T}} A W_i)^{-1} W_i^{\mathrm{T}} b, \end{aligned} \tag{1.3}$$

with $W_i = (w_1, \ldots, w_i)$, $i = 1, \ldots, d$. Hence the partial least squares estimator can be written in the same way as the principal component one but with different matrices W_i, $i = 1, \ldots, d$. It was established in Krämer (2007) that $W_i^{\mathrm{T}} A W_i$ is a positive definite tridiagonal matrix for $i \leq l^*$. Here l^* is what is called the number of relevant eigenvalues of A, i.e., the ones such that $\lambda_i v_i^{\mathrm{T}} b \neq 0$, see Helland (1990).

Based on the weight vectors w_i the score vectors are calculated via $t_i = P^{\perp}_{t_1,\ldots,t_{i-1}} X w_i$. Here $P^{\perp}_{t_1,\ldots,t_{i-1}}$ denotes the orthogonal projection onto $\operatorname{span}\{t_1, \ldots, t_{i-1}\}^{\perp}$. These are used to estimate the latent components N in the model (1.2), which connect X and y, see Martens and Næs (1989).

It was shown in Phatak and de Hoog (2002) that the partial least squares estimator $\widehat{\beta}_i$ solves the optimization problem

$$\widehat{\beta}_i = \arg \min_{v \in \mathcal{K}_i(A,b)} \|y - Xv\|^2, \tag{1.4}$$

with $\|\cdot\|$ denoting the Euclidean norm and $\mathcal{K}_i(A, b) = \operatorname{span}\{b, Ab, A^2 b, \ldots, A^{i-1} b\}$ being the

ith Krylov space with respect to A and b. Hence the estimator $\widehat{\beta}_i$ can be written as $\widehat{\beta}_i = q_{i-1}(A)b$ for some polynomial q_{i-1} of degree $i-1$ with random coefficients. The regularization function for partial least squares is given by $f_i = q_{i-1}$, $q_{-1} = 0$ with regularization parameter $\theta = i \in \{1, \ldots, d\}$, but this function depends on y. This shows that the partial least squares estimator is nonlinear in the response, in contrast to linear methods like ridge regression and principal component regression. For an overview of some other properties of partial least squares we refer to Rosipall and Krämer (2006).

The polynomials q_i establish the link between partial least squares and the conjugate gradient algorithm as derived by Hestenes and Stiefel (1952) applied to the normal equation $Ax = b$ for $x \in \mathbb{R}^d$, see Phatak and de Hoog (2002). In fact, if x_i denotes the conjugate gradient approximation of the solution $x \in \mathbb{R}^d$ after i steps, it holds $x_i = \widehat{\beta}_i$ if $x_0 = 0$ is chosen.

It is well known that conjugate gradient is an efficient algorithm for the solution of normal equations and is part of the wider range of Krylov subspace methods, see Golub and van Loan (1996) for details and it is a well suited algorithm to study ill-posed problems, see Hanke (1995). Computationally this efficiency is due to the fact that only multiplications of matrices and vectors are necessary in the conjugate gradient algorithm. From a theoretical perspective Krämer and Braun (2007) showed that partial least squares uses more degrees of freedom of (1.2) in each iteration than principal component regression does for the calculation of $\widehat{\beta}_i$ and $\widehat{\beta}_i^{PCR}$, respectively. In this sense partial least squares extracts more information about the regression problem in each step.

In iterative methods like principal component regression and partial least squares regularization is achieved by early stopping of the algorithm, i.e., we stop after $i \leq d$ iterations. Here we consider discrepancy principles as stopping rules. In applications discrepancy principles can be difficult to evaluate and other techniques are used, e.g., the number of iterations are derived by cross-validation or the used degrees of freedom of the model.

Discrepancy principles were introduced by Morozov (1984) for the parameter selection in Tikhonov regularization. The main idea for iterative methods like partial least squares is that the smallest $i \leq d$ is chosen such that $\|A\widehat{\beta}_i - b\| \leq \Lambda_n$, where $\{\Lambda_n\}_{n\in\mathbb{N}} \subset [0, \infty)$ is a threshold sequence that converges to zero, making the choice of regularization parameter more adaptive to the data than a-priori parameter choices. Blanchard and Krämer (2010b) use such stopping rules to derive consistency results for a type of kernel conjugate gradient algorithm and state explicit convergence rates.

The consistency of partial least squares was analyzed before in Naik and Tsai (2000) when the number of latent variables of the model is known for independent and identically distributed data without giving explicit rates.

1.1.4 Kernel partial least squares

We will now consider a nonparametric regression model of the form

$$y_t = f^*(X_t) + \varepsilon_t, \;\; t = 1, \ldots, n, \tag{1.5}$$

with $(X_t, y_t)^{\mathrm{T}}$ being independent and identically distributed as $(\tilde{X}, \tilde{y})^{\mathrm{T}}$, $f^* \in \mathcal{L}^2\left(\mathrm{P}^{\tilde{X}}\right)$ and $\varepsilon_1, \ldots, \varepsilon_n$ independent and identically distributed and independent of $X_1, \ldots, X_n$.

There are several approaches to estimate the regression function f^* in the model (1.5). If the dimension d of X_t is small interpolation by splines is often used (under additional smoothness assumptions on the target function f^*). On higher dimensional data conditions are usually imposed to negate the curse of dimensionality (Hastie and Tibshirani, 1990), e.g., f^* follows an additive model $f^*(x) = \sum_{i=1}^d f_i^*(x_i)$, $x = (x_1, \ldots, x_d)^{\mathrm{T}}$, as was done in the extension of the partial least squares algorithm to a spline setting in Krämer et al. (2010).

In the field of machine learning reproducing kernel Hilbert space methods that map the data into abstract spaces in which the nonparametric regression problem is transformed into a linear one are popular (Gyorfi et al., 2002). These are the methods we will deal with in this section.

We consider a reproducing kernel Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions on $\mathbb{R}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, i.e., the property $g(x) = \langle g, k(\cdot, x) \rangle_{\mathcal{H}}$ holds for $g \in \mathcal{H}$ and $x \in \mathbb{R}^d$.

By virtue of the generalized representer theorem of Schölkopf et al. (2001) it is known that the solution of the regularized least squares problem

$$\min_{h \in \mathcal{H}} n^{-1} \sum_{t=1}^{n} \{y_t - h(X_t)\}^2 + \xi \|h\|_{\mathcal{H}}^2 \tag{1.6}$$

with penalization parameter $\xi > 0$ has the form $f_\alpha = \sum_{t=1}^{n} \alpha_t k(\cdot, X_t)$ for some $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. We will use f_α as an approximation of f^* in $\mathcal{H}$. This leads to the problem of estimating $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$. For this purpose linear methods are applied that can be cast into the reproducing kernel Hilbert space setting, like kernel ridge regression (Saunders et al., 1998) and kernel principal component regression (Rosipal et al., 2000).

There is an extension of the partial least squares algorithm to reproducing kernel Hilbert spaces derived by Rosipal and Trejo (2001). The kernel partial least squares estimator $\widehat{\alpha}_i$, $i = 1, \ldots, n$, can be written as the solution of the optimization problem

$$\widehat{\alpha}_i = \arg \min_{v \in \mathcal{K}_i(K_n, y)} \|y - K_n v\|^2,$$

with $K_n = [k(X_t, X_s)]_{t,s=1}^n$ being the kernel matrix, see Krämer and Braun (2007) for the derivation. This is comparable to the linear partial least squares algorithm (1.4) if we write $\widehat{\beta}_i = X^{\mathrm{T}}\widehat{\alpha}_i$ and

$$\widehat{\alpha}_i = \arg \min_{v \in \mathcal{K}_i(XX^{\mathrm{T}}, y)} \|y - XX^{\mathrm{T}}v\|^2.$$

Thus in the linear case we have $K_n = XX^{\mathrm{T}}$, or, in other words, we use the kernel $k(x, y) = x^{\mathrm{T}}y$. Note that an explicit mapping of the data into $\mathcal{H}$ is not necessary due to the kernel trick (Hoffmann et al., 2008) and in the algorithm only the kernel matrix K_n is needed. This is due to an alternative representation of the partial least squares algorithm proposed in Lindgren et al. (1993) that avoids the use of $A = X^{\mathrm{T}}X$ as in (1.3) and only relies on the matrix XX^{T}, which, as demonstrated above, fits perfectly into a kernel setting.

Blanchard and Krämer (2010a) showed the universal consistency of the kernel partial least squares estimator for two different stopping rules without giving explicit convergence rates.

In the study of ill-posed problems it is well known that the convergence rates of conjugate gradient algorithms can be arbitrarily slow if no other assumptions are imposed on the target function f^*, see Hanke (1995), chapter 3. Assuming that the kernel is measurable and bounded and that the target function coincides almost surely with an element $f \in \mathcal{H}$, an a-priori condition on f and the kernel covariance operator $S : \mathcal{H} \to \mathcal{H}, g \mapsto \mathrm{E}\{g(\tilde{X})k(\cdot, \tilde{X})\}$ is given by the Hölder source condition: there exists an $u \in \mathcal{H}$ and $r \geq 1/2$, $R > 0$ such that $f = S^{r-1/2}u$ with $\|u\|_{\mathcal{H}} \leq R$.

This condition is usually interpreted as an abstract smoothness condition for f with respect to S, i.e., the higher r can be chosen the smoother the solution is in $\mathcal{H}$. See Bauer et al. (2007), Section 2.3 for more details and Flemming (2012) for alternative conditions that are used in the ill-posed problems literature.

Under a source condition convergence rates in the $\mathcal{L}^2\left(\mathrm{P}^{\tilde{X}}\right)$-norm of reproducing kernel Hilbert space methods for independent data are of order $O_p\{n^{-r/(2r+1)}\}$, see, e.g., de Vito et al. (2005) for kernel ridge regression and Blanchard and Krämer (2010b) for a kernel conjugate gradient algorithm.

It was shown in Caponnetto and de Vito (2007) that the order optimal convergence rate of kernel ridge regression for independent and identically distributed data is $O_p\{n^{-r/(2r+s)}\}$. Here $s \in (0, 1]$ is the intrinsic (effective) dimensionality parameter measuring the complexity of the data in $\mathcal{H}$. These rates are also achieved for kernel conjugate gradient in Theorem 2.2 of Blanchard and Krämer (2010b). If the parameter s is unknown and only a source condition is assumed as a-priori information on the model we get the worst case rates with respect to this parameter with $s = 1$.

1.1.5 Dependent data

The previously mentioned results dealt either with the case of fixed design (1.1) or with independent and identically distributed data (1.5). A major motivation for this work was the fact that trajectories of atoms in proteins are highly correlated over time as has been discussed in Section 1.1.1.

In this thesis we consider serial dependence in the data given by time series models. Our main focus is on the description of these processes by their autocovariance function to measure the second order dependence in the data. This is due to the fact that the autocovariance function is a popular tool in applications to study dependence and is also investigated in the dynamics of proteins, e.g., Nadler et al. (1987). On the other hand it is an easy to understand and to handle concept in the framework of time series analysis and is closely linked to the spectral density function that is crucial in the analysis of stationary time series (Priestley, 1981). Finally, under the assumption of Gaussianity, it is also all the information that is needed (assuming the data have mean zero) to study the behaviour of the whole process.

A time series $\{Z_t\}_{t\in\mathbb{Z}}$ is stationary if for all choices $h_1,\dots,h_p \in \mathbb{Z}$, $p \in \mathbb{N}$ and $h \in \mathbb{Z}$ the property $\mathrm{P}^{Z_{h_1},\dots,Z_{h_p}} = \mathrm{P}^{Z_{h_1+h},\dots,Z_{h_p+h}}$ holds. If the dependence can be characterized completely by the mean and autocovariance functions, as is the case for Gaussian time series, this can be reduced to $\mathrm{E}(Z_t) = \mathrm{E}(Z_{t+h})$ and $\mathrm{Cov}(Z_t, Z_s) = \mathrm{Cov}(Z_{t+h}, Z_{s+h})$, $t, s, h \in \mathbb{Z}$, and the covariance matrix of $\{Z_t\}_{t=1}^n$ is a symmetric Toeplitz matrix.

Examples of processes we consider are stationary autoregressive moving average time series of order (p, q), $p, q \in \mathbb{N}_0$, i.e., $Z_t = \sum_{i=1}^p \alpha_i Z_{t-i} + \sum_{i=1}^q \beta_i \nu_{t-i} + \nu_t$ for coefficients $\alpha_1,\dots,\alpha_p,\beta_1,\dots,\beta_q \in \mathbb{R}$ and independent and identically distributed innovations $\{\nu_t\}_{t\in\mathbb{Z}}$. In Brockwell and Davis (1991), Chapter 4.4, it is shown that any stationary time series with continuous spectral density can be approximated by an autoregressive moving average process, making these models very flexible.

For stationary time series the distinction between short and long range dependence has to be made. A process is called short range dependent if its autocovariance function is absolutely summable. If this is not the case we speak of long range dependence (Giraitis et al., 2012). If the data are short range dependent, many statistical properties of the time series are similar to the independent case, e.g., the sample mean and the sample variance are $\sqrt{n}$-consistent. Stationary autoregressive moving average time series are examples of short range dependent processes whose autocovariance function decays exponentially fast. See Brockwell and Davis (1991) for more details.

For long range dependent processes convergence rates usually become worse than the rates in the independent case, e.g., $O(n^{-q/2})$ for some $q \in (0,1)$, complicating the statistical analysis of these types of data (Samorodnitsky, 2007). An example of a process that exhibits long range dependence is the fractional Gaussian noise $Z_t = B_H(t) - B_H(t-1)$, $t \in \mathbb{N}$. Here $\{B_H(t), t > 0\}$ is a zero mean Gaussian process in continuous time with $\mathrm{E}[\{B_H(t) - B_H(s)\}^2] = |t-s|^{2H}$, $t, s > 0$, for some $H \in (0,1]$. This H is usually referred to as the Hurst coefficient, after Hurst (1951), and is used as a measure for the degree of long range dependence.

As the motions of atoms in proteins exhibit properties of nonstationarity we also consider integrated models. Assume $\{Z_t\}_{t\in\mathbb{Z}}$ is a stationary process. Then $X_t = \sum_{i=1}^{t} Z_i$, $t = 1, \ldots, n$, is an integrated process of order one. A simple example is the random walk if Z_t, $t \in \mathbb{Z}$, are independent and identically distributed.

In contrast to stationary time series the statistical properties can change at each point in time, e.g., $\mathrm{P}^{X_t} \neq \mathrm{P}^{X_s}$ for $t \neq s = 1, \ldots, n$, and the covariance matrix of $\{X_t\}_{t=1}^{n}$ will in general not have a Toeplitz structure. This makes the statistical treatment of integrated time series difficult and there is no reason to believe that standard statistical estimators, like the sample mean or sample variance, should converge.

If $\{Z_t\}_{t\in\mathbb{N}}$ is an autoregressive moving average time series of order (p,q) we call the corresponding integrated process autoregressive integrated moving average time series of order $(p,1,q)$. These processes are often used to model nonstationarity in the data and, as mentioned before, were also applied to model the dynamics of atoms in Alakent et al. (2004).

1.2 Papers

Here the papers that this work consists of are summarized.

1.2.1 Partial least squares for dependent data

Published in Biometrika, Singer et al. (2016)

We consider a latent variable model of the form

$$X = V(NP^{\mathrm{T}} + \eta_1 F), \ \ y = V(Nq + \eta_2 f), \tag{1.7}$$

where $V \in \mathbb{R}^{n\times n}$ is such that V^2 is a covariance matrix, as is $\Sigma^2 = PP^T + \eta_1^2 I_d \in \mathbb{R}^{d\times d}$ and $l \leq d$. The constants $\eta_1, \eta_2 \geq 0$ denote the noise level in the data. N is an $n \times l$ dimensional random matrix, F an $n \times d$ dimensional random matrix and f an n dimensional random vector. We assume that the noise F and f are independent of the latent variables N and independent of each other. The matrix V^2 is interpreted as the covariance of the observations over time. The partial least squares estimators $\widehat{\beta}_i$ estimate in this model $\beta(\eta_1) = \Sigma^{-2}Pq$, where we understand Σ^{-2} as the Moore-Penrose pseudoinverse if $\eta_1 = 0$ and $l < d$.

We derive the population partial least squares estimators and the corresponding population Krylov space of the model (1.7) and show that they are independent of the temporal covariance. In Theorem 2.1 we establish concentration inequalities for the estimators $A = n^{-1}X^{\mathrm{T}}X$ and $b = n^{-1}X^{\mathrm{T}}y$ under dependence in the data. We see that the mean squared error of A and b does not converge to zero if the ratio of Frobenius norms $\|V\|^{-2}\|V^2\|$ does not go to zero. Otherwise the estimators are consistent and converge to their population counterparts, showing that the population Krylov space can be estimated consistently.

With this result the consistency of the partial least squares estimator is proven in Theorem 2.2 when the algorithm is stopped according to a discrepancy principle. The convergence rate is $O_p(\|V\|^{-2}\|V^2\|)$ if $\|V\|^{-2}\|V^2\|$ goes to zero.

Under the assumption that $\|V\|^{-2}\|V^2\|$ does not converge to zero we prove the inconsistency of the first partial least squares estimator $\widehat{\beta}_1$ in Theorem 2.3. In Theorems 2.4 and 2.5 we consider the convergence of $\|V\|^{-2}\|V^2\|$, showing that if V^2 is the covariance matrix of a stationary time series with autocorrelation function that decays exponentially fast, the partial least squares estimator is $\sqrt{n}$-consistent. If V^2 is the covariance matrix of an integrated process we show, on the other hand, that the ratio converges to some positive constant. Hence $\widehat{\beta}_1$ will be an inconsistent estimator when this type of nonstationarity is present.

We suggest a simple modification of the partial least squares algorithm, called corrected partial least squares, to deal with this shortcoming. Using an estimator $\widehat{V}^2$ for V^2 we consider the partial least squares algorithm with $A(\widehat{V}) = n^{-1}X^{\mathrm{T}}\widehat{V}^{-2}X$ and $b(\widehat{V}) = n^{-1}X^{\mathrm{T}}\widehat{V}^{-2}y$ instead of A and b, respectively. In Theorem 2.6 we establish consistency of the corrected partial least squares estimator and show that the convergence rate depends on the rate with which V^2 can be estimated by $\widehat{V}^2$ in operator norm.

We demonstrate the validity of these results by a simulation study that incorporates several different dependence structures V^2, e.g., independent and identically distributed, autoregressive of order one and autoregressive integrated moving average of order $(1, 1, 1)$.

Finally we apply corrected partial least squares to a protein dynamics problem. The protein aquaporin is a water channel and we consider as the functional value y its opening diameter. We see that corrected partial least squares considerably improves the predictive performance over partial least squares and principal component regression. The first corrected partial least squares estimator already yields a good representation of the dynamics of the protein responsible for changes in the functional value.

1.2.2 Kernel partial least squares for stationary data

We consider the nonparametric regression problem (1.5) when $\{X_t\}_{t\in\mathbb{Z}}$ is a d-dimensional stationary time series. Let $\tilde{X}$ be a random vector that is independent of $\{X_t\}_{t\in\mathbb{Z}}$ and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ and has the same distribution as X_0. We derive properties of the kernel partial least squares estimator given the training set $\{(X_t, y_t)^{\mathrm{T}}\}_{t=1}^{n}$.

In the following we will assume that the reproducing kernel is bounded and that the target function f^* fulfils a source condition with parameter $r \geq 1/2$. In Theorem 3.1 we prove that the kernel partial least squares estimator is consistent in the $\mathcal{L}^2\left(\mathrm{P}^{\tilde{X}}\right)$-norm and the $\mathcal{H}$-norm when the algorithm is stopped early.

The convergence rate depends on two factors: the source parameter $r \geq 1/2$ and the rate at which the estimators of the kernel covariance operator $S_n g = n^{-1}\sum_{t=1}^{n} g(X_t)k(\cdot, X_t)$, $g \in \mathcal{H}$, and the kernel cross covariance $T_n^* y = n^{-1}\sum_{t=1}^{n} y_t k(\cdot, X_t)$ converge to their population counterparts in probability.

In Proposition 3.1 we investigate the set of functions for which a source condition holds. We derive upper bounds for such functions in the $\mathcal{H}$-norm that depend on the parameter r. For univariate normally distributed data and the Gaussian kernel $k(x, y) = \exp\{-s(x-y)^2\}$, $x, y \in \mathbb{R}$, $s > 0$, we give an explicit expression of these functions.

The mean squared error of S_n and $T_n^* y$ is calculated in Theorem 3.2.

Under the assumption that $\{X_t\}_{t\in\mathbb{Z}}$ is a Gaussian process we establish convergence rates for S_n and $T_n^* y$ in Theorem 3.3. These rates depend on the type of stationarity that we have: if $\{X_t\}_{t\in\mathbb{Z}}$ is a short range dependent process we get $\sqrt{n}$-consistency, but for long range dependence the convergence slows down significantly.

Corollary 3.1 combines the previously obtained results and shows that in the considered Gaussian setting the convergence rate of the kernel partial least squares estimator is sensitive to the range of the dependence in the data. The strength of dependence between observations of the process is measured by the polynomial decay of its autocorrelation function ρ. More precisely, we consider $|\rho(h)| \leq (h+1)^{-q}$, $h \in \mathbb{N}_0$, $q > 0$. The case $q > 1$ corresponds to short and $q \in (0,1)$ to long range dependence. We see that the kernel partial least squares estimator has a convergence rate in the $\mathcal{L}^2\left(\mathrm{P}^{\tilde{X}}\right)$-norm of $O_p\{n^{-r/(2r+1)}\}$, if $q > 1$. For $q \in (0,1)$ the rate is only $O_p\{n^{-qr/(2r+1)}\}$.

These theoretical results are supported by a simulation study highlighting how different dependence structures influence the convergence rate. We consider independent and identically distributed, stationary autoregressive and long range dependent data.

1.3 Conclusion

The contribution of this thesis and the papers contained within are threefold. Firstly the partial least squares algorithm was analyzed with respect to its consistency and convergence rate. Secondly the impact of serial dependence in the observations was studied, with focus on long range dependence and nonstationarity. Thirdly a modification was proposed to deal with nonstationarity in the data and was applied to the analysis of the function-dynamic relationship in proteins. In the following we will go into more detail on each of these contributions and outline some possible further research directions.

1. Convergence rates of the (kernel) partial least squares algorithm:

The statistical properties of partial least squares are not well understood, despite an increasing interest in the last decades. One of the main problems in the analysis of the algorithm is its nonlinearity in the response. The fact that the algorithm is consistent was known for some time when the data are independent and identically distributed, yet explicit convergence rates were not available even in this setting.

In the paper Singer et al. (2016) we focused on probabilistic convergence rates and established the $\sqrt{n}$-consistency of the partial least squares estimator if the data are either independent or follow a stationary process with exponentially decaying autocovariance function. These rates are obtained if the algorithm is stopped early using a discrepancy principle stopping rule. This result makes use of the link between partial least squares and the conjugate gradient algorithm, specifically the results obtained by Nemirovskii (1986).

We considered the model (1.7) in which the covariance matrix of the multivariate process $\{X_t\}_{t=1}^n$ is separable, i.e., $\mathrm{Cov}(X_{t,i}, X_{s,j}) = [\Sigma^2]_{i,j}[V^2]_{t,s}$ for $t, s = 1, \ldots, n$ and $i, j = 1, \ldots, d$. The assumption has its origin in the analysis of spatio-temporal data (Cressie and Wikle, 2011). This special covariance structure makes it possible to treat the temporal effects separately and is a reason we obtain such clear convergence rates that depend on the ratio of the Frobenius norms $\|V\|^{-2}\|V^2\|$. The fact that the population Krylov space turns out to be independent of the temporal covariance structure is thanks to this assumption as well.

An extension of our results into a nonseparable setting would be certainly interesting, but not straight forward.

For the kernel partial least squares estimator there is little research into its statistical properties. The universal consistency of the algorithm was proven in Blanchard and Krämer (2010a) without giving explicit convergence rates and the closest result to ours is the one obtained in Blanchard and Krämer (2010b) for a kernel conjugate gradient algorithm that is similar to kernel partial least squares.

We derived the consistency of the kernel partial least squares estimator in both the $\mathcal{L}^2\left(\mathrm{P}^{\tilde{X}}\right)$-norm and the $\mathcal{H}$-norm. Similar to the linear case this is possible because we stop the algorithm early. The stopping rule used for kernel partial least squares is based on the work of Hanke (1995) and is of a more complicated form than the discrepancy principles discussed before. The derivation of this result uses the connection between kernel partial least squares, kernel conjugate gradient and the theory of orthogonal polynomials. The employed techniques are similar to the ones used in Hanke (1995) and Blanchard and Krämer (2010b).

To obtain explicit rates we focus on Gaussian time series. If these time series are short range dependent, i.e., the autocovariance function is absolutely summable, we get convergence rates in the $\mathcal{L}^2\left(\mathrm{P}^{\tilde{X}}\right)$-norm for the kernel partial least squares estimator of order $O_p\{n^{-r/(2r+1)}\}$. These rates were also achieved for kernel ridge regression when only a source condition is assumed and the data are independent and identically distributed.

The best obtainable rates are $O_p\{n^{-r/(2r+s)}\}$ with $s \in (0,1]$ denoting the intrinsic dimensionality parameter as discussed in Section 1.1.4. We obtain the rate for $s = 1$, i.e., when there is no a-priori information about this parameter. Our results could be extended to include this information, but different types of concentration inequalities than the ones established here are needed for this.

2. Properties of the algorithm under long range dependence and nonstationarity:

We investigated the influence of integrated time series, which are inherently nonstationary, on the partial least squares algorithm. We found that the mean squared error of the estimators A and b does in fact not converge to zero in this situation and hence we might be unable to estimate the population Krylov spaces consistently. Furthermore we saw that the first partial least squares estimator $\widehat{\beta}_1$ is inconsistent under this specification. This result highlights the fact that ignoring strong dependencies in the observations leads to incorrect estimation.

An extension of these results would be the study of partial least squares score vectors t_i under nonstationary dependence. The scores are important for the interpretation of latent variable models.

For the kernel partial least squares algorithm we considered stationary but long range dependent observations. We measure the range of the dependence in the data with the degree $q > 0$ of the polynomial decay of the autocorrelation function of the considered Gaussian process. For $q \in (0, 1)$ we are in the situation of long range dependence and the convergence rate of the kernel partial least squares estimator in the $\mathcal{L}^2\left(\mathrm{P}^{\tilde{X}}\right)$-norm is $O_p\{n^{-qr/(2r+1)}\}$. This highlights the fact that for stable statistical results in the long range dependent situation more observations are needed than in the independent case. This is not an unexpected result, as many statistical techniques lose efficiency when long range dependence is present in the data (Samorodnitsky, 2007).

It would be interesting to extend these results to nonstationary depence structures, e.g., integrated processes. There are several technical problems with this approach. It is for example not clear how the kernel covariance operator should be defined, as $S_t g = \mathrm{E}\{g(X_t)k(\cdot, X_t)\}$, $g \in \mathcal{H}$, inherently depends on $t = 1, \dots, n$, in contrast to the stationary case. This operator is crucial for the definition of the source condition that we impose on the target function f^*.

3. Modifications of the algorithm to deal with nonstationarity:

To counter the inconsistency of the partial least squares estimator that results from nonstationary data the corrected partial least squares algorithm was suggested. The idea is to remove correlation in the data in the model (1.7) by multiplying both X and y with the inverse of $\widehat{V}$. Here $\widehat{V}^2$ is an estimator of the temporal covariance matrix V^2. We saw that the corrected partial least squares estimator is consistent if the estimator for the temporal covariance matrix is consistent in operator norm. The feasibility of this approach is again due to the fact that we are dealing with data that has a separable covariance matrix.

The corrected partial least squares algorithm was applied to analyze the function-dynamic relationship of the protein aquaporin. We found an improvement in the predictive power of the algorithm, especially in the first partial least squares components, when using corrected partial least squares compared to ordinary partial least squares and principal component regression. This improvement is especially important in the first estimator $\widehat{\beta}_1$. It corresponds to the ensemble-weighted maximally correlated mode of motion contributing most to the fluctuations in the response y (Krivobokova et al., 2012). Hence corrected partial least squares also improves upon functional mode analysis and helps in identifying relevant underlying dynamics.

A heuristic extension of corrected partial least squares is the corrected multivariate partial least squares algorithm, i.e., we have several response variables. This is the case in model (1.2) for $e > 1$. This method was already implemented and tested on some function-dynamic problems where the function is not represented by a univariate time series. The predictive power of this corrected multivariate partial least squares algorithm substantially improved upon that of ordinary multivariate partial least squares. It could be interesting to do some research in that direction, as there are currently no theoretical results on this algorithm. It has to be noted though that multivariate partial least squares does not share many properties used in the derivation of the results presented in this thesis, making any extensions not straight forward.

At the moment there are no results on a modification of the kernel partial least squares algorithm for long range dependent data. The way the dependence structure enters the algorithm nonlinearly makes this a rather difficult problem.

An interesting feature of our convergence in probability results for both partial least squares and kernel partial least squares is the fact that the convergence rates of the algorithms are based on concentration inequalities for the sample covariance matrix or sample covariance operator and the sample cross covariance. Thus it is possible to include other types of dependence structures than the ones studied in this thesis as long as concentration inequalities can be derived for the aforementioned estimators, making our results in Theorem 2.2 and Theorem 3.1 rather flexible.

The problem of dealing with long range dependence and nonstationarity in the data is of increasing interest in the statistical community as many datasets in applications exhibit these properties, e.g., the dynamics of proteins. The partial least squares algorithm is widely used, especially in the chemometrics but also the biophysics community to analyze regression problems when there is high collinearity present in the regressor matrix or a latent variable model is assumed.

This thesis made contributions to these fields, yet there are still many open questions and possible ways to extend the research presented here. Hopefully the results obtained in this thesis and the corresponding papers will spark further interest in the study of the behaviour of partial least squares when the observations are neither independent nor identically distributed, but rather time series with possible long autocorrelations.

1.4 Own Contribution

Here the contribution by M. Singer to the presented publications are summarized.

The paper “Partial least squares for dependent data” (Singer et al., 2016) is a joint work with T. Krivobokova, A. Munk and B. de Groot. The theory, implementation and simulations were done by M. Singer with some help of T. Krivobokova and A. Munk. The data analysis was done by M. Singer with the aid of B. de Groot. The writing of the paper was done by M. Singer and T. Krivobokova. His own contribution can be judged to 80%.

The paper “Kernel partial least squares for stationary data” is a joint work with T. Krivobokova and A. Munk. This paper was largely done by M. Singer, including model, theory and simulations, with input from T. Krivobokova and A. Munk.

Bibliography

Abraham, M., van der Spoel, D., Lindahl, E., Hess, B., and the GROMACS development team (2014). Gromacs user manual version 5.0.4.

Alakent, B., Doruker, P., and Camurdan, M. (2004). Time series analysis of collective motions in proteins. *J. Chem. Phys.*, 120(2):1072–1088.

Bauer, F., Pereverzev, S., and Rosasco, L. (2007). On regularization algorithms in learning theory. *J. Complexity*, 23:52–72.

Berendsen, H. and Hayward, S. (2000). Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.*, 10:165–169.

Blanchard, G. and Krämer, N. (2010a). Kernel partial least squares is universally consistent. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings*, volume 9, pages 57–64. JMLR.

Blanchard, G. and Krämer, N. (2010b). Optimal learning rates for kernel conjugate gradient regression. *Adv. Neural Inf. Process. Syst.*, 23:226–234.

Bourgeois, D. and Royant, A. (2005). Advances in kinetic protein crystallography. *Curr. Opin. Struct. Biol.*, 15:538–547.

Branden, C. and Tooze, J. (1998). *Introduction to Protein Structure*. Taylor and Francis, New York, 2 edition.

Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*. Springer, New York, 2 edition.

Brooks, B. and Karplus, M. (1983). Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.*, 80:6571–6575.

Caponnetto, A. and de Vito, E. (2007). Optimal rates for regularized least-squares algorithm. *Found. Comp. Math.*, 7:331–368.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-temporal Data*. Wiley, New Jersey, 1 edition.

de Vito, E., Rosasco, L., Caponnetto, A., de Giovanni, U., and Odone, F. (2005). Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.*, 32:407–499.

Flemming, J. (2012). Solution smoothness of ill-posed equations in Hilbert spaces: four concepts and their cross connections. *Appl. Anal.*, 91:1029–1044.

Frank, I. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Giraitis, L., Hira, L., and Surgailis, D. (2012). *Large Sample Inference for Long Memory Processes*. Imperial College Press, London, 1 edition.

Gō, N. and Noguti, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA*, 80:3696–3700.

Golub, G. and van Loan, C. (1996). *Matrix Computations*. John Hopkins University Press, London, 3 edition.

Gorsuch, R. (1983). *Factor Analysis*. Taylor and Francis, Hillsdale, 2 edition.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer, New York, 1 edition.

Hanke, M. (1995). *Conjugate Gradient Type Methods for Ill-posed Problems*. Wiley, New York, 1 edition.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, Berlin, 1 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2 edition.

Hawkins, D. (2004). The problem of overfitting. *J. Chem. Inf. Model.*, 44:1–12.

Helland, I. S. (1988). On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.*, 17(2):581–607.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.*, 17:97–114.

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450:964–972.

Hestenes, M. and Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *J. Re. Nat. Bur. Stand.*, 49:409–436.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 8:27–51.

Hoffmann, T., Schölkopf, B., and Smola, A. (2008). Kernel methods in machine learning. *Ann. Stat.*, 36:1171–1220.

Hub, J. and de Groot, B. (2009). Detection of functional modes in protein dynamics. *PLoS Comput. Biol.*, 5:1029–1044.

Hurst, H. (1951). Long-term storage capacitiy of reservoirs. *Trans. Am. Soc. Civ. Eng.*, 116:770–799.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, New York, 2 edition.

Khalaf, G. and Shukur, G. (2005). Choosing ridge parameter for regression problems. *Commun. Stat. A-Theor.*, 34:1177–1182.

Kitao, A. and Gō, N. (1999). Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.*, 9:143–281.

Krämer, N. (2007). An overview on the shrinkage properties of partial least squares regression. *Comput. Statist.*, 22:249–273.

Krämer, N., Boulesteix, A., and Tutz, G. (2010). Penalized partial least squares based on B-splines transformations. *Chemometr. Intell. Lab.*, 94:60–69.

Krämer, N. and Braun, M. L. (2007). Kernelizing PLS, degrees of freedom, and efficient model selection. In *Proceedings of the 24th International Conference on Machine Learning*, pages 441–448. ACM.

Krivobokova, T., Briones, R., Hub, J., Munk, A., and de Groot, B. (2012). Partial least squares functional mode analysis: application to the membrane proteins AQP1, Aqy1 and CLC-ec1. *Biophys. J.*, 103:786–796.

Lindgren, F., Geladi, P., and Wold, S. (1993). The kernel algorithm for PLS. *J. Chemometrics*, 7:45–59.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*. Wiley, Chichester, 1 edition.

Mittermaier, A. and Kay, L. (2006). New tools provide new insights in NMR studies of protein dynamics. *Science*, 312:224–228.

Morozov, V. (1984). *Methods for Solving Incorrectly Posed Problems*. Springer, New York, 1 edition.

Nadler, W., Brünger, A., Schulten, K., and Karplus, M. (1987). Molecular and stochastic dynamics of proteins. *Proc. Natl. Acad. Sci.*, 84:7933–7937.

Naik, P. and Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 62:763–771.

Nemirovskii, A. (1986). The regularizing properties of the adjoint gradient method in ill-posed problems. *Comput. Math. Math. Phys.*, 26:7–16.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2:559–572.

Phatak, A. and de Hoog, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *J. Chemometr.*, 16:361–367.

Priestley, M. (1981). *Spectral Analysis and Time Series, Volume 1: Univariate Time Series*. Academic Press, London, 1 edition.

Rao, C. and Toutenburg, H. (1999). *Linear Models: Least Squares and Alternatives*. Springer, New York, 2 edition.

Rosipal, R., Girolami, M., and Trejo, L. (2000). Kernel PCA for feature extraction of event-related potentials for human signal detection performance. In *Proceedings of ANNIMAB-1 Conference*, pages 321–326. Springer.

Rosipal, R. and Trejo, L. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.*, 2:97–123.

Rosipall, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. *Lecture Notes in Comput. Sci.*, 3940:34–51.

Samorodnitsky, G. (2007). *Long Range Dependence*. now Publisher, Hanover, 1 edition.

Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann Publishers.

Schölkopf, B., Herbrich, R., and Smola, A. (2001). A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer.

Singer, M., Krivobokova, T., de Groot, B., and Munk, A. (2016). Partial least squares for dependent data. *Biometrika*, 103:351–362.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58:267–288.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solution of Ill-posed Problems*. Winston and Sons, Washington, 1 edition.

Tuckerman, M., Berne, B., and Martyna, G. (1991). Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *J. Chem. Phys.*, 94:6811–6815.

Vinzi, V. E., Trinchera, L., and Amato, S. (2010). *Handbook of Partial Least Squares*. Springer, Berlin, 1 edition.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedure. In *Research papers in statistics: Festschrift for J. Neyman*, pages 411–444. Wiley.

Wold, S., Ruhe, A., Wold, H., and Dunn, I. W. (1984). The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Comput.*, 5:735–743.

Chapter 2

Partial Least Squares for Dependent Data

Partial least squares for dependent data

Marco Singer^a, Tatyana Krivobokova^a, Bert L. de Groot^b, Axel Munk^{a,b}

^{a}Institute for Mathematical Stochastics, Göttingen, Germany

^{b}Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

We consider the partial least squares algorithm for dependent data and study the consequences of ignoring the dependence both theoretically and numerically. Ignoring nonstationary dependence structures can lead to inconsistent estimation, but a simple modification leads to consistent estimation. A protein dynamics example illustrates the superior predictive power of the method.

Key words and phrases:

Dependent data, Latent variable model, Nonstationary process, Partial least squares, Protein dynamics

2.1 Introduction

The partial least squares algorithm introduced by Wold (1966) is a powerful regularized regression tool. It is an iterative technique, which is, unlike most similar methods, nonlinear in the response variable. Consider a linear regression model

$$y = X\beta + \varepsilon, \tag{2.1}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times d}$, $\beta \in \mathbb{R}^d$ and the error term ε is a vector of n independent and identically distributed random variables. To estimate the unknown coefficients β with partial least squares, a base of $i \leq d$ weight vectors $\widehat{w}_1, \dots, \widehat{w}_i$ is iteratively constructed. First, the data are centered, i.e., y and the columns of X are transformed to have mean zero. Then the first vector $\widehat{w}_1$ is obtained by maximizing the empirical covariance between Xw and y in $w \in \mathbb{R}^d$, subject to $\|w\| = 1$. Afterwards, the data are projected into the space orthogonal to $X\widehat{w}_1$ and the procedure is iterated. The ith partial least squares estimator $\widehat{\beta}_i$ for β is obtained by performing a least squares regression of y on X, constrained to the subspace spanned by the columns of $\widehat{W}_i = (\widehat{w}_1, \dots, \widehat{w}_i)$. Helland (1988) summarizes the partial least squares iterations in two steps via

$$\begin{aligned} \widehat{w}_{i+1} &= b - A\widehat{\beta}_i, \;\; \widehat{\beta}_0 = 0, \\ \widehat{\beta}_i &= \widehat{W}_i(\widehat{W}_i^{\mathrm{T}} A \widehat{W}_i)^{-1}\widehat{W}_i^{\mathrm{T}} b, \end{aligned} \tag{2.2}$$

with $b = n^{-1}X^{\mathrm{T}}y$ and $A = n^{-1}X^{\mathrm{T}}X$, under the assumption that $(\widehat{W}_i^{\mathrm{T}} A \widehat{W}_i)^{-1}$ exists. The regularisation is achieved by early stopping, that is, by taking $i \leq d$.

Alternatively, $\widehat{\beta}_i$ can be defined using the fact that $\widehat{w}_i \in \mathcal{K}_i(A, b)$, where $\mathcal{K}_i(A, b)$ is a Krylov space, that is, a space spanned by $\{A^{j-1}b\}_{j=1}^{i}$ (Helland, 1988). Then, one can define partial least squares estimators as $\widehat{\beta}_i = \arg\min_{\beta \in \mathcal{K}_i(A,b)}(y - X\beta)^{\mathrm{T}}(y - X\beta)$. There is also a direct correspondence between partial least squares and the conjugate gradient method with early stopping for the solution of $A\beta = b$.

Frank and Friedman (1993) and Farkas and Héberger (2005) find the partial least squares algorithm to be competitive with regularized regression techniques, such as principal component regression, lasso or ridge regression, in terms of the mean squared prediction error. Also, the optimal number of partial least squares base components is often much lower than that of principal components regression, as found in Almøy (1996).

Partial least squares regression has a long and successful history in various application areas, see e.g., Hulland (1999), Lobaugh et al. (2001), Nguyen and Rocke (2002). However, the statistical properties of this algorithm have been little studied, perhaps because of the nonlinearity of partial least squares estimators in the response variable. Some attempts to understand properties of partial least squares can be found in Höskuldsson (1988), Phatak and de Hoog (2002) and Krämer (2007). Their almost sure convergence was established by Naik and Tsai (2000). For kernel partial least squares, Blanchard and Krämer (2010a) obtained convergence in probability results by early stopping. For the closely linked kernel conjugate gradient algorithm, Blanchard and Krämer (2010b) established order-optimal convergence rates dependent on the regularity of the target function. Delaigle and Hall (2012) compared theoretically the population and sample properties of the partial least squares algorithm for functional data.

Regression techniques typically assume independence of responses, but this is often violated, for example, if the data are observed over time or at dependent spatial locations. We are not aware of any treatment of the partial least squares algorithm for dependent observations. In this work we propose a modification of partial least squares to deal with dependent observations and study the theoretical properties of partial least squares estimators under general dependence in the data. In particular, we quantify the influence of ignored dependence.

Throughout the paper we denote by $\|\cdot\|_{\mathcal{L}}$ the spectral and by $\|\cdot\|$ the Frobenius norm for matrices, $\|\cdot\|$ also denotes the Euclidean norm for vectors.

All proofs are given in Section 2.7.

2.2 Partial least squares under dependence

2.2.1 Latent variable model

In many applications the standard linear model (2.1) is too restrictive. For example, if a covariate that is relevant for the response cannot be observed or measured directly, so-called latent variable or structural equation models are considered (Skrondal and Rabe-Hesketh, 2006): it is assumed that X and y are linked by $l \leq d$ latent vectors and the remaining vectors in the d-dimensional column space of X do not contribute to y. This can be interpreted as if the latent components are of interest, but only X, which contains some unknown nuisance information, can be measured. Such models are relevant in modelling of chemical (Wold et al., 2001), economic (Hahn et al., 2002) and social data (Goldberger, 1972).

We consider a latent variable model with the covariates X and response y connected via a matrix of latent variables N,

$$\begin{aligned} X &= V(NP^{\mathrm{T}} + \eta_1 F), \\ y &= V(Nq + \eta_2 f), \end{aligned} \tag{2.3}$$

where $N$and F are an $n \times l$-dimensional and an $n \times d$-dimensional random matrix, respectively, and f is an n-dimensional random vector. The random elements N, F, f can have different distributions, but are independent of each other, with all entries being independent and identically distributed with expectation zero and unit variance. The matrix $P \in \mathbb{R}^{d\times l}$ and vector $q \in \mathbb{R}^{l}$ are deterministic and unknown, along with the real-valued parameters $\eta_1, \eta_2 \geq 0$. We assume that $n \geq d \geq l$ and that $\operatorname{rank}(N) = \operatorname{rank}(P) = l$, $\operatorname{rank}(F) = d$ almost surely.

The matrix $V \in \mathbb{R}^{n\times n}$ is a deterministic symmetric matrix, such that V^2 is a positive definite covariance matrix. If $V \neq I_n$, then X in model (2.3) can be seen as the matrix form of a d-dimensional time series $\{X_t\}_{t=1}^n$ and y can be seen as a real-valued time series $\{y_t\}_{t=1}^n$. The covariance matrix V^2 determines the dependence between observations, which might be non-stationary. We will call V^2 the temporal covariance matrix of X and define $\Sigma^2 = PP^{\mathrm{T}} + \eta_1^2 I_k$. Setting $l = d$, $\eta_1 = 0$ reduces model (2.3) to standard linear regression with dependent observations.

The latent variables N connect X to y, whereas F can be considered as noise, thus giving a model where not all directions in the column space of X are important for the prediction of y. The representation (2.3) highlights practical settings where the partial least squares algorithm is expected to outperform principal component regression and similar techniques. In particular, if the covariance of $\eta_1 F$ dominates that of NP^{T}, then the first principal components will be largely uncorrelated to y. In contrast, the first partial least squares basis components should by definition be able to recover relevant latent components.

The partial least squares algorithm is run as described in Section 2.1 with matrix X and vector y defined in (2.3). If $\eta_1 = 0$, then model (2.1) is correctly specified with $q = P^{\mathrm{T}}\beta$ and the partial least squares estimator (2.2) estimates β. If $\eta_1 > 0$, then $\beta(\eta_1) = \Sigma^{-2}Pq$ is rather estimated. Note that $\beta(0) = \beta$.

In the standard partial least squares algorithm it is assumed that $V = I_n$. In the subsequent sections we aim to quantify the influence of $V \neq I_n$, which is ignored in the algorithm.

2.2.2 Population and sample partial least squares

The population partial least squares algorithm for independent observations was first introduced by Helland (1990). Under model (2.3) we modify the definition of the population partial least squares basis vectors as

$$w_i = \arg \max_{\substack{w \in \mathbb{R}^d \\ \|w\|=1}} \frac{1}{n^2} \sum_{t,s=1}^{n} \mathrm{Cov}(y_t - X_t^{\mathrm{T}} \beta_{i-1}, X_s^{\mathrm{T}} w), \quad \beta_0 = 0,$$

where $\beta_i \in \mathbb{R}^d$ are the population partial least squares regression coefficients. The average covariances over observations are taken, since the data are neither independent nor identically distributed if $V^2 \neq I_n$. Solving this optimization problem implies that the basis vectors $w_1, \ldots, w_i$ span the Krylov space $\mathcal{K}_i(\Sigma^2, Pq)$: see Section 2.7.1. In particular, under model (2.3), the Krylov space in the population turns out to be independent of the temporal covariance V^2 for all $n \in \mathbb{N}$.

For a given Krylov space, the population partial least squares coefficients are obtained as

$$\beta_i = \arg \min_{\beta \in \mathcal{K}_i(\Sigma^2, Pq)} \mathrm{E} \left\{ \frac{1}{n} \sum_{t=1}^{n} (y_t - X_t^{\mathrm{T}} \beta)^2 \right\}.$$

It is easy to see that the solution to this problem is

$$\beta_i = K_i \left(K_i^{\mathrm{T}} \Sigma^2 K_i \right)^{-1} K_i^{\mathrm{T}} Pq, \quad K_i = (Pq, \Sigma^2 Pq, \ldots, \Sigma^{2(i-1)} Pq),$$

which is independent of V^2 for all $n \in \mathbb{N}$.

To obtain the sample partial least squares estimators $\widehat{\beta}_i$, Σ^2 and Pq are replaced by estimators.

In the standard partial least squares algorithm, under independence of observations, Σ^2 and Pq are estimated by unbiased estimators $n^{-1}X^{\mathrm{T}}X$ and $n^{-1}X^{\mathrm{T}}y$, respectively. However, if the observations are dependent, such naive estimators can lead to $\mathcal{L}^2$-inconsistent estimation, as the following theorem shows.

Theorem 2.1 *Let the model (2.3) hold and the fourth moments of $N_{1,1}$, $F_{1,1}$ exist. Define $A = \|V\|^{-2}X^{\mathrm{T}}X, \quad b = \|V\|^{-2}X^{\mathrm{T}}y$. Then*

$$\begin{aligned} \mathrm{E}\left(\left\|\Sigma^2 - A\right\|^2\right) &= \frac{\|V^2\|^2}{\|V\|^4}\left(C_A + \sum_{t=1}^{n}\frac{\|V_t\|^4}{\|V^2\|^2}c_A\right) \\ \mathrm{E}\left(\|Pq - b\|^2\right) &= \frac{\|V^2\|^2}{\|V\|^4}\left(C_b + \sum_{t=1}^{n}\frac{\|V_t\|^4}{\|V^2\|^2}c_b\right), \end{aligned}$$

where

$$\begin{aligned} C_A &= \|P\|^4 + \|P^{\mathrm{T}}P\|^2 + 4\eta_1^2\|P\|^2 + \eta_1^4 d(1+d) \\ c_A &= \left\{\mathrm{E}\left(N_{1,1}^4\right) - 3\right\}\sum_{i=1}^{l}\|P_i\|^4 + \left\{\mathrm{E}\left(F_{1,1}^4\right) - 3\right\}\eta_1^4 d \\ C_b &= \|Pq\|^2 + \|P\|^2\|q\|^2 + \eta_1^2 d\|q\|^2 + \eta_1^2\eta_2^2 d + \eta_2^2\|P\|^2 \\ c_b &= \left\{\mathrm{E}\left(N_{1,1}^4\right) - 3\right\}\sum_{i=1}^{l}\|P_i\|^2 q_i^2 \end{aligned}$$

and V_t denotes the t-th column of matrix V.

The scaling factors in A and b have no influence on the sample partial least squares estimators in (2.2), so that replacing n^{-1} with $\|V\|^{-2}$ does not affect the algorithm and both A and b are unbiased estimators for Σ^2 and Pq, respectively.

If $\mathrm{E}(N_{1,1}^4) = \mathrm{E}(F_{1,1}^4) = 3$, then constants c_A and c_b vanish, simplifying expressions for the mean squared error of A and b. This is satisfied, for example, for the standard normal distribution.

Thus, these terms can be interpreted as a penalization for non-normality.

Finally, $\sum_{t=1}^{n}\|V_t\|^4 \leq \sum_{t,s=1}^{n}\left(V_t^{\mathrm{T}}V_s\right)^2 = \|V^2\|^2$ implies that the convergence rate of both estimators is driven by the ratio of Frobenius norms $\|V\|^{-2}\|V^2\|$. In particular, if $\|V\|^{-2}\|V^2\|$ converges to zero, then the elements of the population Krylov space Σ^2 and Pq can be estimated consistently. This is the case, for example, for independent observations with $V = I_n$, since $\|I_n^2\| = \|I_n\| = n^{1/2}$. However, if $\|V\|^{-2}\|V^2\|$ fails to converge to zero, ignoring the temporal dependence V^2 may lead to inconsistent estimation.

2.3 Properties of partial least squares estimators under dependence

2.3.1 Concentration inequality for partial least squares estimators

In this section we apply techniques of Blanchard and Krämer (2010b), who derived convergence rates of the kernel conjugate gradient algorithm, which is closely related to kernel partial least squares. Both algorithms approximate the solution on Krylov subspaces, but employ different norms. In particular, Blanchard and Krämer (2010b) have shown that if the conjugate gradient algorithm is stopped early, the convergence in probability of the kernel conjugate gradient estimator to the true regression function can be obtained for bounded kernels. Moreover, the convergence is order-optimal, depending on the regularity of the target function. These results hold for independent identically distributed observations.

We avoid the nonparametric setting of Blanchard and Krämer (2010b) and study a standard linear partial least squares algorithm with a fixed dimension d of the regression space. We allow the observations to be dependent, and, instead of a bounded kernel, consider unbounded random variables with moment conditions. In this setting we derive concentration inequalities for partial least squares estimators that allow us to quantify the influence of the temporal covariance.

We assume that $\eta_1 > 0$ and hence $\operatorname{rank}(A) = d$ almost surely. Regularization of the partial least squares solution is achieved by early stopping, which is characterized by the discrepancy principle, i.e., we stop at the first index $0 < a_0 \leq a$ such that

$$\left\| A^{1/2}\widehat{\beta}_{a_0} - A^{-1/2}b \right\| \leq \tau(\delta\|\widehat{\beta}_{a_0}\| + \epsilon), \tag{2.4}$$

for $\delta, \epsilon > 0$ defined in Theorem 2.2, and some $\tau \geq 1$. Here a denotes the maximal dimension of the sample Krylov space $\mathcal{K}_i(A, b)$ and almost surely equals d. For technical reasons we stop at $a^* = a_0 - 1$ if $p_{a_0}(0) \geq \zeta\delta^{-1}$, where p_i is a polynomial of degree $i - 1$ with $p_i(A)b = \widehat{\beta}_i$ and $\zeta < \tau^{-1}$. The existence of such polynomials was proved by Phatak and de Hoog (2002). If (2.4) never holds, $a^* = a$ is taken. With this stopping index we get the following concentration inequality.

Theorem 2.2 *Assume that model (2.3) with $\eta_1 > 0$ holds and that the fourth moments of $N_{1,1}$, $F_{1,1}$ exist. Furthermore, a^* satisfies (2.4) with $\tau \geq 1$, $\zeta < \tau^{-1}$. For $\nu \in (0, 1]$ let $\delta = \nu^{-1/2}\|V\|^{-2}\|V^2\|C_\delta$ and $\epsilon = \nu^{-1/2}\|V\|^{-2}\|V^2\|C_\epsilon$, such that $\delta, \epsilon \to 0$, where*

$$C_\delta = (2C_A + 2c_A)^{1/2}, \;\; C_\epsilon = (2C_b + 2c_b)^{1/2},$$

with C_A, c_A, C_b and c_b given in Theorem 2.1. Then with probability at least $1-\nu$,

$$\left\|\widehat{\beta}_{a^*}-\beta(\eta_1)\right\| \leq \frac{\|V^2\|}{\|V\|^2}\left\{c_1(\nu)+\frac{\|V^2\|}{\|V\|^2}c_2(\nu)\right\}, \tag{2.5}$$

where

$$c_1(\nu)=\nu^{-1/2}\{c(\tau,\zeta)+o(1)\}\|\Sigma^{-1}\|_{\mathcal{L}}\left(C_\epsilon+\|\Sigma\|_{\mathcal{L}}\|\Sigma^{-3}Pq\|C_\delta\right)$$
$$c_2(\nu)=\nu^{-1}\{c(\tau,\zeta)+o(1)\}\|\Sigma^{-1}\|_{\mathcal{L}}\left(C_\epsilon C_\delta+\|\Sigma^{-3}Pq\|C_\delta^2\right),$$

for some constant $c(\tau,\zeta)$ that depends only on τ and ζ.

If $N_{1,1}$,$F_{1,1}$, $f_1\sim\mathcal{N}(0,1)$, then the expressions for C_δ and C_ϵ are simplified and the scaling factor of $c_1(\nu)$ and $c_2(\nu)$ can be improved from $\nu^{-1/2}$ to $\log(2/\nu)$, which is achieved by using an exponential inequality proved in Theorem 3.3.4 of Yurinsky (1995).

Theorem 2.2 states that the convergence rate of the optimally stopped partial least squares estimator $\widehat{\beta}_{a^*}$ to the true parameter $\beta(\eta_1)$ is driven by the ratio of the Frobenius norms of V^2 and V, similar to the results of Theorem 2.1. In particular, if the data are independent with $V=I_n$ then $\widehat{\beta}_{a^*}$ is square-root consistent. In this case $c_2(\nu)$ is asymptotically negligible. Note that the theorem excludes the case that $\|V\|^{-2}\|V^2\|$ does not converge to zero.

2.3.2 Properties of $\widehat{\beta}_1$ under dependence

Nonlinearity in the response variable of $\widehat{\beta}_i$ hinders its standard statistical analysis, as no closed-form expression for the mean square error of $\widehat{\beta}_i$ is available and concentration inequalities similar to (2.5) are the only results on the convergence rates of partial least squares estimators, to the best of our knowledge. However, if the ratio of $\|V^2\|$ and $\|V\|^2$ does not converge to zero, Theorem 2.2 does not hold.

In this section we study the first partial least squares estimator $\widehat{\beta}_1$, for several reasons. First, the explicit expression for its mean square error can be derived. Second, if there is only one latent component that links X and y, i.e., $l = 1$ in (2.3), then consistent estimation of β_1 is crucial. Finally, $\widehat{\beta}_1$ is collinear to the direction of the maximal covariance between X and y given by $\widehat{w}_1$, which is important for the interpretation of the partial least squares model in applications, see Krivobokova et al. (2012). The next theorem gives conditions under which $\widehat{\beta}_1$ is an inconsistent estimator of β_1.

Theorem 2.3 *Assume that model (2.3) holds, $d > 1$ and eighth moments of $N_{1,1}$, $F_{1,1}$, f_1 exist. Furthermore, suppose that the ratio $\|V\|^{-2}\|V^2\|$ does not converge to zero as $n \to \infty$. Then, for either $l > 1$, $\eta_1 \geq 0$ or $l = 1$, $\eta_1 > 0$, $\widehat{\beta}_1$ is an inconsistent estimator for β_1.*

The case $l = 1$, $\eta_1 = 0$ not treated in Theorem 2.3 corresponds to the standard linear regression model with a single covariate, so the partial least squares estimator coincides with the ordinary least squares estimator, see Helland (1988).

Hence, if there is only one latent component in the model, i.e., $l = 1$, $\eta_1 > 0$, and $\|V\|^{-2}\|V^2\|$ does not converge to zero, then $\beta(\eta_1)$, which in this case equals β_1, cannot be estimated consistently with a standard partial least squares algorithm.

2.3.3 Examples of dependence structures

In all previous theorems the ratio $\|V^2\|\|V\|^{-2}$ plays a crucial role. In this section some special covariance matrices V^2 are studied in order to understand its behaviour. Stationary processes considered in this section are assumed to have expectation zero and to decay exponentially, i.e., for for $c, \rho > 0$ and $\gamma(0) > 0$,

$$|\gamma(t)| \leq \gamma(0) c \exp(-\rho t), \;\; t \in \mathbb{Z}, \tag{2.6}$$

with $\gamma : \mathbb{Z} \to \mathbb{R}$ being the autocovariance function of the process.

Subsequently, $f(n) \sim g(n)$ denotes $c_1 \leq f(n)/g(n) \leq c_2$, for n large, $0 < c_1 < c_2$ and $f, g : \mathbb{N} \to \mathbb{R}$.

Theorem 2.4 *Let $[V^2]_{t,s} = \gamma(|t-s|)$ ($t, s = 1, \ldots, n$) be the covariance matrix of a stationary process, such that the autocovariance function $\gamma : \mathbb{Z} \to \mathbb{R}$ satisfies (2.6). Then $\|V^2\| \sim n^{1/2}$ and $\|V\|^2 \sim n$.*

Hence, if V^2 in model (2.3) is a covariance matrix of a stationary process, then ignoring dependence of observations in the partial least squares algorithm does not affect the rate of convergence of partial least squares estimators, but might affect the constants. Examples of processes with exponentially decaying autocovariances are stationary autoregressive moving average processes.

As examples of nonstationary processes we consider first-order integrated processes. If $\{X_t\}_{t \in \mathbb{Z}}$ is stationary with autocovariance function γ satisfying (2.6), then $\sum_{i=1}^{t} X_i$ is an integrated process of order one.

Theorem 2.5 *Let $\{X_t\}_{t \in \mathbb{Z}}$ be a stationary process with autocovariance function γ satisfying (2.6). If $\gamma(t) < 0$ for some t, we assume additionally $\rho > \log(2c+1)$. Let V^2 be the covariance matrix of $\sum_{i=1}^{t} X_i$. Then $\|V\|^2 \sim n^2$ and $\|V^2\| \sim n^2$.*

The lower bound on ρ for negative $\gamma(t)$ ensures that no element on the diagonal of V^2 becomes negative, so that V^2 is a valid covariance matrix.

This theorem implies that the ratio $\|V\|^{-2}\|V^2\|$ does not converge to zero for certain integrated processes. In particular, combining this result with Theorems 2.1 and 2.3 shows that the elements of the sample Krylov space A and b, as well as $\widehat{\beta}_1$, are inconsistent, if the dependence structure of the data can be described by an integrated process satisfying the conditions of Theorem 2.5, e.g., an integrated autoregressive moving average process of order $(1, 1, 1)$.

2.4 Practical issues

2.4.1 Corrected partial least squares estimator

So far we considered the standard partial least squares algorithm, showing that if certain dependences in the data are ignored, estimation is inconsistent. Hence, it is crucial to take into account the dependence structure of the data in the partial least squares estimators.

Let us define $b\,(S) = n^{-1}X^{\mathrm{T}}S^{-2}y$ and $A\,(S) = n^{-1}X^{\mathrm{T}}S^{-2}X$ for an invertible matrix $S \in \mathbb{R}^{n\times n}$. Furthermore, let $k_i(S) = A(S)^{i-1}b(S)$, $K_i(S) = [k_1(S), \dots, k_i(S)] \in \mathbb{R}^{d\times i}$ and $\widehat{\beta}_i(S) = K_i(S)\left\{K_i(S)^{\mathrm{T}}A(S)K_i(S)\right\}^{-1} K_i(S)^{\mathrm{T}}b(S)$, $i = 1, \dots, d$.

For $S = I_n$ this yields a standard partial least squares estimator. If $S = V$, the temporal dependence matrix, then $b\,(V)$ and $A\,(V)$ are square-root consistent estimators of Pq and Σ^2, respectively, with the mean squared error independent of V, which follows from Theorem 2.1. Hence, the resulting $\widehat{\beta}_i(V)$ is also a consistent estimator of β_i and Theorem 2.2 shows that $\beta(\eta_1)$ can be estimated consistently by early stopping as well. This procedure is equivalent to running the partial least squares algorithm on $V^{-1}y$ and $V^{-1}X$, that is, with the temporal dependence removed from the data.

In practice the true covariance matrix V^2 is typically unknown and is replaced by a consistent estimator $\widehat{V}^2$. We call the estimator $\widehat{\beta}_i(\widehat{V})$ the corrected partial least squares estimator. The next theorem shows that, given a consistent estimator of V^2, the population Krylov space and $\beta(\eta_1)$ can be estimated consistently.

Theorem 2.6 *Let $\widehat{V}^2$ be an estimator for V^2 that is almost surely invertible for $n \in \mathbb{N}$ and $\left\|V\widehat{V}^{-2}V - I_n\right\|_{\mathcal{L}} = O_p(r_n)$, where r_n is some sequence of positive numbers such that $r_n \to 0$ as $n \to \infty$. Then*

$$\|A(\widehat{V}) - \Sigma^2\|_{\mathcal{L}} = O_p(r_n), \quad \|b(\widehat{V}) - Pq\| = O_p(r_n).$$

Moreover, if we assume that $\eta_1 > 0$, we have with probability at least $1 - \nu$, $\nu \in (0, 1]$,

$$\|\widehat{\beta}_{a^*}(\widehat{V}) - \beta(\eta_1)\| = O(r_n),$$

where the definition of a^* in (2.4) is updated by replacing A, b and $\widehat{\beta}_i$ by $A(\widehat{V})$, $b(\widehat{V})$ and $\widehat{\beta}_i(\widehat{V})$, respectively.

Theorem 2.6 states that if a consistent estimator of the covariance matrix V^2 is available, then the elements of the population Krylov space A, b, as well as the coefficient $\beta(\eta_1)$, can be consistently estimated by $A(\widehat{V})$, $b(\widehat{V})$ and $\widehat{\beta}_{a^*}(\widehat{V})$. The convergence rate of these estimators is not faster than that of $\widehat{V}^2$. For example, if the temporal dependence in the data follows some parametric model, then parametric rates of $n^{-1/2}$ are also achieved for $A(\widehat{V})$, $b(\widehat{V})$ and $\widehat{\beta}_{a^*}(\widehat{V})$.

Estimation of V^2 by some nonparametric methods, e.g., with a banding or tapering approach, leads to a slower convergence rates: see Bickel and Levina (2008) or Wu and Xiao (2012). Similar results are well-known in the context of linear regression. For example, Theorem 5.7.1 in Fuller (1996) shows that the convergence rate of feasible generalized least squares estimators is the same as that of the estimator for the covariance matrix of the regression error.

2.4.2 Estimation of covariance matrices

To obtain the corrected partial least squares estimator, some consistent estimator of V^2 based on a single realisation of the process is necessary. In model (2.3) the dependence structure over the observations of X is the same as that of y and V can be estimated from y alone.

If V^2 is the autocovariance matrix of a stationary process, it can be estimated both parametrically and nonparametrically. Many stationary processes can be sufficiently well approximated by an autoregressive moving average process, see Brockwell and Davis (1991), Chapter 4.4. Parameters of autoregressive moving average processes are estimated either by Yule–Walker or maximum likelihood estimators, both attaining parametric rates. Another approach is to band or taper the empirical autocovariance function of y (Bickel and Levina, 2008; Wu and Pourahmadi, 2009; Wu and Xiao, 2012). These nonparametric estimators are very flexible, but are computationally intensive and have slower convergence rates.

If y is an integrated processes of order one, then V^2 can easily be derived from the covariance matrix estimator of the corresponding stationary process.

2.5 Simulations

To verify small sample performance of the partial least squares algorithm under dependence we consider the following simulation setting. To illustrate consistency we choose three sample sizes $n \in \{250, 500, 2000\}$. In the latent variable model (2.3) we set $d = 20$, $l = 1, 5$ and take the elements of P to be independent identically distributed Bernoulli random variables with success probability 0.5. Elements of the vector q are $q_i = 5\, i^{-1}$, $i = 1, \ldots, l$, in order to control the importance of the different latent variables for y. The random variables $N_{1,1}$, $F_{1,1}$ and f_1 are taken to be standard normally distributed. The parameter η_2 is chosen to get the signal to noise ratio in y to be two and η_1 is set so that the signal to noise ratio in X is 0.5. Three matrices V^2 are considered: the identity matrix, the covariance matrix of an autoregressive process of the first order with coefficient 0.9 and the covariance matrix of an autoregressive integrated moving average process of order $(1, 1, 1)$ with both parameters set to 0.9.

First, we ran the standard partial least squares algorithm on the data with the three aforementioned dependence structures to highlight the effect of the ignored dependence in the data. Next, we studied the performance of our corrected partial least squares algorithm applied to nonstationary data. Thereby, the covariance matrix of the autoregressive moving average process has been estimated parametrically, as discussed in Section 2.4.2. A nonparametric estimation of this covariance matrix has lead to qualitative similar results.

The boxplots in Figure 2.1 show the squared distance of $\widehat{\beta}_i$ and $\beta(\eta_1)$ in 500 Monte Carlo replications. Two cases are shown in one panel: the model has just one latent component and $\widehat{\beta}_1$ is considered, i.e., $l = i = 1$ and the model has five latent components and the squared distance of $\widehat{\beta}_5$ to $\beta(\eta_1)$ is studied, i.e., $l = i = 5$.

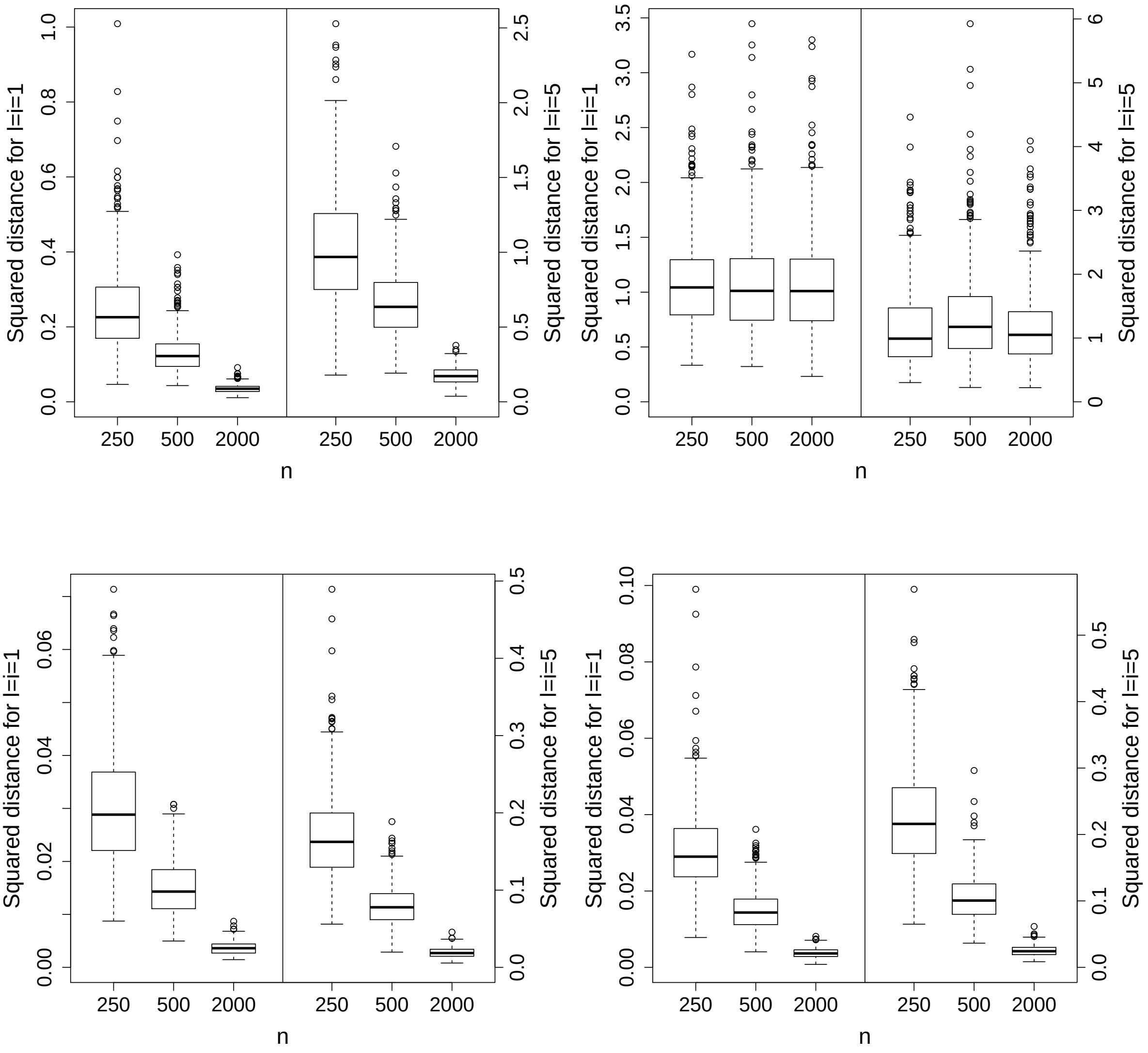

Figure 2.1: Squared distance of partial least squares estimators $\widehat{\beta}_i$ and $\beta(\eta_1)$ in 500 Monte Carlo samples. First three boxplots in each panel correspond to $l = i = 1$, the latter three to $l = i = 5$. The dependence structures are: first order autoregressive (top left), autoregressive integrated moving average of order (1,1,1) (right) and independent, identically distributed (bottom left). The standard partial least squares (top and bottom left) and corrected partial least squares (bottom right) have been employed.

We observe that the mean squared error of $\widehat{\beta}_i$ obtained with the standard partial least squares converges to zero for autoregressive and independent data with the growing sample size. However, an autoregressive dependence in the data leads to a somewhat higher mean squared error, compare the top and bottom left panels. If the data follow an autoregressive integrated moving average process and this is ignored in the partial least squares algorithm, then the mean squared error of $\widehat{\beta}_i$ converges to some positive constant, see the top right boxplots. Taking into account these nonstationary dependencies in the corrected partial least squares leads to consistent estimation, similar to the independent data case, compare the bottom left and right panels.

We conclude that if the observations are dependent, corrected partial least squares improves estimation: in case of stationary dependence the mean squared error is reduced and in case of nonstationary dependence the estimation becomes consistent.

2.6 Application to Protein Dynamics

Proteins fulfil their biological function through particular movements, see Henzler-Wildman and Kern (2007), so a key step in understanding protein functions is a detailed knowledge of the underlying dynamics. Molecular dynamics simulations (de Groot et al., 1998) are routinely used to study the dynamics of biomolecular systems at atomic detail on timescales of nanoseconds to microseconds. Although in principle allowing to directly address function-dynamics relationships, analysis is frequently hampered by the large dimensionality of the protein configuration space, rendering it non-trivial to identify collective modes of motion that are directly related to a functional property of interest.

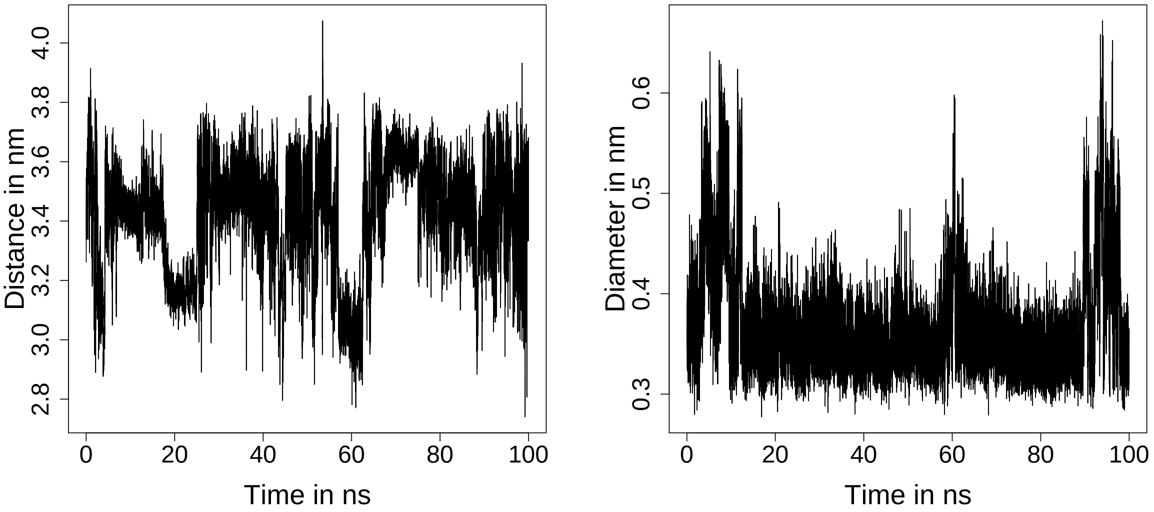

Figure 2.2: Distance between the first backbone atom and the first centre of mass of aquaporine (left) and the opening diameter over time (right).

Krivobokova et al. (2012) have shown that partial least squares helps to identify a hidden relation between atom coordinates of a protein and a functional parameter of interest, yielding robust and parsimonious solutions, superior to principal component regression. In this work we look at a protein studied in the aforementioned paper: the water channel aquaporine as found in the yeast Pichia pastoris. This is a gated channel, i.e., the diameter of the opening can change, controlling the flow of water into the cell. We aim to study which collective motions of protein atoms influence the diameter y_t of the channel at time t, measured by the distance of two centres of mass of the residues of the protein which characterize the opening. For the description of the protein dynamics we use an inner model, i.e. at each point in time we calculate the Euclidean distance d of each backbone atom of the protein and a set of certain four fixed base points. We denote the $p = 739$ atoms by $A_{t,1}, \ldots, A_{t,p} \in \mathbb{R}^3$, the fixed base points by $B_1, \ldots, B_4 \in \mathbb{R}^3$ and take

$$X_t = \{d(A_{t,1}, B_1), \ldots, d(A_{t,p}, B_1), d(A_{t,1}, B_2), \ldots, d(A_{t,p}, B_4)\}^{\mathrm{T}} \in \mathbb{R}^{4p}.$$

The available timeframe has a length of 100 ns split into $n = 20\,000$ equidistant points of observation. Krivobokova et al. (2012) found that a linear relationship between X and y can be assumed.

Taking a closer look at the data reveals that both y_t and $X_{t,i}$, $i = 1, \ldots, 4p$, are nonstationary time series, see Figure 2.2. For the calculation of $\widehat{V}^2$ we used the banding approach mentioned in Section 2.4.2 and found the results to be very similar to a simple autoregressive integrated moving average process with parameters (3,1,1) and corresponding coefficients $(0.1094, 0.0612, 0.0367, -0.9159)$. Autoregressive integrated moving average models have been employed before to study protein time series (Alakent et al., 2004).

To validate our estimators, we used the following procedure. First, the data were split into two equal parts and the models were build on the first half. Then the prediction was done on the test set consisting of the second half of the data and was compared to y_t from the test set. To measure the accuracy of the prediction we used the Pearson correlation coefficient common in the biophysics community and the residual sum of squares, both shown in Figure 2.3. The partial least squares estimator clearly outperforms principal components regression. The corrected partial least squares algorithm, which takes temporal dependence into account, delivers better prediction than standard partial least squares. The improvement is strongly present in the first components. High predictive power of the first corrected partial least squares components is particularly relevant for the interpretation of the underlying protein dynamics. Krivobokova et al. (2012) established that the first partial least squares regression coefficient $\widehat{\beta}_1$ corresponds to the so-called ensemble-weighted maximally correlated mode of motion contributing most to the fluctuation in the response y.

Altogether, due to the low dimensionality, corrected partial least squares greatly facilitates the

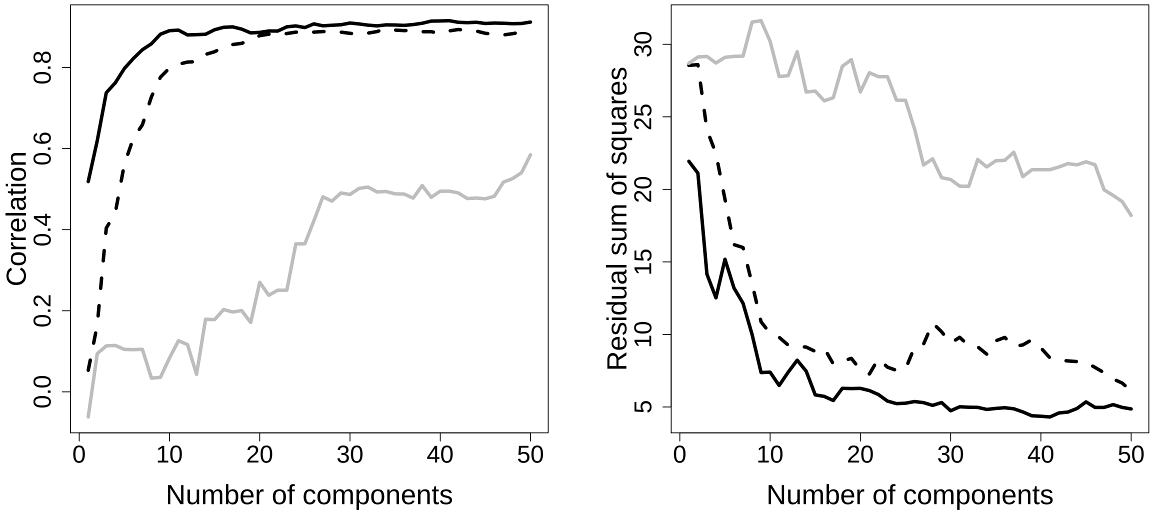

Figure 2.3: Correlation (left) and residual sum of squares (right) of the predicted opening diameter and the real data on the test set. Compared methods are principal component regression (grey), corrected partial least squares (black, solid) and partial least squares (black, dashed).

interpretation of the underlying relevant dynamics, compared to partial least squares and principal component regression, where many more components are required to obtain the same predictive power.

Acknowledgements

The authors are grateful to the reviewers for their valuable comments which helped to improve the article. The support of the German Research Foundation is gratefully acknowledged.

2.7 Proofs

2.7.1 Derivation of the population partial least squares components

Let $K_i \in \mathbb{R}^{d \times i}$ denote the matrix representation of a base for $\mathcal{K}_i(\Sigma^2, Pq)$. Then

$$\sum_{t=1}^{n} \mathrm{E}\left(y_t - X_t^{\mathrm{T}} K_i \alpha\right)^2 = \sum_{t=1}^{n} [V^2]_{t,t} \left(\|q\|^2 + \eta_2^2 - 2\alpha^{\mathrm{T}} K_i^{\mathrm{T}} Pq + \alpha^{\mathrm{T}} K_i^{\mathrm{T}} \Sigma^2 K_i \alpha\right).$$

Taking the derivative with respect to $\alpha \in \mathbb{R}^i$ and setting the resulting equation to zero gives $K_i^{\mathrm{T}} \Sigma^2 K_i \alpha = K_i Pq$. Since the matrix $K_i^{\mathrm{T}} \Sigma^2 K_i$ is invertible, we get the least squares fit β_i in Section 2.2.

Assume now that the first $i < a$ partial least squares base vectors $w_1, \ldots, w_i$ have been calculated and consider for $\lambda \in \mathbb{R}$ the Lagrange function

$$\sum_{t,s=1}^{n} \mathrm{Cov}\left(y_t - X_t^{\mathrm{T}} \beta_i, X_s^{\mathrm{T}} w\right) - \lambda(\|w\|^2 - 1) = w^{\mathrm{T}} \left(Pq - \Sigma^2 \beta_i\right) \sum_{t,s=1}^{n} [V^2]_{t,s} - \lambda(\|w\|^2 - 1).$$

Maximizing with respect to w yields

$$w_{i+1} = (2\lambda)^{-1} \left(Pq - \Sigma^2 \beta_i\right) \sum_{t,s=1}^{n} [V^2]_{t,s} \propto Pq - \Sigma^2 \beta_i.$$

Since $\beta_i \in \mathcal{K}_i(\Sigma^2, Pq)$, we get $w_{i+1} \in \mathcal{K}_{i+1}(\Sigma^2, Pq)$ and w_{i+1} is orthogonal to $w_1, \ldots, w_i$.

2.7.2 Proof of Theorem 2.1

First consider

$$\begin{aligned}
\mathrm{E}\left(\|b-Pq\|^2\right) =& \mathrm{E}\left[\left\|\frac{1}{\|V\|^2}\left\{(PN^{\mathrm{T}}+\eta_1F^{\mathrm{T}})V^2Nq+\eta_2(PN^{\mathrm{T}}+\eta_1F^{\mathrm{T}})V^2f\right\}-Pq\right\|^2\right]\\
=& \left\{\mathrm{E}\left(\left\|\frac{1}{\|V\|^2}PN^{\mathrm{T}}V^2Nq-Pq\right\|^2\right)+\frac{\eta_2^2}{\|V\|^4}\mathrm{E}\left(\left\|PN^{\mathrm{T}}V^2f\right\|^2\right)\right\}\\
&+\frac{\eta_1^2}{\|V\|^4}\left\{\mathrm{E}\left(\left\|F^{\mathrm{T}}V^2Nq\right\|^2\right)+\eta_2^2\,\mathrm{E}\left(\left\|F^{\mathrm{T}}V^2f\right\|^2\right)\right\}=S_1+S_2,
\end{aligned}$$

due to the independence of N, F and f. It is easy to see that

$$S_2=\frac{\|V^2\|^2}{\|V\|^4}\eta_1^2d\left(\|q\|^2+\eta_2^2\right).$$

Furthermore, with $A_0=N^{\mathrm{T}}V^2N$, we get

$$S_1=\frac{1}{\|V\|^4}\mathrm{E}\left(q^{\mathrm{T}}A_0P^{\mathrm{T}}PA_0q\right)-\|Pq\|^2+\frac{\eta_2^2}{\|V\|^4}\mathrm{E}\left(\left\|PN^{\mathrm{T}}V^2f\right\|^2\right).$$

Consider now $\mathrm{E}\left(q^{\mathrm{T}}A_0P^{\mathrm{T}}PA_0q\right)$ as a quadratic form with respect to the matrix $P^{\mathrm{T}}P$. Denote $\kappa=\mathrm{E}\left(N_{1,1}^4\right)-3$. First, $\mathrm{E}\left(A_0q\right)=\mathrm{E}\left(N^{\mathrm{T}}V^2Nq\right)=\|V\|^2q$ and

$$\begin{aligned}
\mathrm{Var}(A_0q)&=\left[\sum_{a,b=1}^{l}q_aq_b\sum_{t,s,u,v=1}^{n}V_u^{\mathrm{T}}V_sV_t^{\mathrm{T}}V_v\,\mathrm{E}(N_{s,i}N_{u,a}N_{t,j}N_{v,b})\right]_{i,j=1}^{l}-\|V\|^4qq^{\mathrm{T}}\\
&=\left[q_iq_j\|V\|^4+\left(q_iq_j+\delta_{i,j}\|q\|^2\right)\|V^2\|^2+\kappa\sum_{t=1}^{n}\|V_t\|^4\delta_{i,j}q_i^2\right]_{i,j=1}^{l}-\|V\|^4qq^{\mathrm{T}}\\
&=\|V^2\|^2\left(qq^{\mathrm{T}}+\|q\|^2I_l\right)+\kappa\sum_{t=1}^{n}\|V_t\|^4\mathrm{diag}\left(q_1^2,\dots,q_l^2\right),
\end{aligned}$$

where $\operatorname{diag}(v_1, \ldots, v_l)$ denotes the diagonal matrix with entries $v_1, \ldots, v_l \in \mathbb{R}$ on its diagonal and δ is the Kronecker delta. In the second equation we made use of $\mathrm{E}\left(N_{s,i}N_{u,a}N_{t,j}N_{v,b}\right) = \delta_{i,a}\delta_{j,b}\delta_{s,u}\delta_{t,v} + \delta_{i,b}\delta_{j,a}\delta_{s,v}\delta_{t,u} + \delta_{i,j}\delta_{a,b}\delta_{t,s}\delta_{u,v} + \kappa\,\delta_{t,s}\delta_{s,u}\delta_{u,v}\delta_{i,j}\delta_{j,a}\delta_{a,b}$, $t, s, u, v = 1, \ldots, n$, $i, j, a, b = 1, \ldots, d$.

Hence,

$$\begin{aligned}\frac{1}{\|V\|^4}\,\mathrm{E}\left(q^{\mathrm{T}}A_0P^{\mathrm{T}}PA_0q\right) =& \frac{1}{\|V\|^4}\operatorname{tr}\left\{P^{\mathrm{T}}P\operatorname{Var}\left(A_0q\right)\right\} - \frac{1}{\|V\|^4}\,\mathrm{E}\left(q^{\mathrm{T}}A_0\right)P^{\mathrm{T}}P\,\mathrm{E}\left(A_0q\right)\\ =& \frac{\|V^2\|^2}{\|V\|^4}\left(q^{\mathrm{T}}P^{\mathrm{T}}Pq + \|P\|^2\|q\|^2\right) + q^{\mathrm{T}}P^{\mathrm{T}}Pq + \kappa\sum_{t=1}^{n}\frac{\|V_t\|^4}{\|V\|^4}\sum_{i=1}^{l}\|P_i\|^2q_i^2.\end{aligned}$$

The remaining term in S_1 follows trivially, proving the result. $\mathrm{E}\,\|\Sigma^2 - A\|^2$ is obtained using similar calculations. □

2.7.3 Proof of Theorem 2.2

Lemma 2.1 *Assume that for $\nu \in (0, 1]$, $\eta_1 > 0$ and some constants $\delta, \epsilon > 0$ it holds that $\mathrm{P}\left(\|A - \Sigma^2\|_{\mathcal{L}} \leq \delta\right) \geq 1 - \nu/2$ and $\mathrm{P}\left(\|b - Pq\| \leq \epsilon\right) \geq 1 - \nu/2$. Then the inequalities*

$$\begin{aligned}\|A^{1/2} - \Sigma\| &\leq 2^{-1}\delta\|\Sigma^{-1}\|\{1 + o(1)\},\\ \|A^{-1/2}b - \Sigma^{-1}Pq\| &\leq \epsilon\|\Sigma^{-1}\|_{\mathcal{L}} + 2^{-1}\delta(\|Pq\| + \epsilon)\|\Sigma^{-2}\|\|\Sigma^{-1}\|\left\{1 + o(1)\right\}\end{aligned}$$

hold simultaneously with probability at least $1 - \nu$.

Proof: We show the result by using the Fréchet-derivative for functions $F : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$. Due to the fact that $\eta_1 > 0$ it holds that Σ^2 is positive definite and thus invertible. It holds $\operatorname{rank}(A) = d$ almost surely since $\operatorname{rank}(F) = d$ almost surely and NP^{T} and F are independent.

We assume that we are on the set where $\|A - \Sigma^2\|_{\mathcal{L}} \leq \delta$ and $\|b - Pq\| \leq \epsilon$ hold both with probability at least $1 - \nu$.

It holds due to Higham (2008), Problem 7.4, that $F'(\Sigma^2)B$ for an arbitrary $B \in \mathbb{R}^{d\times d}$ is given as the solution in $Z \in \mathbb{R}^{d\times d}$ of $B = \Sigma Z + Z\Sigma$, i.e., due to the symmetry and positive definiteness of Σ we have $F'(\Sigma^2)B = 2^{-1}\Sigma^{-1}B$. We take the orthonormal base $\{E_{i,j}, i, j = 1, \dots, d\}$ for the space $(\mathbb{R}^{d\times d}, \|\cdot\|)$ with $E_{i,j}$ corresponding to the matrix that has zeros everywhere except at the position (i, j), where it is one. The Hilbert-Schmidt norm $\|F'(\Sigma^2)\|_{\text{HS}}$ is

$$\|F'(\Sigma^2)\|^2_{\text{HS}} = 4^{-1}\sum_{i,j=1}^{d} \|\Sigma^{-1}E_{i,j}\|^2 = 4^{-1}\sum_{i,j=1}^{d} [\Sigma^{-1}]^2_{i,j} = 4^{-1}\|\Sigma^{-1}\|^2.$$

This yields with the Taylor expansion for Fréchet-differentiable maps

$$\|A^{1/2} - \Sigma\|_{\mathcal{L}} \leq \|F'(\Sigma)(A - \Sigma^2)\| + o(\|A - \Sigma^2\|) \leq 2^{-1}\|\Sigma^{-1}\|\delta\{1 + o(1)\}.$$

For the second inequality we see first that

$$\|A^{-1/2}b - \Sigma^{-1}Pq\| \leq \epsilon\|\Sigma^{-1}\|_{\mathcal{L}} + \left\|(A^{-1/2} - \Sigma^{-1})b\right\|. \tag{2.7}$$

The Fréchet-derivative of the map $F : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}, A \mapsto A^{-1/2}$ is $F'(\Sigma^2)B = -2^{-1}\Sigma^{-2}B\Sigma^{-1}$ for $B \in \mathbb{R}^{d\times d}$ and

$$\|F'(\Sigma^2)\|^2_{\text{HS}} = 4^{-1}\sum_{i,j=1}^{d} \|\Sigma^{-2}E_{i,j}\Sigma^{-1}\|^2 \leq 4^{-1}\|\Sigma^{-2}\|^2\|\Sigma^{-1}\|^2.$$

Here we used the submultiplicativity of the Frobenius norm with the Hadamard product of matrices. Thus we get via Taylor's theorem

$$\|A^{-1/2} - \Sigma^{-1}\| \leq 2^{-1}\|\Sigma^{-2}\|\|\Sigma^{-1}\|\|A - \Sigma^2\|\delta\{1 + o(1)\}.$$

Plugging this into (2.7) yields

$$\|A^{-1/2}b - \Sigma^{-1}Pq\| \leq \epsilon\|\Sigma^{-1}\|_{\mathcal{L}} + 2^{-1}\delta(\|Pq\| + \epsilon)\|\Sigma^{-2}\|\|\Sigma^{-1}\| \left\{1 + o(1)\right\},$$

where we used that $\|b\| \leq \|Pq\| + \epsilon$. □

Equivalence of conjugate gradient and partial least squares: We denote $\tilde{A} = A^{1/2}$ and $\tilde{b} = A^{-1/2}b$. The partial least squares optimization problem is

$$\min_{v \in \mathcal{K}_i(A,b)} \|y - Xv\|^2,$$

whereas the conjugate gradient problem studied in Nemirovskii (1986) is

$$\min_{v \in \mathcal{K}_i(\tilde{A}^2, \tilde{A}\tilde{b})} \|\tilde{b} - \tilde{A}v\|^2. \tag{2.8}$$

It is easy to see that the Krylov space $\mathcal{K}_i(\tilde{A}^2, \tilde{A}\tilde{b}) = \mathcal{K}_i(A, b)$, $i = 1, \ldots, d$. We have

$$\arg \min_{v \in \mathcal{K}_i(\tilde{A}^2, \tilde{A}\tilde{b})} \|\tilde{b} - \tilde{A}v\|^2 = \arg \min_{\mathcal{K}_i(A,b)} \|y - Xv\|^2, i = 1, \ldots, d.$$

Thus it holds

$$\widehat{\beta}_i = \arg \min_{v \in \mathcal{K}_i(\tilde{A}^2, \tilde{A}\tilde{b})} \|\tilde{b} - \tilde{A}v\|^2.$$

Furthermore we have $\Sigma\beta(\eta_1) = \Sigma^{-1}Pq$, i.e., the correct problem in the population is solved by $\beta(\eta_1)$ as well. Now we will restate the main result in Nemirovskii (1986) in our context:

Theorem 2.7 *Nemirovskii*

Assume that we have $\eta_1 > 0$ and there are $\tilde{\delta} = \tilde{\delta}(\nu, n) > 0$, $\tilde{\epsilon} = \tilde{\epsilon}(\nu, n) > 0$ such that for $\nu \in (0,1]$ it holds that $\mathrm{P}\left(\|\Sigma - A^{1/2}\|_{\mathcal{L}} \leq \tilde{\delta}, \|\Sigma^{-1}Pq - A^{-1/2}b\| \leq \tilde{\epsilon}\right) \geq 1 - \nu$. Assume furthermore that there is a vector $u \in \mathbb{R}^d$ and constants $R, \mu > 0$ such that $\beta(\eta_1) = \Sigma^{\mu}u$, $\|u\| \leq R$ is satisfied.

If we stop according to the stopping rule a^* as defined in (2.4) with $\tau \geq 1$ and $\zeta < \tau^{-1}$ then we have for any $\theta \in [0,1]$ with probability at least $1 - \nu$

$$\left\|\Sigma^{\theta}\{\widehat{\beta}_{a^*} - \beta(\eta_1)\}\right\|^2 \leq C^2(\mu, \tau, \zeta) R^{2(1-\theta)/(1+\mu)} \left(\tilde{\epsilon} + \tilde{\delta} R L^{\mu}\right)^{2(\theta+\mu)/(1+\mu)}.$$

Proof: Note first that on the set where $\|\Sigma - A^{1/2}\|_{\mathcal{L}} \leq \tilde{\delta}$ holds we also have $\max\{\|A^{1/2}\|_{\mathcal{L}}, \|\Sigma\|_{\mathcal{L}}\} \leq L$. Constrained on the set where all the conditions of the theorem hold with probability at least $1 - \nu$ we consider Nemirovskii's $(\Sigma, A^{1/2}, \Sigma^{-1}Pq, A^{-1/2}b)$ problem with errors $\tilde{\delta}$ and $\tilde{\epsilon}$. Furthermore by assumption Nemirovskii's $(2\theta, R, L, 1)$ conditions hold and thus the theorem follows by a simple application of the main theorem in Nemirovskii (1986). □

We will now apply Theorem 2.7 to our problem. Due to the fact that $\eta_1 > 0$ it holds that Σ^2 is positive definite and thus invertible. We note that the spectral norm is dominated by the

Frobenius norm. From Markov's inequality we get

$$\mathrm{P}\left(\|A-\Sigma^2\| \geq \delta\right) \leq \delta^{-2}\,\mathrm{E}\left(\|A-\Sigma^2\|^2\right).$$

Using Theorem 2.1, $\sum_{t=1}^{n} \|V_i\|^4 \leq \|V^2\|^2$ and setting the right hand side to $\nu/2$ for $\nu \in (0,1]$ gives $\delta = \nu^{-1/2}\|V\|^{-2}\|V^2\|C_\delta$. In the same way $\epsilon = \nu^{-1/2}\|V\|^{-2}\|V^2\|C_\epsilon$. Lemma 2.1 gives with probability at least $1-\nu$ the concentration results required by Theorem 2.7 with

$$\begin{aligned}
\tilde{\delta} &= \nu^{-1/2}\frac{\|V^2\|}{\|V\|^2}C_\delta\{1+o(1)\} \\
\tilde{\epsilon} &= \left(\nu^{-1/2}\frac{\|V^2\|}{\|V\|^2}C_\epsilon + \nu^{-1}\frac{\|V^2\|^2}{\|V\|^4}C_\epsilon C_\delta\right)\{1+o(1)\}
\end{aligned}$$

The remaining condition of Theorem 2.7 holds by choosing $\mu = 1$ and $R = \|\Sigma^{-3}Pq\|$. Here we used that $\beta(\eta_1) = \Sigma^{-2}Pq$. Thus the theorem yields for $\theta = 1$

$$\left\|\Sigma\{\beta(\eta_1) - \widehat{\beta}_{a^*}\}\right\| \leq C(1,\tau,\zeta)\left(\tilde{\epsilon} + \tilde{\delta}RL\right),$$

with $L = \|\Sigma\|_{\mathcal{L}} + \tilde{\delta}$. Denote $c(\tau,\zeta) = C(1,\tau,\zeta)$. Finally we have $\|\Sigma^{-1}\|_{\mathcal{L}}^{-1}\|v\| \leq \|\Sigma v\|$ for any $v \in \mathbb{R}^d$ and thus the theorem is proven with

$$\begin{aligned}
c_1(\nu) &= \nu^{-1/2}\{c(\tau,\zeta)+o(1)\}\|\Sigma^{-1}\|_{\mathcal{L}}\left(C_\epsilon + \|\Sigma\|_{\mathcal{L}}\|\Sigma^{-3}Pq\|C_\delta\right) \\
c_2(\nu) &= \nu^{-1}\{c(\tau,\zeta)+o(1)\}\|\Sigma^{-1}\|_{\mathcal{L}}\left(C_\epsilon C_\delta + \|\Sigma^{-3}Pq\|C_\delta^2\right).
\end{aligned}$$

□

2.7.4 Proof of Theorem 2.3

The theorem is proved by contradiction. Assume that $\widehat{\beta}_1 \longrightarrow \beta_1$ in probability. Choosing $v \in \mathbb{R}^d$, $v \neq 0$, orthogonal to β_1 implies that $v^{\mathrm{T}}\widehat{\beta}_1$ converges in probability to zero. Note that we assume $d > 1$. Next we show that the second moment vanishes as well.

Let $M_r(z) = \max_{i \in I_r} \prod_{j=1}^{r} \mathrm{E}(z^{i_j})$ for a random variable z with existing mixed rth moments, $r \in \mathbb{N}$ and $I_r = \{i \in \{0, \dots, r\}^r : i_1 + i_2 + \cdots + i_r = r\}$. Consider

$$\begin{aligned}\mathrm{E}\left(\|PN^{\mathrm{T}}V^2Nq\|^4\right) &\leq \|P\|^4\|q\|^4\,\mathrm{E}(\|N^{\mathrm{T}}V^2N\|^4)\\ &\leq \|P\|^4\|q\|^4 \sum_{i\in\{1,\dots,l\}^4}\ \sum_{s\in\{1,\dots,n\}^8} \prod_{j=1}^{8}\|V_{s_j}\| \left|\mathrm{E}\left(\prod_{h=1}^{4}\prod_{j=2h-1}^{2h} N_{i_h,s_j}\right)\right|\\ &\leq l^4\|P\|^4\|q\|^4\left\{\sum_{t=1}^{n}\|V_t\|^2\right\}^4 M_8(N_{1,1}) = l^4\|P\|^4\|q\|^4\|V\|^8 M_8(N_{1,1}).\end{aligned}$$

In the last inequality we used the fact that the $N_{t,i}$, $t = 1, \dots, n$, $i = 1, \dots, l$, are independent and identically distributed and $\mathrm{E}(N_{t,i}) = 0$. Thus $\mathrm{E}(\prod_{h=1}^{4}\prod_{j=2h-1}^{2h} N_{i_h,s_j})$ is zero if the random variables N_{i_h,s_j} do not appear at least in pairs in the product. We see that then the norms $\|V_{s_j}\|$ have to appear at least in pairs as well. Finally we can use the fact that $\sum_{t=1}^{n}\|V_t\|^u \leq (\sum_{t=1}^{n}\|V_t\|^2)^{u/2}$ for $u \geq 2$ and the definition of $M_8(N_{1,1})$.

Now using $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$ and the independence of N, F and f we obtain

$$\begin{aligned}\mathrm{E}\,(v^{\mathrm{T}}b)^4 \leq &\frac{8^2\|v\|^4}{\|V\|^8}\,\mathrm{E}\left(\left\|PN^{\mathrm{T}}V^2Nq\right\|^4 + \eta_1^4\left\|F^{\mathrm{T}}V^2Nq\right\|^4 + \eta_2^4\left\|PN^{\mathrm{T}}V^2f\right\|^4 + \eta_1^4\eta_2^4\left\|F^{\mathrm{T}}V^2f\right\|^4\right)\\ \leq &8^2\|v\|^4\left\{M_8(N_{1,1})\|q\|^4 l^4\|P\|^4 + M_4(N_{1,1})M_4(F_{1,1})\eta_1^4\|q\|^4 l^2 d^2\right.\\ &\left.+M_4(N_{1,1})M_4(f_1)\eta_2^4 l^2\|P\|^4 + M_4(F_{1,1})M_4(f_1)\eta_1^4\eta_2^4 d^2\right\} < \infty,\ \ n \in \mathbb{N}.\end{aligned}$$

Thus, $(v^{\mathrm{T}}b)^2$ is uniformally integrable by the theorem of de la Vallée-Poussin and it follows that the directional variance $\mathrm{Var}(v^{\mathrm{T}}b)$ has to vanish in the limit as well. Now, calculations similar to Theorem 2.1 yield

$$\begin{aligned}\mathrm{Var}(v^{\mathrm{T}}b) =& \frac{\|V^2\|^2}{\|V\|^4}\left\{\eta_1^2\|v\|^2\left(\|q\|^2+\eta_2^2\right)+\|P^{\mathrm{T}}v\|^2\left(\|q\|^2+\eta_2^2\right)+(v^{\mathrm{T}}Pq)^2\right\}\\ &+\sum_{t=1}^{n}\frac{\|V_t\|^4}{\|V\|^4}\sum_{i=1}^{l}q_i^2\left(v^{\mathrm{T}}P_i\right)^2\{\mathrm{E}(N_{1,1}^4)-3\},\quad v\in\mathbb{R}^d.\end{aligned}$$

We assumed that $\|V\|^{-2}\|V^2\|$ does not converge to zero. It remains to check under which conditions $\mathrm{Var}(v^{\mathrm{T}}b)$ is larger than zero. This will always be the case if $v\neq 0$ and $\eta_1>0$, $l=1$. For $\eta_1=0$ and $l>1$ a vector v that lies in the range of P and is orthogonal to $\beta_1\propto Pq$ exists, thus contradicting $\widehat{\beta}_1\longrightarrow\beta_1$ in probability. □

2.7.5 Proof of Theorem 2.4

It is easy to verify that $\|V\|^2=\mathrm{tr}(T^2)=n\gamma(0)$ and $\|V^2\|^2=n\gamma^2(0)+2\sum_{t=1}^{n-1}\gamma^2(t)(n-t)$. If (2.6) is fulfilled, then

$$n\gamma^2(0)\leq\left\|V^2\right\|^2\leq n\gamma^2(0)\left\{1+2c^2\frac{1-\exp(-2\rho(n-1))}{\exp(2\rho)-1}\right\}\leq n\gamma^2(0)\left\{1+\frac{2c^2}{\exp(2\rho)-1}\right\}.$$

It follows that $\|V^2\|\sim n^{1/2}$. □

2.7.6 Proof of Theorem 2.5

Let $\gamma:\mathbb{N}\to\mathbb{R}$ be the autocovariance function of a stationary time series that has zero mean. For the autocovariance matrix V^2 of the corresponding integrated process of order one we get

$[V^2]_{t,s} = \sum_{i,j=1}^{t,s} \gamma(|i-j|),\ t,s = 1,\dots,n$. Let $t \geq s$. By splitting the sum into parts with $i < j$ and $i > j$ we get $[V^2]_{t,s} = s\gamma(0) + \sum_{j=1}^{s}\sum_{i=1}^{t-j}\gamma(i) + \sum_{j=2}^{s}\sum_{i=1}^{j-1}\gamma(i)$. Due to symmetry, $[V^2]_{t,s} = [V^2]_{s,t}$ for $s > t$.

First, consider the case that all $\gamma(j)$, $j > 0$, are negative. Using (2.6) we obtain

$$\gamma(0)s \geq \left[V^2\right]_{t,s} \geq \gamma(0)\left\{s - c\sum_{j=1}^{s}\sum_{i=1}^{t-j}\exp(-\rho j) - c\sum_{j=2}^{s}\sum_{i=1}^{j-1}\exp(-\rho j)\right\},\ t \geq s.$$

Evaluation of the geometric sums gives

$$\left[V^2\right]_{t,s} \geq \gamma(0)\left(s\left\{1 - \frac{2c}{\exp(\rho)-1}\right\} + c\frac{\exp(\rho)}{\{\exp(\rho)-1\}^2}\{1-\exp(-\rho s)\}\left[1+\exp\{\rho(s-t)\}\right]\right).$$

The second term on the right is always positive and the positivity of the first term is ensured by the condition $\rho > \log(2c+1)$. Hence, $\gamma(0)\left[1 - 2c\{\exp(\rho)-1\}^{-1}\right]s \leq [V^2]_{t,s} \leq \gamma(0)s,\ s \geq 1$. If $\gamma(t)$, $t \geq 1$, is not purely negative, it can be bound by

$$\gamma(0)\left[1 - 2c\{\exp(\rho)-1\}^{-1}\right]s \leq \left[V^2\right]_{t,s} \leq \gamma(0)\left[1 + 2c\{\exp(\rho)-1\}^{-1}\right]s.$$

We write δ_1 and δ_2 for the constants in the lower and upper bound, respectively, so that $\delta_1 \min\{s,t\} \leq [V^2]_{t,s} \leq \delta_2 \min\{s,t\}$, $t,s = 1,\dots,n$. This yields upper and lower bounds on the trace of V^2 and shows that $\|V\|^2 \sim n^2$. Additionally,

$$\left[V^4\right]_{t,t} = \sum_{l=1}^{n}\left[V^2\right]_{t,l}\left[V^2\right]_{l,t} = \sum_{l=1}^{t}\left[V^2\right]_{t,l}^2 + \sum_{l=t+1}^{n}\left[V^2\right]_{l,t}^2 \leq \frac{\delta_2^2}{6}t\left(6nt - 4t^2 + 3t + 1\right)$$
$$\left[V^4\right]_{t,t} \geq \frac{\delta_1^2}{6}t\left(6nt - 4t^2 + 3t + 1\right).$$

This implies upper and lower bounds on the trace of V^4 in the form $\tilde{c}\,n(n+1)(n^2+n+1)$ for $\tilde{c} \in \{\delta_1^2/6, \delta_2^2/6\}$ and thus $\|V^2\| \sim n^2$. □

2.7.7 Proof of Theorem 2.6

First consider $n^{-1}X^{\mathrm{T}}\widehat{V}^{-2}y$. Define $X_u = (X_{u,1}, \ldots, X_{u,n})^{\mathrm{T}} = NP^{\mathrm{T}} + \eta_1 F$ and $y_u = (y_{u,1}, \ldots, y_{u,n})^{\mathrm{T}} = Nq + \eta_2 f$ such that $X = VX_u$ and $y = Vy_u$. By the triangle inequality we have

$$\left\|n^{-1}X^{\mathrm{T}}\widehat{V}^{-2}y - Pq\right\| \leq \left\|n^{-1}X^{\mathrm{T}}V^{-2}y - Pq\right\| + \left\|n^{-1}X^{\mathrm{T}}\left(\widehat{V}^{-2} - V^{-2}\right)y\right\|.$$

The first term on the right hand side is convergent to zero in probability due to Theorem 2.1. The second term can be bound by

$$n^{-2}\left\|X^{\mathrm{T}}\left(\widehat{V}^{-2} - V^{-2}\right)y\right\|^2 \leq \|V\widehat{V}^{-2}V - I_n\|_{\mathcal{L}}^2\, n^{-1}\|X_u^{\mathrm{T}}\|^2\, n^{-1}\|y_u\|^2.$$

Since both $X_{u,1}, \ldots, X_{u,n}$ and $y_{u,1}, \ldots, y_{u,n}$ are independent and identically distributed, it follows that $n^{-1}\|y_u\|^2$ is a strongly consistent estimator for $\mathrm{E}(y_{u,1}^2)$, as well as that $n^{-1}\|X_u^{\mathrm{T}}\|^2$ is a strongly consistent estimator of $\mathrm{E}(\|X_{u,1}\|^2)$. Convergence in probability of $\left\|V\widehat{V}^{-2}V - I_n\right\|_{\mathcal{L}}^2$ to zero implies the convergence of $b(\widehat{V})$ to Pq in probability by Slutsky's lemma.

To obtain the convergence rate $\|n^{-1}X^{\mathrm{T}}V^{-2}y - Pq\| = O_p(r_n)$, use Theorem 2.1 and $\|V\widehat{V}^{-2}V - I_n\|_{\mathcal{L}} = O_p(r_n)$. The convergence of $\|n^{-1}X^{\mathrm{T}}\widehat{V}^{-2}X - \Sigma^2\|$ is proven in a similar way.

To show the consistency and the rate of the corrected partial least squares estimator, we follow the same lines as in the proof of Theorem 2.2. First, $\delta = r_n c_A(\nu)$ and $\epsilon = r_n c_b(\nu)$ for $\nu \in (0, 1]$

with constants $c_A(\nu)$, $c_b(\nu)$ are taken, such that the inequalities

$$\|A(\widehat{V})^{1/2} - \Sigma\|_{\mathcal{L}} \leq r_n c_A(\nu), \quad \|A(\widehat{V})^{-1/2} b(\widehat{V}) - \Sigma^{-1} Pq\| \leq r_n c_b(\nu)$$

hold simultaneously with probability at least $1 - \nu$. As the product of three matrices that have almost surely full rank d due to $\eta_1 > 0$ it holds that $A(\widehat{V})$ is almost surely invertible. Moreover, $R = \|\Sigma^{-3} Pq\|$, $\mu = 1$ fulfils the remaining condition in Theorem 2.7. Thus, with probability at least $1 - \nu$ we get by setting $\theta = 1$

$$\left\|\widehat{\beta}_{a^*}(\widehat{V}) - \beta(\eta_1)\right\| \leq r_n \, C(1, \tau, \zeta)\{1+o(1)\} \|\Sigma^{-1}\|_{\mathcal{L}} \left[c_b(\nu) + c_A(\nu) \|\Sigma^{-3} Pq\| \left\{\|\Sigma\|_{\mathcal{L}} + r_n c_A(\nu)\right\}\right],$$

where the constants ζ, τ are taken from the definition of a^*. □

Bibliography

Alakent, B., Doruker, P., and Camurdan, M. (2004). Time series analysis of collective motions in proteins. *J. Chem. Phys.*, 120(2):1072–1088.

Almøy, T. (1996). A simulation study on the comparison of prediction methods when only a few components are relevant. *Comput. Statist. Data Anal.*, 21:87–107.

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

Blanchard, G. and Krämer, N. (2010a). Kernel partial least squares is universally consistent. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings*, volume 9, pages 57–64. JMLR.

Blanchard, G. and Krämer, N. (2010b). Optimal learning rates for kernel conjugate gradient regression. *Adv. Neural Inf. Process. Syst.*, 23:226–234.

Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*. Springer, New York, 2 edition.

de Groot, B. L., Hayward, S., van Aalten, D. M. F., Amadei, A., and Berendsen, H. J. C. (1998).

Domain motions in bacteriophage T4 lysozyme; a comparison between molecular dynamics and crystallographic data. *Proteins*, 31:116–127.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.*, 40:322–352.

Farkas, O. and Héberger, K. (2005). Comparison of ridge regression, partial least-squares, pairwise correlation, forward-and best subset selection methods for prediction of retention indices for aliphatic alcohols. *J. Chem. Inf. Model.*, 45:339–346.

Frank, I. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Fuller, W. (1996). *Introduction to Statistical Time Series*. Wiley, New York, 2 edition.

Goldberger, A. (1972). Structural equation methods in the social sciences. *Econometrica*, 40:979–1001.

Hahn, C., Johnson, M., Herrmann, A., and Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Business Review*, 54:243–269.

Helland, I. S. (1988). On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.*, 17(2):581–607.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.*, 17:97–114.

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450:964–972.

Higham, N. (2008). *Functions of Matrices: Theory and Computation*. SIAM, Phildadelphia, 1 edition.

Höskuldsson, A. (1988). PLS regression methods. *J. Chemometr.*, 2:211–228.

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strateg. Manage. J.*, 20:195–204.

Krämer, N. (2007). An overview on the shrinkage properties of partial least squares regression. *Comput. Statist.*, 22:249–273.

Krivobokova, T., Briones, R., Hub, J., Munk, A., and de Groot, B. (2012). Partial least squares functional mode analysis: application to the membrane proteins AQP1, Aqy1 and CLC-ec1. *Biophys. J.*, 103:786–796.

Lobaugh, N., West, R., and McIntosh, A. (2001). Spatiotemporal analysis of experimental differences in event-related potential data with partial least squares. *Psychophysiology*, 38:517–530.

Naik, P. and Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 62:763–771.

Nemirovskii, A. (1986). The regularizing properties of the adjoint gradient method in ill-posed problems. *Comput. Math. Math. Phys.*, 26:7–16.

Nguyen, D. and Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39–50.

Phatak, A. and de Hoog, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *J. Chemometr.*, 16:361–367.

Skrondal, A. and Rabe-Hesketh, S. (2006). Latent variable modelling: A survey. *Scand. J. Statist.*, 34:712–745.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedure. In *Research papers in statistics: Festschrift for J. Neyman*, pages 411–444. Wiley.

Wold, S., Sjøstrøma, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab.*, 58:109–130.

Wu, W. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sinica*, 19:1755–1768.

Wu, W. and Xiao, H. (2012). Covariance matrix estimation in time series. *Ann. Statist.*, 40(1):466–493.

Yurinsky, V. (1995). *Sums and Gaussian Vectors*. Springer, Berlin, 1 edition.

Chapter 3

Kernel Partial Least Squares for Stationary Data

Kernel partial least squares for stationary data

Marco Singer[a], Tatyana Krivobokova[a], Axel Munk[a,b]

[a]Institute for Mathematical Stochastics, Göttingen, Germany

[b]Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

We consider the kernel partial least squares algorithm for the solution of nonparametric regression problems when the data exhibit dependence in their observations in the form of stationary time series. Probabilistic convergence rates of the kernel partial least squares estimator to the true regression function are established under a source condition. The impact of long range dependence in the data is studied both theoretically and in simulations.

Key words and phrases:

Kernel partial least squares, Long range dependence, Nonparametric regression, Source condition, Stationary process

3.1 Introduction

We study the statistical regularization properties of the kernel partial least squares algorithm for the solution of nonparametric regression problems

$$y_t = f^*(X_t) + \varepsilon_t, \;\; t \in \mathbb{Z}. \tag{3.1}$$

For fixed $d \in \mathbb{N}$ we consider the d-dimensional stationary time series $\{X_t\}_{t\in\mathbb{Z}}$ on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an independent and identically distributed sequence of real valued random variables with expectation zero and variance $\sigma^2 > 0$ that is independent of $\{X_t\}_{t\in\mathbb{Z}}$. Furthermore let X be a random vector that is independent of $\{X_t\}_{t\in\mathbb{Z}}$ and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ with the same distribution as X_0.

The aim is to estimate the regression function $f^* \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ based on a training sample $\{(X_t, y_t)^{\mathrm{T}}\}_{t=1}^n$, $n \in \mathbb{N}$.

The focus of this work is on the kernel based learning approach. Due to the representer theorem of Wahba (1999) and its generalization in Schölkopf et al. (2001), reproducing kernel Hilbert space methods have gained popularity in recent years, especially in the machine learning community, and many regularized regression techniques like ridge regression, principal component regression and partial least squares have been adapted to this nonparametric setting (Saunders et al., 1998; Rosipal et al., 2000; Rosipal and Trejo, 2001).

In reproducing kernel Hilbert space methods the data $\{X_t\}_{t=1}^n$ are mapped into a Hilbert space $\mathcal{H}$ of functions on $\mathbb{R}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that the nonparametric regression problem (3.1) becomes linear in a function space. The main advantage of these methods is the kernel trick, i.e., this mapping can be done implicitly via the kernel matrix $K_n = n^{-1}[k(X_t, X_s)]_{t,s=1}^n$. Linear methods are then applied to the problem in the reproducing kernel Hilbert space.

Partial least squares was derived for the solution of linear regression problems with collinearities in the regressor matrix by Wold et al. (1984). The algorithm works similar to principal component regression with the main difference being that in each step the covariance between response and regressor is maximized instead of the variance of the regressor, see Helland (1988) for a detailed description. Similar to principal component regression partial least squares is an iterative regularized regression technique and regularisation is achieved by stopping the algorithm early. It is also closely related to the conjugate gradient algorithm, see Phatak and de Hoog

(2002).

In several studies it has been seen that partial least squares is competitive with other regression methods like ridge regression and principal component regression and it needs generally fewer iterations than the latter to achieve good parameter estimation and prediction, see, e.g., Frank and Friedman (1993) and Krämer and Braun (2007). For an overview of further properties of partial least squares we refer to Rosipall and Krämer (2006).

The method was adapted to the kernel setting in Rosipal and Trejo (2001) by using the reformulation of the algorithm presented in Lindgren et al. (1993). The relationship to kernel conjugate gradient methods was highlighted in Blanchard and Krämer (2010a). It can be seen in Hanke (1995) that conjugate gradient methods are well suited for handling ill-posed problems, as they arise for example in kernel learning (De Vito et al., 2006).

Rosipal (2003) investigated the performance of kernel partial least squares for nonlinear discriminant analysis. Blanchard and Krämer (2010a) proved the consistency of kernel partial least squares when the algorithm is stopped early without giving convergence rates.

For a variant of kernel conjugate gradient explicit probabilistic convergence rates were proven in Blanchard and Krämer (2010b). The question what convergence rates kernel partial least squares achieves remained open.

To the best of our knowledge there was no previous research on the performance of the algorithm when instead of independent and identically distributed observations stationary time series are considered. We investigate this case and see that under short range dependence the convergence rates of kernel partial least squares are the same as in the independent case, whereas long range dependence leads to slower rates.

All the proofs are given in the Sections 3.6 and 3.7.

3.2 Kernel partial least squares

We consider the nonparametric regression model (3.1) with the stationary time series $\{X_t\}_{t\in\mathbb{Z}}$ and assume that we have a training sample $\{(X_t, y_t)^{\mathrm{T}}\}_{t=1}^n$ for $n \in \mathbb{N}$.

Define with $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ the reproducing kernel Hilbert space of functions on $\mathbb{R}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, i.e., the property

$$g(x) = \langle g, k(\cdot, x)\rangle_{\mathcal{H}}, \quad x \in \mathbb{R}^d, g \in \mathcal{H}, \tag{3.2}$$

holds.

The corresponding norm in $\mathcal{H}$ will be denoted by $\|\cdot\|_{\mathcal{H}}$. We refer to Berlinet and Thomas-Agnan (2004) for examples of Hilbert spaces and their reproducing kernels and specifically to Steinwart et al. (2005) for a derivation of the reproducing kernel Hilbert space belonging to the Gaussian kernel $k(x, y) = \exp\{-s(x-y)^{\mathrm{T}}(x-y)\}$, $x, y \in \mathbb{R}^d$, $s > 0$.

To find an approximation of f^* in $\mathcal{H}$ given the training sample $\{(X_t, y_t)^{\mathrm{T}}\}_{t=1}^n$ we consider the regularized least squares problem

$$\min_{h\in\mathcal{H}} n^{-1}\sum_{t=1}^{n}\{y_t - h(X_t)\}^2 + \xi\|h\|_{\mathcal{H}}^2$$

with penalty term $\xi > 0$. By virtue of the generalized representer theorem of Schölkopf et al. (2001) the solution of this optimization problem is $f_\alpha = \sum_{t=1}^n \alpha_t k(\cdot, X_t)$ for some $\alpha_1, \ldots \alpha_n \in \mathbb{R}$.

There are a variety of different approaches to estimate the coefficients $\alpha_1, \ldots, \alpha_n$, including kernel ridge regression, kernel principal component regression or kernel partial least squares.

Kernel partial least squares was derived by Rosipal and Trejo (2001). Essentially the data X_t are mapped into $\mathcal{H}$ via $\phi_t = k(\cdot, X_t)$. This mapping can be done implicitly by using the kernel trick $\langle\phi_t, \phi_s\rangle_{\mathcal{H}} = k(X_t, X_s)$ and thus only the $n \times n$ dimensional kernel matrix $K_n =$

$n^{-1}[k(X_t, X_s)]_{t,s=1}^n$ is needed in the algorithm.

It was shown by Krämer and Braun (2007) that the kernel partial least squares algorithm solves the optimization problem

$$\widehat{\alpha}_i = \arg \min_{v \in \mathcal{K}_i(K_n, y)} \|y - K_n v\|^2, \;\; i = 1, \ldots, n, \tag{3.3}$$

with $y = (y_1, \ldots, y_n)^{\mathrm{T}}$. Here $\mathcal{K}_i(K_n, y) = \operatorname{span}\{y, K_n y, K_n^2 y, \ldots, K_n^{i-1} y\}$, $i = 1, \ldots, n$, is the ith Krylov space with respect to K_n and y and $\|\cdot\|$ denotes the Euclidean norm rescaled by n^{-1}.

The following assumptions on $\mathcal{H}$ and k are crucial for our further analysis:

(K1) $\mathcal{H}$ is separable,

(K2) There exists a $\kappa > 0$ such that $|k(x, y)| \leq \kappa$ for all $x, y \in \mathbb{R}^d$ and k is measurable.

Under condition (K1) the Hilbert-Schmidt norm $\|\cdot\|_{\mathrm{HS}}$ for operators mapping from $\mathcal{H}$ to $\mathcal{H}$ is well defined. If condition (K2) holds all functions in $\mathcal{H}$ are bounded, see Berlinet and Thomas-Agnan (2004), chapter 2. Thus we have $\mathcal{H} \subseteq \mathcal{L}^2\left(\mathrm{P}^X\right)$ as every bounded function is integrable with respect to a probability measure. The condition is satisfied for a variety of popular kernels, e.g., Gaussian or triangular.

The change of space operator $T : \mathcal{H} \to \mathcal{L}^2\left(\mathrm{P}^X\right), g \mapsto g$ is well defined if (K2) holds and the kernel integral operator is given by $T^* : \mathcal{L}^2\left(\mathrm{P}^X\right) \to \mathcal{H}, g \mapsto \mathrm{E}\{k(\cdot, X) g(X)\}$. It is easy to see that T, T^* are adjoint, i.e., for $u \in \mathcal{H}$ and $v \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ it holds $\langle T^* v, u\rangle_{\mathcal{H}} = \langle v, Tu\rangle_2$ with $\langle\cdot,\cdot\rangle_2$ being the inner product in $\mathcal{L}^2\left(\mathrm{P}^X\right)$. This property is useful as we will measure the estimation error of $f_{\widehat{\alpha}_i}$ in the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-norm $\|\cdot\|_2$.

The sample analogues of T, T^* are $T_n : \mathcal{H} \to \mathbb{R}^n, g \mapsto \{g(X_1), \ldots, g(X_n)\}^{\mathrm{T}}$ and $T_n^* : \mathbb{R}^n \to \mathcal{H}, (v_1, \ldots, v_n)^{\mathrm{T}} \mapsto n^{-1} \sum_{t=1}^n v_t k(\cdot, X_t)$, respectively. In the rescaled Euclidean product $\langle u, v\rangle = n^{-1} u^{\mathrm{T}} v$, $u, v \in \mathbb{R}^d$, both of these operators are adjoint as well.

Finally we define the sample kernel covariance operator $S_n = T_n^* T_n : \mathcal{H} \to \mathcal{H}$ and the population kernel covariance operator $S = T^* T : \mathcal{H} \to \mathcal{H}$. Note that it holds $K_n = T_n T_n^*$. Under (K1) and (K2) S is a self-adjoint compact operator with operator norm $\|S\|_{\mathcal{L}} \leq \kappa$, see Caponnetto and de Vito (2007).

In the literature of ill-posed problems it is well known that without further conditions on the target function f^* the convergence rate of the conjugate gradient algorithm can be arbitrarily slow, see (Hanke, 1995), chapter 3.2. One common a-priori assumption on the regression function f^* is the source condition:

(S) There exist $r \geq 1/2$, $R > 0$ and $u \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ such that $f^* = (TT^*)^r u$ and $\|u\|_2 \leq R$.

It is well known that in the case $r \geq 1/2$ the target function $f^* \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ coincides almost surely with a function $f \in \mathcal{H}$ and we can write $f^* = Tf$, see Cucker and Smale (2002). Hence we can define $M = \sup_{x \in \mathbb{R}^d} |f(x)|$.

For $r \geq 1/2$ the source condition can be restated for the function $f \in \mathcal{H}$ as what is also known as the Hölder source condition with $\mu = r - 1/2$

(SH) There exist $\mu \geq 0$, $R > 0$ and $u \in \mathcal{H}$ such that $f = S^\mu u$ and $\|u\|_{\mathcal{H}} \leq R$.

The condition (SH) measures the smoothness of the solution f with respect to S in $\mathcal{H}$, see Bauer et al. (2007) for more details.

3.3 Consistency of kernel partial least squares

The kernel conjugate gradient algorithm as described by Blanchard and Krämer (2010b) is consistent when stopped early and explicit convergence rates can be obtained when a source condition (S) holds. Here we will proof the same property for kernel partial least squares. Early stopping in this context means that we stop the algorithm at some $a \leq n$ that depends on n and consider the estimator $f_{\widehat{\alpha}_a}$.

For $x \in \mathbb{R}^n$ we define $\|x\|_{K_n}^2 = n^{-1}x^{\mathrm{T}}K_n x$. The difference between the two methods is the norm to be optimized. The kernel conjugate gradient algorithm of Blanchard and Krämer (2010b) estimates the coefficients $\alpha \in \mathbb{R}^n$ of f_α via $\widehat{\alpha}_i^{CG} = \arg\min_{v\in\mathcal{K}_i(K_n,y)} \|y - K_n v\|_{K_n}^2$ as opposed to kernel partial least squares $\widehat{\alpha}_i = \arg\min_{v\in\mathcal{K}_i(K_n,y)} \|y - K_n v\|^2$, $i = 1, \dots, n$, see (3.3).

It is easy to see that the optimization problems can be rewritten for the function f_α as $\min_{g\in\mathcal{K}_i(S_n,T_n^*y)} \|T_n^* y - S_n g\|_{\mathcal{H}}^2$ and $\min_{g\in\mathcal{K}_i(S_n,T_n^*y)} \|y - T_n g\|^2$, respectively. Thus kernel conjugate gradient obtains the least squares solution g in the $\mathcal{H}$-norm for the normal equation $T_n^*y = S_n g$ and kernel partial least squares finds a function that minimizes the residual sum of squares $\|y - \{g(X_1), \dots, g(X_n)\}^{\mathrm{T}}\|^2$. In both methods the solutions are restricted to functions $g \in \mathcal{K}_i(S_n, T_n^*y)$.

An advantage of the formulation for the kernel conjugate gradient estimator is that concentration inequalities can be established for both T_n^*y and S_n and applied directly as the optimization function contains both quantities. The stopping index for the regularization can be chosen by a discrepancy principle as $a^* = \min\{1 \le i \le n : \|S_n f_{\widehat{\alpha}_i^{CG}} - T_n^* y\| \le \Lambda_n\}$ with Λ_n being a threshold sequence that goes to zero as n increases.

For the kernel partial least squares optimization problem, on the other hand, the function to be optimized contains only y and $T_n g = \{g(X_1), \dots, g(X_n)\}^{\mathrm{T}}$ for which statistical properties are not readily available. Thus we need to find a way to apply the concentration inequalities for T_n^*y and S_n to this slightly different problem. This leads to complications in the proof of consistency and a rather different and more technical stopping rule for choosing the optimal regularization parameter a^* is used, as can be seen in Theorem 3.1. This stopping rule has its origin in Hanke (1995).

The next theorem states the convergence properties of the kernel partial least squares algorithm when a source condition holds.

Theorem 3.1 *Assume that conditions (K1), (K2), (S) hold with* $r \geq 3/2$ *and there are constants* $C_\delta, C_\epsilon > 0$ *and a sequence* $\{\gamma_n\}_{n\in\mathbb{N}} \subset [0,\infty)$*,* $\gamma_n \to 0$*, such that we have for* $\nu \in (0,1]$

$$\mathrm{P}\left(\|S_n - S\|_{\mathcal{L}} \leq C_\delta \gamma_n\right) \geq 1 - \nu/2, \quad \mathrm{P}\left(\|T_n^* y - Sf\|_{\mathcal{H}} \leq C_\epsilon \gamma_n\right) \geq 1 - \nu/2.$$

Define the stopping index a^* *by*

$$a^* = \min\left\{1 \leq a \leq n : \sum_{i=0}^{a} \|S_n f_{\widehat{\alpha}_i} - T_n^* y\|_{\mathcal{H}}^{-2} \geq (C\gamma_n)^{-2}\right\}, \tag{3.4}$$

with $C = C_\epsilon + R\{1 + C_\delta(r + 1/2)\kappa^{r-1/2}\}$*.*

Then it holds with probability at least $1 - \nu$ *that*

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = O\left\{\gamma_n^{2r/(2r+1)}\right\},$$

$$\|f_{\widehat{\alpha}_{a^*}} - f\|_{\mathcal{H}} = O\left\{\gamma_n^{(2r-1)/(2r+1)}\right\}.$$

It can be shown that the stopping rule (3.4) always determines a finite index, i.e., the set the minimum is taken over is not empty, see Hanke (1995), chapter 4.3.

The convergence rate of the kernel partial least squares estimator depends crucially on the sequence γ_n and the source parameter r. If $\gamma_n = O(n^{-1/2})$, this yields the same convergence rate as Theorem 2.1 of Blanchard and Krämer (2010b) for kernel conjugate gradient or de Vito et al. (2005) for kernel ridge regression with independent and identically distributed data.

The optimal stopping index a^* is of a theoretical nature. The constants C_δ and C_ϵ require decent knowledge about the estimators S_n and $T_n^* y$ and the stopping index also depends on the source parameter r, which is unlikely to be available in every application.

Recall again that the target function fulfils $f^* = Tf$ almost surely for an $f \in \mathcal{H}$ under (S) and that $\mu = r + 1/2$. As the source condition is crucial for Theorem 3.1 the following proposition will derive a more explicit representation for $f \in \mathcal{H}$ if (SH) holds:

Proposition 3.1 *Assume that (K1),(K2) and (SH) hold. Then we have for $\mu \in \mathbb{N}$*

(i) $\|f\|_{\mathcal{H}} \leq R\kappa^{\mu}$.

If additionally $d = 1$ and $X_0 \sim \mathcal{N}(0, \sigma^2), \sigma^2 > 0$, holds and the Gaussian kernel $k(x,y) = \exp\{-s(x-y)^2\}$ for $x, y \in \mathbb{R}$, $s > 0$, is used we have for $\mu \in \mathbb{N}$

(ii) $\|f\|^2_{\mathcal{H}} \leq R \left\{ \sum\limits_{i=0}^{\mu} \beta_{i,\mu}(s\sigma^2)^i \right\}^{-1/2}$,

for coefficients $\{\beta_{i,\mu}\}_{i=0}^{\mu} \subset (0,\infty)$ and $\beta_{0,\mu} = 1$,

(iii) f can be expressed via $f(x) = \sum_{i=1}^{\infty} c_i L_\mu(x, z_i)$ for certain $\{z_i\}_{i=1}^{\infty}, \{c_i\}_{i=1}^{\infty} \subset \mathbb{R}$ such that $\sum_{i,j=1}^{\infty} c_i c_j k(z_i, z_j) \leq R^2$. Here we have for $x, z \in \mathbb{R}$

$$L_\mu(x,z) = \{\sigma^{2\mu} \det(\Lambda_{1:\mu})\}^{-1/2} \exp\left[-1/2 \left\{ \frac{\det(\Lambda)(x^2+z^2) - 2s^{\mu+1}xz}{\det(\Lambda_{1:\mu})} \right\}\right],$$

with $\Lambda \in \mathbb{R}^{(\mu+1)\times(\mu+1)}$ being a tridiagonal matrix with elements

$$\Lambda_{i,j} = \begin{cases} \sigma^{-2} + 2s & , \quad |i-j| = 0, i,j < \mu+1 \\ s & , \quad i = j = \mu+1 \\ -s & , \quad |i-j| = 1 \\ 0 & , \quad else \end{cases}$$

for $i, j = 1, \dots, \mu+1$ and $\Lambda_{1:\mu}$ is the $\mu \times \mu$-dimensional submatrix of Λ including the fist μ columns and rows.

Given an $u \in \mathcal{H}$ with $\|u\|_{\mathcal{H}} \leq R$ and $s\sigma^2 > 1$ the higher $\mu \in \mathbb{N}$ is chosen the smaller the norm $\|f\|_{\mathcal{H}}$ becomes, which can be interpreted as a higher degree of smoothness of the solution in $\mathcal{H}$. The first part can be applied for $\kappa < 1$ and shows the same property, whereas for $\kappa \geq 1$ this inequality becomes too coarse.

The explicit representation of the solution f under a source condition will be useful for simulations as the convergence rate in Theorem 3.1 explicitly depends on $r = \mu + 1/2$. It shows that any function satisfying the source condition is a linear combination of very specific exponential functions $L_\mu(y_i, \cdot)$, $i \in \mathbb{N}$.

3.4 Concentration inequalities

Crucial assumptions of Theorem 3.1 are the concentration inequalities for S_n and $T_n^* y$ and convergence of the sequence $\{\gamma_n\}_{n\in\mathbb{N}}$.

In this section we will derive concentration inequalities by an application of Markov's inequality and the mean squared error of both estimators is of interest.

Theorem 3.2 *Under Assumptions (K1) and (K2) it holds that*

$$\begin{aligned}
\mathrm{E}\,\|S_n - S\|_{\mathrm{HS}}^2 &= \frac{2}{n^2}\sum_{h=1}^{n-1}(n-h)\int\limits_{\mathbb{R}^{2d}} k^2(x,y)\left\{\mathrm{dP}^{X_h,X_0}(x,y) - \mathrm{dP}^{X_0}(x)\mathrm{dP}^{X_0}(y)\right\} \\
&\quad + n^{-1}\left\{\mathrm{E}\,k^2(X_0,X_0) - \|S\|_{\mathrm{HS}}^2\right\}, \\
\mathrm{E}\,\|T_n^* y - Sf\|_{\mathcal{H}}^2 &= \frac{2}{n^2}\sum_{h=1}^{n-1}(n-h)\int\limits_{\mathbb{R}^{2d}} k(x,y)f(x)f(y)\left\{\mathrm{dP}^{X_h,X_0}(x,y) - \mathrm{dP}^{X_0}(x)\mathrm{dP}^{X_0}(y)\right\} \\
&\quad + n^{-1}\left[\mathrm{E}\left\{k(X_0,X_0)f^2(X_0)\right\} - \|Sf\|_{\mathcal{H}}^2 + \sigma^2\,\mathrm{E}\,k(X_0,X_0)\right].
\end{aligned}$$

It is obvious that the convergence rate is controlled by the sums appearing on the right hand side of both equations in Theorem 3.2. If these sums are of $O(n)$ then the mean squared error of both S_n and $T_n^* y$ will converge to zero with a rate of n^{-1}. On the other hand if the sums are of order $O(n^{2-q})$ for some $q \in (0,1)$, the mean squared errors will converge with the reduced rate n^{-q}.

In the next theorem we will derive explicit convergence rates for these sums in a Gaussian setting, i.e.,

(D1) $(X_h, X_0)^{\mathrm{T}} \sim \mathcal{N}_{2d}(0, \Sigma_h)$, $h = 1, \dots, n-1$, with

$$\Sigma_h = \begin{bmatrix} \tau_0 & \tau_h \\ \tau_h & \tau_0 \end{bmatrix} \otimes \Sigma.$$

Here $\Sigma \in \mathbb{R}^{d \times d}$ and $[\tau_{|i-j|}]_{i,j=1}^n \in \mathbb{R}^{n \times n}$ are positive definite, symmetric matrices and $\otimes$ denotes the Kronecker product between matrices. Furthermore $X_0 \sim \mathcal{N}_d(0, \tau_0 \Sigma)$.

(D2) For the autocorrelation function $\rho_h = \tau_0^{-1} \tau_h$ there exists a $q > 0$ such that $|\rho_h| \leq (h+1)^{-q}$ for $h = 0, \dots, n-1$.

Condition (D1) is a separability condition for the covariance matrices Σ_h, $h = 0, \dots, n-1$. Under condition (D2) it is easy to see that from $q > 1$ follows the absolute summability of the autocorrelation function ρ and thus $\{X_t\}_{t \in \mathbb{Z}}$ is a short memory process.

On the other hand $q \in (0, 1]$ yields long memory, see, e.g., Definition 3.1.2 in Giraitis et al. (2012). Examples of long memory processes are the fractional Gaussian noise with an autocorrelation function that behaves like $(h+1)^{-2(1-H)}$, with $H \in (0, 1]$ being the Hurst coefficient. See chapter 2 of Samorodnitsky (2007) for details.

The next theorem gives concentration inequalities for both estimators S_n and $T_n^* y$ with convergence rates depending on the parameter $q > 0$.

Theorem 3.3 *Under the assumptions (K1),(K2),(D1) and (D2) it holds for* $\nu \in (0, 1]$ *with probability at least* $1 - \nu$

$$\|S_n - S\|_{\mathrm{HS}}^2 \leq \varphi_n(q) \frac{\kappa^2 d^{1/2}}{\nu \{(2\pi)^d \det(\Sigma)\}^{1/2}} (1 - 4^{-q})^{-1/4(d-2)} + n^{-1} \frac{\kappa^2 - \|S\|_{\mathrm{HS}}^2}{\nu},$$
$$\|T_n^* y - Sf\|_{\mathcal{H}}^2 \leq \varphi_n(q) \frac{\kappa M d^{1/2}}{\nu \{(2\pi)^d \det(\Sigma)\}^{1/2}} (1 - 4^{-q})^{-1/4(d-2)} + n^{-1} \frac{\kappa [M + \sigma^2] - \|Sf\|_{\mathcal{H}}^2}{\nu}.$$

The function $\varphi_n(q)$, $q > 0$, is defined as

$$\varphi_n(q) = \begin{cases} n^{-1}\zeta(q) & , \quad q > 1 \\ n^{-1}\log(n)\{5 - \log(4)\} & , \quad q = 1 \\ n^{-q}\left[\{2(1-q)^{-1} - (2-q)^{-1}\} + (2-q)^{-1}2^{2-q}\right] & , \quad q \in (0,1), \end{cases}$$

with ζ being the Riemann-zeta function.

The theorem shows that for $q > 1$ both estimators are $\sqrt{n}$ consistent. For $q \in (0,1]$ the convergence rates slow down significantly as the long range dependence affects the estimation procedure.

Together with Theorem 3.1 Theorem 3.3 implies

Corollary 3.1 *Assume that the conditions of Theorem 3.1 hold. Assume also (D1) and (D2). Then we have with probability at least* $1 - \nu$

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = \begin{cases} O\{n^{-r/(2r+1)}\}, & q > 1, \\ O\{n^{-qr/(2r+1)}\}, & q \in (0,1). \end{cases}$$

3.5 Simulations

We set $d = 1$ and $\Sigma = \sigma^2 = 4$.

For the matrix $V^2 = [\tau_{|i-j|}]_{i,j=1}^n \in \mathbb{R}^{n \times n}$ we choose three different structures. The first setting is $\tau_h = I(h = 0)$, corresponding to independent data, the second $\tau_h = 0.9^{-h}$ for an autoregressive process of order one and the third is the long range dependent case with $\tau_h = (1+h)^{-1/4}$, $h = 0, \ldots, n-1$ and $n = 200, 400, \ldots, 1000$.

We use the Gaussian kernel $k(x_1, x_2) = \exp\{-s(x_1 - x_2)^2\}$, $x_1, x_2 \in \mathbb{R}$ for $s = 2$. The regression function is chosen as $f(x) = 4.37^{-1}\{3\tilde{L}_4(x,-4) - 2\tilde{L}_4(x,3) + 1.5\tilde{L}_4(x,9)\}$, $x \in \mathbb{R}$.

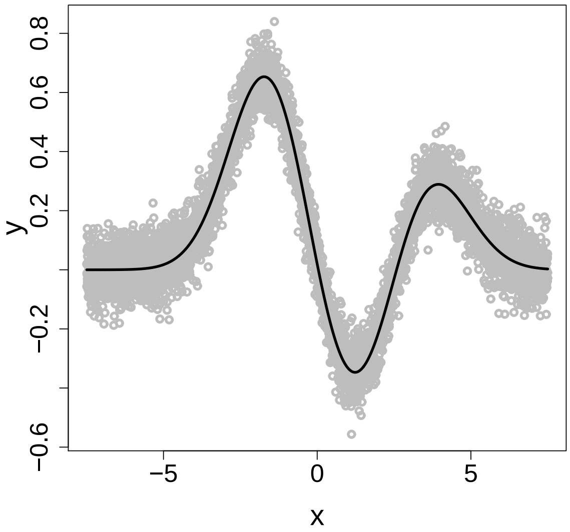

Figure 3.1: The function f evaluated on $[-7.5, 7.5]$ (black) and one realisation of the noisy data $y = f(x) + \varepsilon$ (grey).

The normalization constant is chosen such that f takes values in $[-0.35, 0.65]$. The function $\tilde{L}_4$ is the exponential function L_4 given in Proposition 3.1 normalized to take values in $[0, 1]$ and hence the source condition (SH) is fulfilled with $\mu = 4$.

The function can be seen in Figure 3.1.

In a Monte Carlo simulation with $l = 1000$ repetitions the time series $\{X_t^{(j)}\}_{t=1}^n$ are generated via $X^{(j)} = VN^{(j)}$ with $N^{(j)} \sim \mathcal{N}_n(0, \sigma^2 I_n)$, $j = 1, \ldots, l$. We denote with I_n the $n \times n$-dimensional identity matrix. All Monte-Carlo samples are independent of each other.

The residuals $\varepsilon_1^{(j)}, \ldots, \varepsilon_n^{(j)}$ are generated as independent standard normally distributed random variables and independent of $\{X_t^{(j)}\}_{t=1}^n$. The response is defined as $y_t^{(j)} = f(X_t^{(j)}) + \eta\, \varepsilon_t^{(j)}$, $t = 1, \ldots, n$, $j = 1, \ldots, l$ with $\eta = 1/16$.

The kernel partial least squares and kernel conjugate gradient algorithms are run for each sample $\{(X_t^{(j)}, y_t^{(j)})^{\mathrm{T}}\}_{t=1}^n$, $j = 1, \ldots, l$, as described in Rosipal and Trejo (2001) and Blanchard and Krämer (2010b), respectively, with a maximum of 40 iteration steps. We denote the estimated coefficients with $\widehat{\alpha}_1^{(j,m)}, \ldots, \widehat{\alpha}_{40}^{(j,m)}$, $j = 1, \ldots, l$, with $m = CG$ meaning that the kernel

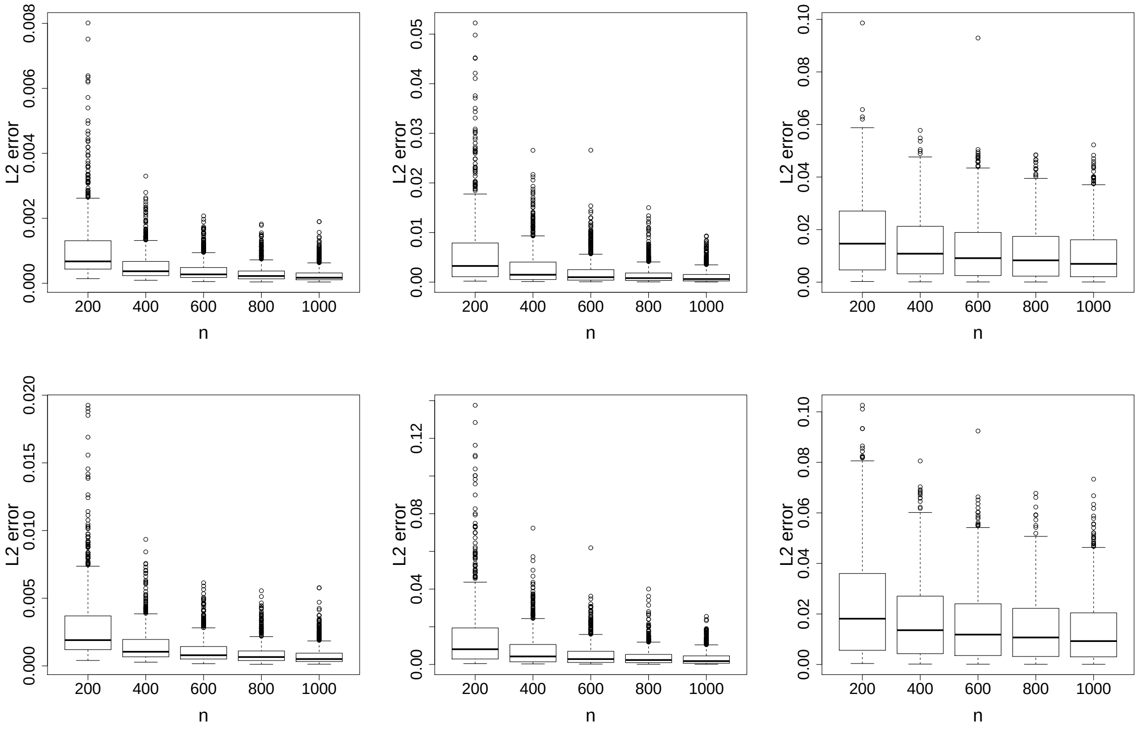

Figure 3.2: Boxplots of the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-errors $\{\widehat{e}^{(j,m)}_{n,\tau}\}_{j=1}^{1000}$ of kernel partial least squares (top) and kernel conjugate gradient (bottom) for different autocovariance functions τ and $n = 200, 400, \ldots, 1000$. On the left is $\tau_h = I(h = 0)$, in the middle $\tau_h = 0.9^{-h}$ and on the right $\tau_h = (h+1)^{-1/4}$.

conjugate gradient algorithm was employed and $m = PLS$ that kernel partial least squares was used to estimate $\alpha_1, \ldots, \alpha_n$.

The squared error in the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-norm is calculated via

$$\widehat{e}^{(j,m)}_{n,\tau} = \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} \left\{ f_{\widehat{\alpha}_a^{(j,m)}}(x) - f(x) \right\}^2 \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}x,$$

for $j = 1, \ldots, l$, $n = 200, 400, \ldots, 1000$ and $m \in \{CG, PLS\}$.

The results of the Monte-Carlo simulations can be seen in the boxplots of Figure 3.2. For kernel partial least squares it can be seen that the independent case on the left and the case of autoregressive dependence have roughly the same convergence rates, although the latter case has a generally higher error. On the other hand in the case of long range dependence we see

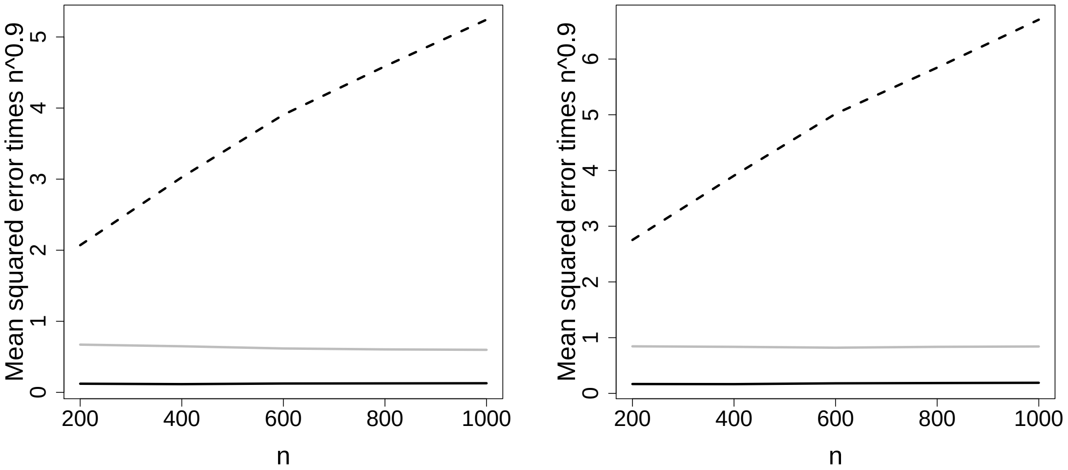

Figure 3.3: Mean of the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-errors $\{\widehat{e}_{n,\tau}^{(j,m)}\}_{j=1}^{1000}$ of kernel partial least squares (left) and kernel conjugate gradient (right) for $n = 200, 400, \dots, 1000$ multiplied by $n^{2r/(2r+1)}$ for $r = 4.5$. The solid black line is for $\tau_h = I(h = 0)$, the grey line for $\tau_h = 0.9^{-h}$ and the dashed black line for $\tau_h = (h+1)^{-1/4}$.

that the convergence is slower and the interquartile range is larger at the same time, supporting the theoretical results of Corollary 3.1.

On the other hand the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-error of kernel conjugate gradient is generally a bit higher in this simulation study than that of kernel partial least squares for all settings. Nonetheless, qualitatively both of them are very similar.

Figure 3.3 shows the mean of the estimated $\mathcal{L}^2\left(\mathrm{P}^X\right)$ errors $\{\widehat{e}_{n,\tau}^{(j,m)}\}_{j=1}^{1000}$ for different n, τ and $m \in \{CG, PLS\}$. The errors were multiplied by $n^{2r/(2r+1)}$ for $r = \mu + 0.5 = 4.5$ to illustrate the convergence rates. According to Corollary 3.1 we expect the rates for the independent and autoregressive cases to be $n^{-2r/(2r+1)}$ which is verified by the fact that the solid black and grey lines are roughly constant. For the long range dependent case we expect worse convergence rates which are also illustrated by the divergence of the dashed black line.

3.6 Proof of Theorem 3.1

The proof of Theorem 3.1 makes use of the connection between the partial least squares and the conjugate gradient algorithm. This section is structured as follows: First we will introduce the link between kernel partial least squares and kernel conjugate gradient. We will state some key facts about orthogonal polynomials and their relationship to the algorithm in Lemma 3.1. Then the consistency of kernel partial least squares is shown with the help of three error bounds that are obtained in Lemmas 3.3 – 3.5.

With a slight abuse of notation we define $f_i = f_{\widehat{\alpha}_i}$ for $i = 1, \dots, n$. We consider the kernel partial least squares algorithm as an optimization problem $\widehat{\alpha}_i = \arg\min_{v \in \mathcal{K}_i(K_n, y)} \|y - K_n v\|^2,\ i = 1, \dots, n$, and thus $f_i = T_n^* \widehat{\alpha}_i$. It is easy to see that $K_n = T_n T_n^*$ and thus

$$f_i = \arg \min_{g \in \mathcal{K}_i(S_n, T_n^* y)} \|y - T_n g\|^2,\ \ i = 1, \dots, n. \tag{3.5}$$

This is the conjugate gradient algorithm CGNE discussed in chapter 2.2 of Hanke (1995).

3.6.1 Orthogonal polynomials and some notation

Denote with $\mathcal{P}_i$ the set of polynomials of degree at most $i = 0, \dots, n$. For functions $\psi, \phi : \mathbb{R} \to \mathbb{R}$ and $r \in \mathbb{N}_0$ define the inner products $[\psi, \phi]_r = \langle \psi(S_n) T_n^* y, S_n^r \phi(S_n) T_n^* y \rangle_{\mathcal{H}}$. From the definition of the Krylov space it is immediate that every element $v \in \mathcal{K}_i(S_n, T_n^* y)$, $i = 1, \dots, n$, can be represented by a polynomial $q \in \mathcal{P}_{i-1}$ via $v = q(S_n) T_n^* y$.

The following discussion is based on Hanke (1995), chapter 2. There exist two sequences of polynomials $p_i, q_i \in \mathcal{P}_i$, $i = 0, \dots, n$, such that $f_i = q_{i-1}(S_n) T_n^* y$ with $q_{-1} = 0$ and $T_n^* y - S_n f_i = p_i(S_n) T_n^* y$. Both sequences are connected by the equation $p_i(x) = 1 - x q_{i-1}(x)$, $x \in \mathbb{R}$, and the polynomials $\{p_i\}_{i=0}^n$ are orthogonal with respect to $[\cdot, \cdot]_0$.

We will also consider other sequences of polynomials, namely $q_i^{[r]}, p_i^{[r]} \in \mathcal{P}_i$, $i = 0, \dots, n,$,

$q_{-1}^{[r]} = 0$, such that $p_i^{[r]}(x) = 1 - xq_{i-1}^{[r]}(x)$, $x \in \mathbb{R}$, and the sequence $\{p_i^{[r]}\}_{i=0}^n$ is orthogonal with respect to $[\cdot,\cdot]_r$. This yields for every $r \in \mathbb{N}_0$ a separate conjugate gradient algorithm with solution $f_i^{[r]} = q_{i-1}^{[r]}(S_n)T_n^*y \in \mathcal{K}_i(S_n, T_n^*y)$ and residuals $T_n^*y - S_nf_i^{[r]} = p_i^{[r]}(S_n)T_n^*y$, $i = 1, \dots, n$.

As S_n is self-adjoint, positive semi-definite and the kernel is bounded by κ we know that its spectrum is a subset of $[0, \kappa]$, see Caponnetto and de Vito (2007). This also implies that $\max\{\|S\|_{\mathcal{L}}, \|S_n\|_{\mathcal{L}}\} \leq \kappa$, with $\|\cdot\|_{\mathcal{L}}$ denoting the operator norm. The i distinct roots of $p_i^{[r]}$ will be denoted by $0 < x_{1,i}^{[r]} < \dots x_{i,i}^{[r]} < \kappa$, $i = 1, \dots, n$.

We will summarize some key facts about the orthogonal polynomials in the next lemma.

Lemma 3.1 *Let $r, s \in \mathbb{N}_0$ and $i = 1, \dots, n$. Then we have:*

(i) The roots of consecutive orthogonal polynomials interlace, i.e., for $j = 1, \dots, i$ it holds

$$0 < x_{j,i+1}^{[r]} < x_{j,i}^{[r]} < x_{j,i}^{[r+1]} < x_{j+1,i+1}^{[r]} < x_{j+1,i}^{[r]} < \dots < x_{i,i}^{[r+1]} < x_{i+1,i+1}^{[r]} < \kappa,$$

(ii) the optimality property $[p_i^{[1]}, p_i^{[1]}]_0^{1/2} = \|T_n^*y - S_nf_i^{[1]}\|_{\mathcal{H}} \leq \|T_n^*y - S_nh\|_{\mathcal{H}}$ holds for all $h \in \mathcal{K}_i(S_n, T_n^*y)$,

(iii) on $x \in [0, x_{1,i}^{[r]}]$ it holds $0 \leq p_i^{[r]}(x) \leq 1$ and $q_i^{[r]}(x) \leq \left|\left(p_i^{[r]}\right)'(0)\right|$,

(iv) $p_n^{[r]} = p_n^{[s]}$,

(v) $\left(p_i^{[r]}\right)'(0) = -\sum_{j=1}^i \left(x_{j,i}^{[r]}\right)^{-1}$,

(vi) define $\phi_i(x) = p_i^{[r]}(x)\left(x_{1,i}^{[r]}\right)^{1/2}\left(x_{1,i}^{[r]} - x\right)^{-1/2}$ for $x \in [0, x_{1,i}^{[r]}]$, $i = 1, \dots, n$. Then it holds for $u \geq 0$ that $x^u\phi_i^2(x) \leq u^u\left|\left(p_i^{[r]}\right)'(0)\right|^{-u}$ with the convention $0^0 = 1$.

Proof: (i) See Hanke (1995), Corollary 2.7.

(ii) See Hanke (1995), Proposition 2.1.

(iii) Due to part (i) we know that all i roots of the polynomial $p_i^{[r]}$ are contained in $(0, \kappa)$. Furthermore $p_i^{[r]}(0) = 1 - 0q_i^{[r]} = 1$. Thus $p_i^{[r]}$ is convex and falling in $[0, x_{1,i}^{[r]}]$ and the first assertion follows.

Because of the convexity of $p_i^{[r]}$ on $[0, x_{1,i}^{[r]}]$ we get $q_i^{[r]}(x) = x^{-1}\{1 - p_i^{[r]}(x)\} \leq \left| \left(p_i^{[r]}\right)' (0) \right|$.

(iv) See the discussion in Hanke (1995) preceding Proposition 2.1 and use the facts that $T_n^* y \in \mathrm{range}(S_n)$ and S_n is an operator of rank n.

(v) Write $p_i^{[r]}(x) = \prod_{j=1}^{i}(1 - x/x_{j,i}^{[r]})$, $x \in [0, \kappa]$, and the result is immediate.

(vi) See equation (3.10) in Hanke (1995). □

We denote for $x \geq 0$ by P_x the orthogonal projection operator on the eigenspace corresponding to the eigenvalues of S_n that are smaller or equal x and $P_x^\perp = I_\mathcal{H} - P_x$ with $I_\mathcal{H} : \mathcal{H} \to \mathcal{H}$ being the identity operator.

3.6.2 Preparation for the proof

An important technical result that will be useful in the upcoming proof is

Lemma 3.2 *Let $B, C : \mathcal{H} \to \mathcal{H}$ be two positive semi-definite, self-adjoint operators with $\max\{\|B\|_\mathcal{L}, \|C\|_\mathcal{L}\} \leq \kappa$. Then it holds for any $r \geq 0$ with $\zeta = \max\{r - 1, 0\}$*

$$\|B^r - C^r\|_\mathcal{L} \leq (\zeta + 1)\kappa^\zeta \|B - C\|_\mathcal{L}^{r-\zeta}.$$

Proof: See Blanchard and Krämer (2010b), Lemma A.6. □

For the remainder of the proof we assume that we are on the set where it holds with probability at least $1 - \nu$, $\nu \in (0, 1]$, that $\|S_n - S\|_\mathcal{L} \leq \delta$ and $\|T_n^* y - Sf\| \leq \epsilon$. Here we take $\delta = C_\delta \gamma_n$, $\epsilon = C_\epsilon \gamma_n$ for a sequence $\{\gamma\}_n$ converging to zero and constants $C_\delta, C_\epsilon > 0$.

The stopping rule (3.4) is given by $a^* = \min\left\{a = 1, \ldots, n : \sum_{i=0}^{a} \|S_n f_i - T_n^* y\|_\mathcal{H}^{-2} \geq (C\gamma_n)^{-2}\right\}$ for $C = C_\epsilon + R + C_\delta(\mu + 1)\kappa^\mu R$.

With Lemma 2.4 in Hanke (1995) we see that this stopping iteration can also be expressed as

$$a^* = \min\left\{1 \leq a \leq n : \|S_n f_a^{[1]} - T_n^* y\|_{\mathcal{H}} \leq C\gamma_n\right\}, \tag{3.6}$$

i.e., we stop the kernel partial least squares algorithm when a discrepancy principle for $f_a^{[1]}$ holds.

Recall that $\mathcal{H} \subseteq \mathcal{L}^2\left(\mathrm{P}^X\right)$ and $T : \mathcal{H} \to \mathcal{L}^2\left(\mathrm{P}^X\right)$ is the change of space operator. Using the fact that T, T^* are adjoint operators, $f_{a^*} = Tf_{a^*}$ and $f^* = Tf$ we see

$$\|f_{a^*} - f^*\|_2 = \|T(f_{a^*} - f)\|_2 = \langle S(f_{a^*} - f), f_{a^*} - f\rangle_{\mathcal{H}} = \|S^{1/2}(f_{a^*} - f)\|_{\mathcal{H}}.$$

An application of Lemma 3.2 with $r = 1/2$ and the definition of δ yields

$$\begin{aligned}\|f_{a^*} - f^*\|_2 &= \|S^{1/2}(f_{a^*} - f)\|_{\mathcal{H}} \leq \|S^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} + \|S^{1/2}(f_{a^*}^{[1]} - f)\|_{\mathcal{H}} \\ &\leq \delta^{1/2}\left(\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} + \|f_{a^*}^{[1]} - f\|_{\mathcal{H}}\right) + \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} + \|S_n^{1/2}(f_{a^*}^{[1]} - f)\|_{\mathcal{H}}. \end{aligned} \tag{3.7}$$

The following lemmas will deal with bounding the quantities in (3.7). Recall that $\mu = r - 1/2$ is the source parameter in (SH).

Lemma 3.3 *Assume $C_x \in (0, 1]$ such that $x_* = (C_x\gamma)^{1/(\mu+1)} < x_{1,a^*-1}^{[1]}$ and $C > C_\epsilon + C_x R + C_\delta(\mu + 1)\kappa^\mu R$. Under the conditions of the theorem it holds*

$$\begin{aligned}\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} &\leq \gamma^{\mu/(\mu+1)} \frac{C}{C_x^{1/(\mu+1)}\left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]^2} \\ \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} &\leq \gamma^{(2\mu+1)/(2\mu+2)} \frac{C}{C_x^{1/(2\mu+2)}\left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]}.\end{aligned}$$

Proof: If the inner products $[\cdot,\cdot]_0$ and $[\cdot,\cdot]_1$ are the same the proof is done because both polynomial sequences are identical.

We now observe that we have for $a^* = n$ due to Lemma 3.1 (iv) $q_{n-1}(x) - q^{[1]}_{n-1}(x) = x^{-1}\{p^{[1]}_n(x) - p_n(x)\} = 0$, i.e., $\|f_{a^*} - f^{[1]}_{a^*}\|_{\mathcal{H}} = 0$ and $\|S_n^{1/2}(f_{a^*} - f^{[1]}_{a^*})\|_{\mathcal{H}} = 0$ and the proof is done.

If the inner products differ and we have $0 < a^* < n$ it holds $f_{a^*} \neq f^{[1]}_{a^*}$.

Proposition 2.8 in Hanke (1995) can now be applied for $0 < a^* < n$ and yields $q_{a^*-1}(x) - q^{[1]}_{a^*-1}(x) = x^{-1}\{p^{[1]}_{a^*}(x) - p_{a^*}(x)\} = \theta_{a^*}p^{[2]}_{a^*-1}(x)$, $x \geq 0$, with $\theta_{a^*} = (p^{[1]}_{a^*})'(0) - (p^{[0]}_{a^*})'(0) > 0$. We get $f_{a^*} - f^{[1]}_{a^*} = q_{a^*-1}(S_n)T^*_n y - q^{[1]}_{a^*-1}(S_n)T^*_n y = \theta_{a^*}p^{[2]}_{a^*-1}(S_n)T^*_n y$.

Proposition 2.9 in Hanke (1995) yields $\theta_{a^*} = \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{-1}\left[p^{[1]}_{a^*}, p^{[1]}_{a^*}\right]_0$. The optimality property of $f^{[1]}_{a^*}$ in Lemma 3.1 (ii) shows that

$$\|T^*_n y - S_n f^{[1]}_{a^*}\|_{\mathcal{H}} = \|p^{[1]}_{a^*}(S_n)T^*_n y\|_{\mathcal{H}} = \left[p^{[1]}_{a^*}, p^{[1]}_{a^*}\right]_0^{1/2} \leq \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2}. \tag{3.8}$$

Combining these results yields

$$\|f_{a^*} - f^{[1]}_{a^*}\|_{\mathcal{H}} = \frac{\left[p^{[1]}_{a^*}, p^{[1]}_{a^*}\right]_0}{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} \leq \frac{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0}{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1}\|p^{[1]}_{a^*}(S_n)T^*_n y\|_{\mathcal{H}}. \tag{3.9}$$

Recall that $x^{[2]}_{1,a^*-1}$ denotes the first root of $p^{[2]}_{a^*-1}$. It holds for any $0 \leq x \leq x^{[2]}_{1,a^*-1}$ that $0 \leq p^{[2]}_{a^*-1}(x) \leq 1$, see Lemma 3.1 (iii), and thus

$$\begin{aligned}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} &\leq \|P_x p^{[2]}_{a^*-1}(S_n)\{T^*_n y - Sf + Sf\}\|_{\mathcal{H}} + \|P_x^{\perp}p^{[2]}_{a^*-1}(S_n)T^*_n y\|_{\mathcal{H}} \\ &\leq \epsilon + \|P_x p^{[2]}_{a^*-1}(S_n)S^{\mu+1}u\|_{\mathcal{H}} + x^{-1/2}\|P_x^{\perp}S_n^{1/2}p^{[2]}_{a^*-1}(S_n)T^*_n y\|_{\mathcal{H}} \\ &\leq \epsilon + x^{\mu+1}R + \|P_x p^{[2]}_{a^*-1}(S^{\mu+1} - S_n^{\mu+1})u\|_{\mathcal{H}} + x^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}.\end{aligned}$$

In the second inequality (SH) with $\mu \geq 0$ and the definitions of δ and ϵ were applied.

By assumption $x_* = (C_x\gamma)^{1/(\mu+1)} \leq x^{[1]}_{1,a^*-1} < x^{[2]}_{1,a^*-1}$ due to the interlacing property of the roots of the polynomials $p_i^{[r]}$, $i = 1, \dots, n$, $r \in \mathbb{N}_0$, see Lemma 3.1 (i).

Using Lemma 3.2 we get $\|S^{\mu+1} - S_n^{\mu+1}\|_{\mathcal{L}} \leq (\mu+1)\kappa^\mu \delta$ and setting $x = x_*$ we get

$$\begin{aligned}\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2} &\leq \epsilon + x_*^{\mu+1} R + \delta(\mu+1)\kappa^\mu R + x_*^{-1/2}\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1^{1/2} \\ &= \gamma\left\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\right\} + x_*^{-1/2}\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1^{1/2}. \qquad (3.10)\end{aligned}$$

Due to (3.6) and (3.8) we have additionally $C\gamma \leq \|S_n f_{a^*-1}^{[1]} - T_n^* y\|_{\mathcal{H}} = \|p_{a^*-1}^{[1]}(S_n) T_n^* y\|_{\mathcal{H}} \leq \left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2}$.

Plugging this into (3.10) yields

$$\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2} \leq C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2} + x_*^{-1/2}\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1^{1/2},$$

or equivalently

$$\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2} \leq \gamma^{-1/(2\mu+2)} \frac{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1^{1/2}}{C_x^{1/(2\mu+2)}\left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]}, \qquad (3.11)$$

where by assumption $C > C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R$ and $x_* = (C_x\gamma)^{1/(\mu+1)}$.

Combining (3.9), (3.11) and $\|p_{a^*}^{[1]}(S_n) T_n^* y\|_{\mathcal{H}} \leq C\gamma$ due to the stopping index (3.6) yields

$$\begin{aligned}\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} &\leq \gamma^{-1/(\mu+1)} \frac{\|p_{a^*}^{[1]}(S_n) T_n^* y\|_{\mathcal{H}}}{C_x^{1/(\mu+1)}\left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]^2} \\ &\leq \gamma^{\mu/(\mu+1)} \frac{C}{C_x^{1/(\mu+1)}\left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]^2}.\end{aligned}$$

For the second part of the proof we derive in the same way as (3.9)

$$\|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} \leq \frac{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2}}{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1^{1/2}} \|p_{a^*}^{[1]}(S_n)T_n^* y\|_{\mathcal{H}}.$$

Using again (3.11) and $\|p_{a^*}^{[1]}(S_n)T_n^* y\|_{\mathcal{H}} \leq C\gamma$ gives

$$\|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} \leq \gamma^{(2\mu+1)/(2\mu+2)} \frac{C}{C_x^{1/(2\mu+2)} \left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]},$$

finishing the proof. □

Lemma 3.4 *For any* $i = 1, \ldots, n$, $\mu \geq 1$ *and any* $0 < x \leq x_{1,i}^{[1]}$ *we have under the conditions of the theorem*

$$\begin{aligned}\|f - f_i^{[1]}\|_{\mathcal{H}} &\leq R\left\{x^\mu + \delta\mu\kappa^{\mu-1}\right\} + x^{-1}\left(\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + \epsilon + \delta\kappa^\mu R\right) \\ &\quad + (\epsilon + \delta\kappa^\mu R)|(p_i^{[1]})'(0)|, \\ \|S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}} &\leq R\left\{x^{\mu+1/2} + x^{1/2}\delta\mu\kappa^{\mu-1}\right\} + x^{-1/2}\left(\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + \epsilon + \delta\kappa^\mu R\right) \\ &\quad + x^{1/2}(\epsilon + \delta\kappa^\mu R)|(p_i^{[1]})'(0)|.\end{aligned}$$

Proof: Denote $\bar{f}_i = q_{i-1}^{[1]}(S_n)S_n f$ and consider

$$\|f - f_i^{[1]}\|_{\mathcal{H}} \leq \|P_x(f - \bar{f}_i)\|_{\mathcal{H}} + \|P_x(\bar{f}_i - f_i^{[1]})\|_{\mathcal{H}} + \|P_x^\perp(f - f_i^{[1]})\|_{\mathcal{H}}. \qquad (3.12)$$

The first term of (3.12) can be bound by an application of Lemma 3.2 and (SH) with $\mu \geq 1$

$$\begin{aligned} \|P_x(f - \bar{f}_i)\|_{\mathcal{H}} &= \|P_x\{I - q_{i-1}^{[1]}(S_n)S_n\}f\|_{\mathcal{H}} = \|P_x p_i^{[1]}(S_n)f\|_{\mathcal{H}} = \|P_x p_i^{[1]}(S_n)S^\mu u\|_{\mathcal{H}} \\ &\leq \|P_x p_i^{[1]}(S_n)S_n^\mu u\|_{\mathcal{H}} + \|P_x p_i^{[1]}(S_n)(S^\mu - S_n^\mu)u\|_{\mathcal{H}} \\ &\leq R\left\{x^\mu + \delta\mu\kappa^{\mu-1}\right\}. \end{aligned}$$

In the last inequality we used that on $0 \leq x \leq x_{1,i}^{[1]}$ we have $0 \leq p_i^{[1]}(x) \leq 1$.

For the second term of (3.12) we use Lemma 3.1 (iii) $q_i^{[1]}(x) \leq |(p_i^{[1]})'(0)|$ on $x \in [0, x_{1,i}^{[1]}]$. This yields

$$\begin{aligned} \|P_x(f_i^{[1]} - \bar{f}_i)\|_{\mathcal{H}} &= \|P_x q_i^{[1]}(S_n)(S_n f - T_n^* y)\|_{\mathcal{H}} \\ &\leq \|P_x q_i^{[1]}(S_n)(Sf - T_n^* y)\|_{\mathcal{H}} + \|P_x q_i^{[1]}(S_n)(S_n - S)f\|_{\mathcal{H}} \\ &\leq (\epsilon + \delta\kappa^\mu R)\left|\left(p_i^{[1]}\right)'(0)\right|. \end{aligned}$$

Finally, we have

$$\begin{aligned} \|P_x^\perp(f - f_i^{[1]})\|_{\mathcal{H}} \leq x^{-1}\|P_x^\perp S_n(f - f_i^{[1]})\|_{\mathcal{H}} &\leq x^{-1}\left\{\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + \|P_x(T_n^* y - S_n f)\|_{\mathcal{H}}\right\} \\ &\leq x^{-1}\left(\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + \epsilon + \delta\kappa^\mu R\right) \end{aligned}$$

and thus the first inequality is proven.

For the second inequality we use

$$\|S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}} \leq \|P_x S_n^{1/2}(f - \bar{f}_i)\|_{\mathcal{H}} + \|P_x S_n^{1/2}(\bar{f}_i - f_i^{[1]})\|_{\mathcal{H}} + \|P_x^\perp S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}}.$$

In the same way as before we derive bounds for the three terms:

$$\|P_x S_n^{1/2}(f - \bar{f}_i)\|_{\mathcal{H}} \le x^{1/2}\delta\mu\kappa^{\mu-1}R + x^{\mu+1/2}R,$$

$$\|P_x S_n^{1/2}(\bar{f}_i - f_i^{[1]})\|_{\mathcal{H}} \le x^{1/2}(\epsilon + \delta R\kappa^{\mu})\left|\left(p_i^{[1]}\right)'(0)\right|,$$

$$\|P_x^{\perp} S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}} \le x^{-1/2}(\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + \epsilon + \delta\kappa^{\mu}R),$$

completing the proof. □

Lemma 3.5 *Assume that* $C_x \in (0,1]$ *is such that* $x_* = (C_x\gamma)^{1/(\mu+1)} < x_{1,a^*-1}^{[1]}$ *and* $C > C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R$*. Under the conditions of the theorem it holds for* $\mu \ge 1$

$$\left|\left(p_{a^*}^{[1]}\right)'(0)\right| \le \gamma^{-1/(\mu+1)}\left[C_x^{-1/(\mu+1)}\left\{1 - \frac{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R}{C}\right\}^{-2} + \left\{\frac{(2\mu+2)^{\mu+1}R}{C - C_\delta(\mu+1)\kappa^\mu R + C_\epsilon}\right\}^{1/(\mu+1)}\right]$$

Proof: The proof is done in two steps by using the inequality $\left|\left(p_{a^*}^{[1]}\right)'(0)\right| \le \left|\left(p_{a^*-1}^{[1]}\right)'(0)\right| + \left|\left(p_{a^*}^{[1]}\right)'(0) - \left(p_{a^*-1}^{[1]}\right)'(0)\right|$.

1. Consider first $a^* > 1$.

We will bound $\|S_n f_{a^*-1}^{[1]} - T_n^* y\|_{\mathcal{H}}$ from above. Define $z = x_{1,a^*-1}^{[1]}$ and $\phi_i(x) = p_i^{[1]}(x)(z - x)^{-1/2}z^{1/2}$, $0 \le x \le z$. Due to Lemma 3.1 (vi) it holds that $\sup_{0\le x\le z} x^\nu \phi_{a^*-1}^2(x) \le \nu^\nu |(p_{a^*-1}^{[1]})'(0)|^{-\nu}$, $\nu \ge 0$. The proof of Lemma 3.7 in Hanke (1995) shows that

$\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1}\right]_0^{1/2} \leq \|P_z \phi_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}}$. This yields with (SH)

$$\begin{aligned}
\|S_n f^{[1]}_{a^*-1} - T_n^* y\|_{\mathcal{H}} &= \left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1}\right]_0^{1/2} \leq \|P_z \phi_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} \\
&\leq \|P_z \phi_{a^*-1}(S_n) S f\|_{\mathcal{H}} + \|P_z \phi_{a^*-1}(S_n)(T_n^* y - S f)\|_{\mathcal{H}} \\
&\leq \|P_z \phi_{a^*-1}(S_n) S f\|_{\mathcal{H}} + \epsilon \left(\sup_{0 \leq x \leq z} \phi^2_{a^*-1} \right)^{1/2} \\
&\leq \|P_z \phi_{a^*-1}(S_n) S_n^{\mu+1} u\|_{\mathcal{H}} + \|P_z \phi_{a^*-1}(S_n)(S_n^{\mu+1} - S^{\mu+1}) u\|_{\mathcal{H}} + \epsilon \\
&\leq R \left\{ \left(\sup_{0 \leq x \leq z} x^{2\mu+2} \phi^2_{a^*-1} \right)^{1/2} + \delta(\mu+1)\kappa^\mu \left(\sup_{0 \leq x \leq z} \phi^2_{a^*-1} \right)^{1/2} \right\} + \epsilon \\
&\leq \left| \left(p^{[1]}_{a^*-1} \right)'(0) \right|^{-\mu-1} (2\mu+2)^{\mu+1} R + \delta(\mu+1)\kappa^\mu R + \epsilon.
\end{aligned}$$

This gives together with $C\gamma \leq \|S_n f^{[1]}_{a^*-1} - T_n^* y\|_{\mathcal{H}}$

$$C\gamma \leq \left| \left(p^{[1]}_{a^*-1} \right)'(0) \right|^{-\mu-1} (2\mu+2)^{\mu+1} R + \gamma \left\{ C_\delta(\mu+1)\kappa^\mu R + C_\epsilon \right\}.$$

If $C > C_\delta(\mu+1)\kappa^\mu R + C_\epsilon$ we finally have

$$\left| \left(p^{[1]}_{a^*-1} \right)'(0) \right| \leq \gamma^{-1/(\mu+1)} \left\{ \frac{(2\mu+2)^{\mu+1} R}{C - C_\delta(\mu+1)\kappa^\mu R + C_\epsilon} \right\}^{1/(\mu+1)}. \tag{3.13}$$

If $a^* = 1$ it holds $p^{[1]}_{a^*-1} = 1$ and thus $\left| \left(p^{[1]}_{a^*-1} \right)'(0) \right| = 0$ and the inequality (3.13) is true as well.

2. We will derive an upper bound on $\left| \left(p^{[1]}_{a^*} \right)'(0) - \left(p^{[1]}_{a^*-1} \right)'(0) \right|$. Due to Corollary 2.6 of Hanke (1995) we have

$$\left| \left(p^{[1]}_{a^*-1} \right)'(0) - \left(p^{[1]}_{a^*} \right)'(0) \right| \leq \frac{\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1}\right]_0}{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1}. \tag{3.14}$$

We have $0 \leq x \leq x^{[1]}_{1,a^*-1} < x^{[2]}_{1,a^*-1}$ due to the interlacing property of the roots in Lemma 3.1 (i) and thus $0 \leq p^{[2]}_{a^*-1}(x) \leq 1$ for $0 \leq x \leq x^{[2]}_{1,a^*-1}$. With that we get with (SH)

$$\begin{aligned}
&\|p^{[1]}_{a^*-1}(S_n)T_n^*y\|_{\mathcal{H}} \leq \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} \\
&\leq \|P_x p^{[2]}_{a^*-1}(S_n)T_n^*y\|_{\mathcal{H}} + x^{-1/2}\|P_x^{\perp} S_n^{1/2} p^{[2]}_{a^*-1}(S_n)T_n^*y\|_{\mathcal{H}} \\
&\leq \|P_x p^{[2]}_{a^*-1}(S_n)(T_n^*y - Sf)\|_{\mathcal{H}} + \|P_x p^{[2]}_{a^*-1}(S_n)S^{\mu+1}u\|_{\mathcal{H}} + x^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2} \\
&\leq \epsilon + R\left\{\delta(\mu+1)\kappa^{\mu} + x^{\mu+1}\right\} + x^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}.
\end{aligned}$$

For the choice $x_* = (C_x\gamma)^{1/(\mu+1)}$ we get

$$\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1}\right]_0^{1/2} \leq \gamma\left\{C_\epsilon + C_\delta(\mu+1)\kappa^{\mu}R + C_x\right\} + x_*^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}.$$

It holds $\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1}\right]_0^{1/2} = \|S_n f^{[1]}_{a^*-1} - T_n^*y\|_{\mathcal{H}} \geq C\gamma$. This yields with $C > C_\epsilon + C_xR + C_\delta(\mu+1)\kappa^{\mu}R$

$$\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1}\right]_0 \leq \gamma^{-1/(\mu+1)}C_x^{-1/(\mu+1)}\left\{1 - \frac{C_\epsilon + C_xR + C_\delta(\mu+1)\kappa^{\mu}R}{C}\right\}^{-2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1.$$

Together with (3.14) we have

$$\left|\left(p^{[1]}_{a^*-1}\right)'(0) - \left(p^{[1]}_{a^*}\right)'(0)\right| \leq \gamma^{-1/(\mu+1)}C_x^{-1/(\mu+1)}\left\{1 - \frac{C_\epsilon + C_xR + C_\delta(\mu+1)\kappa^{\mu}R}{C}\right\}^{-2}.$$

Combining this with (3.13) completes the proof. □

3.6.3 Proof of Theorem 3.1

The proof is an application of Lemmas 3.3 - 3.5 to (3.7). First note that $r \geq 3/2$ implies $\mu \geq 1$ and thus this condition in Lemma 3.4 holds.

Let us choose $x_* = (C_x\gamma)^{1/(\mu+1)}$. Lemma 3.1 (v) shows that $\left|\left(p_i^{[r]}\right)'(0)\right| = \sum_{j=1}^{i}(x_{j,i}^{[r]})^{-1}$ for $i = 1, \ldots, n$, $r \in \mathbb{N}_0$. Thus it holds $\left|\left(p_i^{[1]}\right)'(0)\right|^{-1} \leq x_{1,i}^{[1]}$.

Equation (3.13) thus shows that C_x can be chosen small enough such that

$$x_* \leq \left|\left(p_{a^*-1}^{[1]}\right)'(0)\right|^{-1} \leq x_{1,a^*-1}^{[1]}$$

and $C_x < 1$, which makes the first condition in Lemma 3.3 and 3.5 hold true. The choice $C = C_\epsilon + R + C_\delta(\mu+1)\kappa^\mu R$ gives the second condition.

Now we need to check the remaining condition of Lemma 3.4, namely that a C_z can be chosen such that $(C_z\gamma)^{1/(\mu+1)} \leq x_{1,a^*}^{[1]}$ is true. Lemma 3.5 yields a $C_z > 0$ such that $C_z\gamma^{1/(\mu+1)} \leq \left|\left(p_{a^*}^{[1]}\right)'(0)\right|^{-1} \leq x_{1,a^*}^{[1]}$. More precisely we want

$$C_z^{1/(\mu+1)} \leq \left[C_x^{-1/(\mu+1)}\left\{1 - \frac{C_\epsilon + C_xR + C_\delta(\mu+1)\kappa^\mu R}{C}\right\}^{-2} + \left\{\frac{(2\mu+2)^{\mu+1}R}{C - C_\delta(\mu+1)\kappa^\mu R + C_\epsilon}\right\}^{1/(\mu+1)}\right]^{-1}.$$

Denote $z_* = (C_z\gamma)^{1/(\mu+1)}$ and with $x = z_*$ Lemma 3.4 can be applied.

To ease notation we will denote everything in the derived bounds that does not depend on γ as a constant c_j, $j \in \mathbb{N}$. Thus we get by combining Lemmas 3.4 and 3.5 that with probability at least $1 - \nu$

$$\begin{aligned}\|f - f_{a^*}^{[1]}\|_{\mathcal{H}}^2 &\leq c_1\gamma^{\mu/(\mu+1)} + c_2\gamma + c_3\gamma^{1-1/(\mu+1)} + c_4\gamma\left|\left(p_{a^*}^{[1]}\right)'(0)\right| \\ &\leq c_1\gamma^{\mu/(\mu+1)} + c_2\gamma + c_3\gamma^{\mu/(\mu+1)} + c_5\gamma^{1-1/(\mu+1)} = O\{\gamma^{\mu/(\mu+1)}\}\end{aligned}$$

and

$$\begin{aligned}
\|S_n^{1/2}(f - f_{a^*}^{[1]})\|_{\mathcal{H}}^2 &\leq c_6\gamma^{(\mu+1/2)/(\mu+1)} + c_7\gamma^{1/(2\mu+2)}\gamma + c_8\gamma^{-1/(2\mu+2)}\gamma + c_9\gamma^{1/(2\mu+1)}\gamma\left|\left(p_{a^*}^{[1]}\right)'(0)\right| \\
&\leq c_6\gamma^{(\mu+1/2)/(\mu+1)} + c_7\gamma^{(2\mu+3)/(2\mu+2)} + c_8\gamma^{(2\mu+1)/(2\mu+2)} + c_10\gamma^{1+1/(2\mu+2)-1/(\mu+1)} \\
&= O\{\gamma^{(2\mu+1)/(2\mu+2)}\}.
\end{aligned}$$

Finally Lemma 3.3 gives

$$\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}}^2 = O\{\gamma^{\mu/(\mu+1)}\}, \quad \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} = O\{\gamma^{(2\mu+1)/(2\mu+2)}\}.$$

Combining the above with (3.7) yields

$$\begin{aligned}
&\|f - f_{a^*}\|_{\mathcal{H}}^2 = O\{\gamma^{\mu/(\mu+1)}\}, \\
&\|f^* - f_{a^*}\|_2^2 = O\{\gamma^{1/2}\gamma^{\mu/(\mu+1)}\} + O\{\gamma^{(2\mu+1)/(2\mu+2)}\} = O\{\gamma^{(2\mu+1)/(2\mu+2)}\},
\end{aligned}$$

completing the proof with $\mu = r - 1/2$. □

3.7 Additional proofs

3.7.1 Proof of Theorem 3.2

We denote with $\mathrm{tr}(A^*B)$ the trace inner product of two Hilbert-Schmidt operators $A, B : \mathcal{H} \to \mathcal{H}$ and the tensor product $(f_1 \otimes f_2)h = \langle f_1, h\rangle_{\mathcal{H}} f_2$ for functions $f_1, f_2, h \in \mathcal{H}$. We use the notation $k_t = k(\cdot, X_t)$. Note that it holds $\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^*A)$ for a Hilbert-Schmidt operator A.

Lemma 3.6 *Under the assumptions (K1) and (K2) the following hold*

(i) $\operatorname{tr}\{(k_t \otimes k_t)(k_s \otimes k_s)\} = k^2(X_t, X_s)$*,*

(ii) $\|S\|_{\mathrm{HS}}^2 = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k^2(x,y)\mathrm{dP}^{X_0}(x)\mathrm{dP}^{X_0}(y)$*,*

(iii) $\mathrm{E}[\operatorname{tr}\{(k_0 \otimes k_0)S\}] = \|S\|_{\mathrm{HS}}^2$.

Proof: (i) Let $\{v_i\}_{i\in\mathbb{N}}$ denote an orthonormal base of $\mathcal{H}$. Then it holds due to the reproducing property (3.2)

$$\operatorname{tr}\{(k_t \otimes k_t)(k_s \otimes k_s)\} = \sum_{i=1}^{\infty} \langle v_i, k_t\rangle_{\mathcal{H}} \langle v_i, k_s\rangle_{\mathcal{H}} k(X_t, X_s) = \left\langle \sum_{i=1}^{\infty} \langle v_i, k_s\rangle_{\mathcal{H}} v_i, k_t \right\rangle_{\mathcal{H}} k(X_t, X_s).$$

(ii)

$$\begin{aligned} \|S\|_{\mathrm{HS}}^2 &= \sum_{i=1}^{\infty} \langle S v_i, S v_i\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \int\limits_{\mathbb{R}^d} \langle S v_i, k(\cdot, x)\rangle_{\mathcal{H}} \langle v_i, k(\cdot, x)\rangle_{\mathcal{H}} \mathrm{dP}^X(x) \\ &= \int\limits_{\mathbb{R}^d}\int\limits_{\mathbb{R}^d} \left\langle \sum_{i=1}^{\infty} \langle v_i, k(\cdot, x)\rangle_{\mathcal{H}} v_i, k(\cdot, y) \right\rangle_{\mathcal{H}} k(x,y)\mathrm{dP}^X(x)\mathrm{dP}^X(y). \end{aligned}$$

The assertion follows because $\mathrm{P}^X = \mathrm{P}^{X_0}$.

(iii)

$$\begin{aligned} \mathrm{E}[\operatorname{tr}\{(k_0 \otimes k_0)S\}] &= \mathrm{E}(\langle S k_0, k_0\rangle_{\mathcal{H}}) = \mathrm{E}\left(\int\limits_{\mathbb{R}^d} \langle k_0, k(\cdot, x)\rangle_{\mathcal{H}}^2 \mathrm{dP}^X(x) \right) \\ &= \int\limits_{\mathbb{R}^d}\int\limits_{\mathbb{R}^d} k^2(x,y)\mathrm{dP}^X(x)\mathrm{dP}^{X_0}(y) = \|S\|_{\mathrm{HS}}^2. \end{aligned}$$

□

Proof of the theorem: It holds due to $S_n = n^{-1}\sum_{t=1}^{n} k_t \otimes k_t$

$$\mathrm{E}\left(\|S_n - S\|_{\mathrm{HS}}^2\right) = \frac{1}{n^2}\sum_{t,s=1}^{n}\left(\mathrm{E}[\mathrm{tr}\{(k_t \otimes k_t)(k_s \otimes k_s)\}] - 2\,\mathrm{E}[\mathrm{tr}\{(k_0 \otimes k_0)S\}] + \|S\|_{\mathrm{HS}}^2\right).$$

For the first summand we get $\mathrm{E}[\mathrm{tr}\{(k_t \otimes k_t)(k_s \otimes k_s)\}] = \mathrm{E}\{k^2(X_t, X_s)\}$, due to Lemma 3.6 (i). Using the stationarity of $\{X_t\}_{t=1}^n$ and Lemma 3.6 (iii) we get

$$\mathrm{E}\left(\|S_n - S\|_{\mathrm{HS}}^2\right) = \frac{1}{n}\left\{\mathrm{E}\{k^2(X_0, X_0)\} - \|S\|_{\mathrm{HS}}^2\right\} + \frac{2}{n^2}\sum_{h=1}^{n-1}(n-h)\left[\mathrm{E}\{k^2(X_h, X_0)\} - \|S\|_{\mathrm{HS}}^2\right],$$

yielding the first result by an application of Lemma 3.6 (ii).

For the second equation we see due to the independence of $\{X_t\}_{t=1}^n$ and $\{\varepsilon_t\}_{t=1}^n$ that

$$\|T_n^* y - Sf\|_{\mathcal{H}}^2 = \sigma^2 n^{-1}\,\mathrm{E}\{k(X_0, X_0)\} + \mathrm{E}\left(\|S_n f - Sf\|_{\mathcal{H}}^2\right).$$

The rest follows along the same lines as the first part of the proof. □

3.7.2 Proof of Proposition 3.1

Recall that $Su = \mathrm{E}\,u(X_0)k(\cdot, X_0)$ for $u \in \mathcal{H}$. Define the independent random variables $Y_1, \ldots, Y_\mu$ that are all distributed as X_0.

First consider the following observation for $\mu \in \mathbb{N}$:

$$\begin{aligned} S^\mu u = S(S^{\mu-1}u) &= \mathrm{E}_{Y_1}(S^{\mu-1}u)(Y_1)k(\cdot, Y_1) = \mathrm{E}_{Y_2}\,\mathrm{E}_{Y_1}(S^{\mu-2}u)(Y_2)k(Y_1, Y_2)k(\cdot, Y_1) = \ldots \\ &= \mathrm{E}_{Y_\mu}\cdots\mathrm{E}_{Y_1}\prod_{\nu=1}^{\mu-1}k(Y_\nu, Y_{\nu+1})u(Y_\mu)k(\cdot, Y_1). \end{aligned} \tag{3.15}$$

(i) The inequality follows trivially: $\|f\|_{\mathcal{H}} = \|S^\mu u\| \leq R\|S\|_{\mathcal{L}}^\mu \leq R\kappa^\mu$.

(ii) We derive with (3.15) and the Cauchy-Schwarz inequality

$$\begin{aligned}\|S^\mu u\|_{\mathcal{H}} &\leq \kappa \, \mathrm{E}_{Y_\mu} \cdots \mathrm{E}_{Y_1} \prod_{\nu=1}^{\mu-1} k(Y_\nu, Y_{\nu+1}) \langle u, k(\cdot, Y_\mu) \rangle_{\mathcal{H}} \\ &\leq R \, \mathrm{E}_{Y_\mu} \cdots \mathrm{E}_{Y_1} \prod_{\nu=1}^{\mu-1} k(Y_\nu, Y_{\nu+1}).\end{aligned}$$

Note that for the Gaussian kernel we have $\kappa = 1$.

Define the matrix $\Gamma \in \mathbb{R}^{\mu \times \mu}$ via

$$\Gamma_{i,j} = \begin{cases} \sigma^{-2} + 2s & , \quad i = j = 2, \ldots, \mu - 1 \\ \sigma^{-2} + s & , \quad i = j = 1, \mu \\ -s & , \quad |i - j| = 1 \\ 0 & , \quad else. \end{cases}$$

We have

$$\begin{aligned}\mathrm{E}_{Y_\mu} \cdots \mathrm{E}_{Y_1} \prod_{\nu=1}^{\mu-1} k(Y_\nu, Y_{\nu+1}) &= \{2\pi\sigma^2\}^{-\mu/2} \int\limits_{\mathbb{R}^\mu} \exp\left(-1/2 x^{\mathrm{T}} \Gamma x\right) \mathrm{d}x \\ &= \{\sigma^{2\mu} \det(\Gamma)\}^{-1/2}.\end{aligned}$$

We denote with $\Gamma_{i:j}$, $i \leq j$ the $(j-i+1) \times (j-i+1)$-dimensional submatrix of Γ that contains only the columns and rows $i, i+1, \ldots, j-1, j$.

For the representation we consider the three term recursion that holds for determinants of tridiagonal matrices. Denote with $D_{i,\mu} = \sigma^{2i}\det(\Gamma_{1:i})$ for $i = 1, \dots, \mu$. Then we have

$$D_{i,\mu} = \begin{cases} 1 + s\sigma^2 & , \quad i = 1 \\ (1 + 2s\sigma^2)D_{i-1,\mu} - s^2\sigma^4 D_{i-2,\mu} & , \quad i = 2, \dots, \mu - 1 \\ (1 + s\sigma^2)D_{\mu-1,\mu} - s^2\sigma^4 D_{\mu-2,\mu} & , \quad i = \mu. \end{cases}$$

It is immediate that $D_{i,\mu}$ is a polynomial of degree i in $s\sigma^2$ with coefficients $\beta_{0,i,\mu}, \dots, \beta_{i,i,\mu}$. It is also straight forward that $\beta_{0,i,\mu} = 1$, $i = 1, \dots, \mu$. So all left to show is that these coefficients are always positive via induction.

For $D_{1,1}$ this is obviously true. We write $x = s\sigma^2$. Assume now that it is true for some $\mu \in \mathbb{N}$ and $i = 1, \dots, \mu$. As $D_{i,\mu+1} = D_{i,\mu}$ for $i < \mu$ we have

$$\begin{aligned} D_{\mu+1,\mu+1} &= (1 + 3x + x^2)D_{\mu-1,\mu} - (1 + x)x^2 D_{\mu-2,\mu} \\ &= (1 + x)D_{\mu-1,\mu} - x^2 D_{\mu-2,\mu} + x\{(2 + x)D_{\mu-1,\mu} - x^2 D_{\mu-2,\mu}\} \\ &= D_{\mu,\mu} + x(D_{\mu,\mu} + D_{\mu-1,\mu}). \end{aligned}$$

Thus we have the sum of polynomials with positive coefficients according to the induction hypothesis and the result is proven.

(iii) We take $u = \sum_{i=1}^{\infty} c_i k(\cdot, z_i)$ for $\{z_i\}_{i\in\mathbb{N}}, \{c_i\}_{i\in\mathbb{N}} \subset \mathbb{R}$ such that $\|u\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{\infty} c_i c_j k(z_i, z_j) \leq R^2$. The fact that a function $u \in \mathcal{H}$ can be represented as a linear combination of kernel functions is clear due to the Moore-Aronszajn Theorem, see Berlinet and Thomas-Agnan (2004).

Define the matrix $\Gamma = [\Gamma_{i,j}]_{i,j=1}^{\mu+2} \in \mathbb{R}^{(\mu+2)\times(\mu+2)}$ via

$$\Gamma_{i,j} = \begin{cases} \sigma^{-2} + 2s & , \quad i = j = 2, \ldots, \mu+1 \\ s & , \quad i = j = 1, \mu+2 \\ -s & , \quad |i-j| = 1 \\ 0 & , \quad else \end{cases} .$$

Then we have via the integration of Gaussian functions and (3.15)

$$\begin{aligned} f(x) &= \frac{1}{(2\pi\sigma^2)^{\mu/2}} \sum_{i=1}^{\infty} c_i \int\limits_{\mathbb{R}^\mu} \exp\left\{-1/2(x, x_1, \ldots, x_\mu, z_i)\Gamma(x, x_1, \ldots, x_\mu, z_i)^{\mathrm{T}}\right\} \mathrm{d}(x_1, \ldots, x_\mu) \\ &= \frac{1}{\sigma^\mu \det(\Gamma_{2:\mu+1})^{1/2}} \sum_{i=1}^{\infty} c_i \exp\left[-1/2 \det(\Gamma_{2:\mu+1})^{-1} \left\{\det(\Lambda_{1:\mu+1})(x^2 + z_i^2) - 2s^{\mu+1} x z_i\right\}\right] . \end{aligned}$$

Here we used the symmetry property $\det(\Gamma_{2:\mu+2}) = \det(\Gamma_{1:\mu+1})$ as the first and last rows and columns of Γ are identical. This concludes the proof. □

3.7.3 Proof of Theorem 3.3

Denote with g_h the common density of $(X_h, X_0)^{\mathrm{T}}$ and g_0 the density of X_0. We need some intermediate results to prove the theorem.

Lemma 3.7 *Assume that condition (D2) holds. Then we have*

$$n^{-2} \sum_{h=1}^{n-1} (n-h)|\rho_h| \leq c \begin{cases} n^{-1}\zeta(q) & , \quad q > 1 \\ n^{-1}\log(n)\{5 - \log(4)\} & , \quad q = 1 \\ n^{-q}\left[\{2(1-q)^{-1} - (2-q)^{-1}\} + (2-q)^{-1}2^{2-q}\right] & , \quad q \in (0,1). \end{cases} \tag{3.16}$$

Here ζ denotes the Riemann zeta function.

Proof: Recall that by condition (D2) we have $|\rho(h)| \leq (h+1)^{-q}$, $h = 0, \ldots, n-1$ for some $q > 0$.

First assume $q \in (0, 1]$. The integral test for series convergence gives lower and upper bounds for the hyperharmonic series as

$$(1-q)^{-1}\{(n+1)^{1-q} - 2^{1-q}\} \leq \sum_{h=2}^{n} h^{-q} \leq 2^{-q} + (1-q)^{-1}\{n^{1-q} - 2^{1-q}\}.$$

This yields

$$n^{-2}\sum_{h=1}^{n-1}(n-h)(h+1)^{-q} = n^{-2}\sum_{h=2}^{n}(n+1-h)h^{-q} = n^{-2}\left\{(n+1)\sum_{h=2}^{n} h^{-q} - \sum_{h=2}^{n} h^{-(q-1)}\right\}$$
$$\leq n^{-2}\left[(n+1)\left\{2^{-q} + (1-q)^{-1}(n^{1-q} - 2^{1-q}\right\} - (2-q)^{-1}\left\{(n+1)^{2-q} - 2^{2-q}\right\}\right]. \quad (3.17)$$

We need to separate two cases. First let $q \in (0, 1)$, then it holds from (3.17) and the fact that $n^{-2} \leq n^{-1} \leq n^{-q}$

$$\begin{aligned}
&n^{-2}\sum_{h=1}^{n-1}(n-h)(h+1)^{-q}\\
\leq&\frac{n+1}{n^2}\left\{\frac{2^{-q}(1-q) - 2^{1-q}}{1-q}\right\} + \frac{n+1}{n^{1+q}}(1-q)^{-1} - \frac{(n+1)^{2-q}}{n^2}(2-q)^{-1} + \frac{1}{n^2}\frac{2^{2-q}}{2-q}\\
\leq& n^{-q}[\{2(1-q)^{-1} - (2-q)^{-1}\} + (2-q)^{-1}2^{2-q}],
\end{aligned}$$

due to $2^{-q}(1-q) - 2^{1-q} < 0$.

For $q = 1$ we evaluate the limit

$$\begin{aligned}
&\lim_{q\to 1\pm} n^{-2}\left[(n+1)\left\{2^{-q} + (1-q)^{-1}(n^{1-q} - 2^{1-q}\right\} - (2-q)^{-1}\left\{(n+1)^{2-q} - 2^{2-q}\right\}\right] \\
&= (2n^2)^{-1}[3 - \log(4) - n\{1 + \log(4)\}] + n^{-2}(n+1)\log(n) \\
&\leq \frac{\log(n)}{n}\left[5 - \log(4)\right].
\end{aligned}$$

Finally, the case $q > 1$ is trivial as the zeta-function $\zeta(q)$ is defined as the hyperharmonic series with coefficient q. □

The next lemma and the subsequent corollary show that the quantities appearing in the sums of Theorem 3.2 can be linked to the autocorrelation function ρ:

Lemma 3.8 *Under the assumptions (K1), (K2) and (D1) it holds for $h > 0$ with $\rho_h = \tau_0^{-1}\tau_h$*

$$\int\limits_{\mathbb{R}^{2d}} k^2(x,y)\{g_h(x,y) - g_0(x)g_0(y)\}\mathrm{d}(x,y) \leq \frac{\kappa^2}{\{(4\pi\tau_0)^d\det(\Sigma)\}^{1/2}}\theta^{1/2}(\rho_h),$$

$$\int\limits_{\mathbb{R}^{2d}} k(x,y)f(x)f(y)\{g_h(x,y) - g_0(x)g_0(y)\}\mathrm{d}(x,y) \leq \frac{\kappa M}{\{(4\pi\tau_0)^d\det(\Sigma)\}^{1/2}}\theta^{1/2}(\rho_h),$$

with $\theta(\rho) = 1 + (1-\rho^2)^{-d/2} - 2^{d+1}(4-\rho^2)^{-d/2}$, $\rho \in [0,1)$.

Proof: We will only proof the first inequality, the second one follows in the same way.

By Jensen's inequality and (K2) we know

$$\begin{aligned}
&\int\limits_{\mathbb{R}^{2d}} k^2(x,y)\{g_h(x,y) - g_0(x)g_0(y)\}\mathrm{d}(x,y) \\
&\leq \kappa^2\left[\int\limits_{\mathbb{R}^{2d}}\left\{g_h^2(x,y) - 2g_h(x,y)g_0(x)g_0(y) + g_0^2(x)g_0^2(y)\right\}\mathrm{d}(x,y)\right]^{1/2}.
\end{aligned}$$

The first and third integral term can readily be calculated as

$$\int\limits_{\mathbb{R}^{2d}} g_h^2(x,y)\mathrm{d}(x,y) = [(4\pi)^d(\tau_0^2-\tau_h^2)^{d/2}\det(\Sigma)]^{-1}$$

$$\left\{\int\limits_{\mathbb{R}^d} g_0^2(x)\mathrm{d}x\right\}^2 = \{(4\pi)^d\tau_0^d\det(\Sigma)\}^{-1}.$$

For the first equality we use $\det(A\otimes\Sigma)=\det(A)^d\det(\Sigma)^2$ for $A\in\mathbb{R}^{2\times 2}$ and thus

$$\int\limits_{\mathbb{R}^{2d}} g_h(x,y)g_0(x)g_0(y)\mathrm{d}(x,y) = \frac{\int_{\mathbb{R}^{2d}}\exp\left(-1/2z^{\mathrm{T}}G^{-1}z\right)\mathrm{d}z}{(2\pi)^{2d}\det(\Sigma)^2\tau_0^d(\tau_0^2-\tau_h^2)^{d/2}}, \tag{3.18}$$

with

$$G^{-1}=\left\{\begin{pmatrix}\tau_0 & \tau_h\\ \tau_h & \tau_0\end{pmatrix}^{-1}+\begin{pmatrix}\tau_0^{-1} & 0\\ 0 & \tau_0^{-1}\end{pmatrix}\right\}\otimes\Sigma^{-1}.$$

It holds $\det(G)=(4\tau_0^2-\tau_h^2)^{-d}(\tau_0^4-\tau_0^2\tau_h^2)^d\det(\Sigma)^2$. Thus we get with (3.18)

$$\begin{aligned}\int\limits_{\mathbb{R}^{2d}} g_h(x,y)g_0(x)g_0(y)\mathrm{d}(x,y) &= \frac{(2\pi)^d\tau_0^d(\tau_0^2-\tau_h^2)^{d/2}\det(\Sigma)}{(2\pi)^{2d}\det(\Sigma)^2(4\tau_0^2-\tau_h^2)^{d/2}\tau_0^d(\tau_0^2-\tau_h^2)^{d/2}}\\ &= \left\{(2\pi)^d(4\tau_0^2-\tau_h^2)^{d/2}\det(\Sigma)\right\}^{-1},\end{aligned}$$

completing the proof by multiplying all integrals with $\tau_0^{-d}\tau_0^d$. □

Corollary 3.2 *Under the assumptions (K1), (K2), (D1) and (D2) it holds for all* $h>0$ *and* $q>0$

$$\int\limits_{\mathbb{R}^{2d}} k^2(x,y)\{g_h(x,y)-g_0(x)g_0(y)\}\mathrm{d}(x,y) \le \frac{\kappa^2 d^{1/2}}{\{(2\pi)^d\det(\Sigma)\}^{1/2}}(1-4^{-q})^{-1/4(d-2)}|\rho_h|$$

$$\int\limits_{\mathbb{R}^{2d}} k(x,y)f(x)f(y)\{g_h(x,y)-g_0(x)g_0(y)\}\mathrm{d}(x,y) \le \frac{\kappa M d^{1/2}}{\{(2\pi)^d\det(\Sigma)\}^{1/2}}(1-4^{-q})^{-1/4(d-2)}|\rho_h|.$$

Proof: Recall that $\theta(\rho) = 1 + \{1-\rho^2\}^{-d/2} - 2^{d+1}\{4-\rho^2\}^{-d/2}$ for $\rho \in [0,1)$. We seek to find bounds on θ and the corollary can be proven by an application of Lemma 3.8.

By assumption (D2) we know there is a ρ_* such that $\rho_h^2 \leq \rho_*^2 < 1$ for all $h > 0$. Thus consider $\rho \in [0, \rho_*]$. We start by finding a constant $C > 0$ with

$$\theta'(\rho) = \rho\left\{(1-\rho^2)^{-d/2-1} - 2^{d+1}(4-\rho^2)^{-d/2-1}\right\} d \leq \rho 2C.$$

Thus C can be taken as $C = d\left\{(1-\rho_*^2)^{-d/2-1} - 2^{d+1}(4-\rho_*^2)^{-d/2-1}\right\}$.

Thus we know that the slope of θ is always less than that of $C\rho^2$. Finally it holds that $\theta(0) = 0$ and thus $0 \leq \theta(\rho) \leq C\rho^2$, $\rho \in [0, \rho_*]$.

Under condition (D2) it holds $\{1-\rho_*^2\}^{-d/2} \leq \{1-2^{-2q}\}^{-d/2}$, completing the proof by using Lemma 3.8. □

Proof of the theorem: First note that the the operator norm is dominated by the Hilbert-Schmidt norm. By Markov's inequality we have for $\nu \in (0,1]$

$$\mathrm{P}\left(\|S_n - S\|_{\mathrm{HS}}^2 \leq \nu^{-1}\,\mathrm{E}\,\|S_n - S\|_{\mathrm{HS}}^2\right) \geq 1-\nu,$$

$$\mathrm{P}\left(\|T_n^* y - Sf\|_{\mathcal{H}}^2 \leq \nu^{-1}\,\mathrm{E}\,\|T_n^* y - Sf\|_{\mathcal{H}}^2\right) \geq 1-\nu.$$

An application of Theorem 3.2, Corollary 3.2 and Lemma 3.7 completes the proof. □

Bibliography

Bauer, F., Pereverzev, S., and Rosasco, L. (2007). On regularization algorithms in learning theory. *J. Complexity*, 23:52–72.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Boston.

Blanchard, G. and Krämer, N. (2010a). Kernel partial least squares is universally consistent. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings*, volume 9, pages 57–64. JMLR.

Blanchard, G. and Krämer, N. (2010b). Optimal learning rates for kernel conjugate gradient regression. *Adv. Neural Inf. Process. Syst.*, 23:226–234.

Caponnetto, A. and de Vito, E. (2007). Optimal rates for regularized least-squares algorithm. *Found. Comp. Math.*, 7:331–368.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39:1–49.

De Vito, E., Caponnetto, A., and Rosasco, L. (2006). Discretization error analysis for Tikhonov regularization in learning theory. *Anal. Appl.*, 4:81–99.

de Vito, E., Rosasco, L., Caponnetto, A., de Giovanni, U., and Odone, F. (2005). Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904.

Frank, I. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Giraitis, L., Hira, L., and Surgailis, D. (2012). *Large Sample Inference for Long Memory Processes*. Imperial College Press, London, 1 edition.

Hanke, M. (1995). *Conjugate Gradient Type Methods for Ill-posed Problems*. Wiley, New York, 1 edition.

Helland, I. S. (1988). On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.*, 17(2):581–607.

Krämer, N. and Braun, M. L. (2007). Kernelizing PLS, degrees of freedom, and efficient model selection. In *Proceedings of the 24th International Conference on Machine Learning*, pages 441–448. ACM.

Lindgren, F., Geladi, P., and Wold, S. (1993). The kernel algorithm for PLS. *J. Chemometrics*, 7:45–59.

Phatak, A. and de Hoog, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *J. Chemometr.*, 16:361–367.

Rosipal, R. (2003). Kernel partial least squares for nonlinear regression and discrimination. *Neural Netw. World*, 13:291–300.

Rosipal, R., Girolami, M., and Trejo, L. (2000). Kernel PCA for feature extraction of event-related potentials for human signal detection performance. In *Proceedings of ANNIMAB-1 Conference*, pages 321–326. Springer.

Rosipal, R. and Trejo, L. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.*, 2:97–123.

Rosipall, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. *Lecture Notes in Comput. Sci.*, 3940:34–51.

Samorodnitsky, G. (2007). *Long Range Dependence*. now Publisher, Hanover, 1 edition.

Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann Publishers.

Schölkopf, B., Herbrich, R., and Smola, A. (2001). A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer.

Steinwart, I., Hush, D., and Scovel, C. (2005). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. Technical report, IEEE Trans. Inform. Theory.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. In *Advances in Kernel Methods - Support Vector Learning*, pages 69–88. MIT Press.

Wold, S., Ruhe, A., Wold, H., and Dunn, I. W. (1984). The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Comput.*, 5:735–743.

Name: Marco Singer

Address: Eisenbahnstraße 15, 37073 Göttingen.

Email: msinger@gwdg.de

Date of birth: 24.07.1986, Braunschweig, Germany.

Marital status: single

Curriculum vitae

10.2006–09.2009	Study of Mathematics (Bachelor), Leibniz-Universität Hannover
09.2009	Bachelor degree
	Thesis name: SQP-Methods for equality constrained nonlinear optimization problems
10.2009–11.2012	Study of Mathematics (Master), Leibniz-Universität Hannover
11.2012	Master degree
	Thesis name: Goodness of fit tests for nonlinear time series models
01.2013–	PhD studies in Mathematics in the GAUSS program, Georg-August-Universität Göttingen
	Thesis name: Partial least squares for serially dependent data.