Biometrika (2016), **103**, 2, *pp.* 351–362
Printed in Great Britain

doi: 10.1093/biomet/asw010
Advance Access publication 29 April 2016

Partial least squares for dependent data

By MARCO SINGER, TATYANA KRIVOBOKOVA, AXEL MUNK

Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany

msinger@gwdg.de tkrivob@uni-goettingen.de munk@math.uni-goettingen.de

AND BERT DE GROOT

Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

bgroot@gwdg.de

Summary

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 yields consistent estimation. A protein dynamics example illustrates the superior predictive power of the proposed method.

Some key words: Dependent data; Latent variable model; Nonstationary process; Partial least squares; Protein dynamics.

1. Introduction

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

$$y = X\beta + \varepsilon, \tag{1}$$

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

$$\begin{aligned} \hat{w}_{i+1} &= b - A\hat{\beta}_i, \quad \hat{\beta}_0 = 0, \\ \hat{\beta}_i &= \hat{W}_i(\hat{W}_i^{\mathrm{T}} A \hat{W}_i)^{-1}\hat{W}_i^{\mathrm{T}} b \end{aligned} \tag{2}$$

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

Alternatively, $\hat{\beta}_i$ can be defined using the fact that $\hat{w}_i \in \mathcal{K}_i(A, b)$, where $\mathcal{K}_i(A, b)$ is a Krylov space, i.e., a space spanned by $\{A^{j-1}b\}_{j=1}^{i}$ (Helland, 1988). Then, one can define the partial least squares estimators as $\hat{\beta}_i = \arg\min_{\beta \in \mathcal{K}_i(A,b)} (y - X\beta)^{\mathrm{T}}(y - X\beta)$. 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 & Friedman (1993) and Farkas & Héberger (2005) found the partial least squares algorithm to be competitive with regularized regression techniques, such as principal component regression, the 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 component regression, as found in Almøy (1996).

Partial least squares regression has a long and successful history in various areas of application; see, for example, Hulland (1999), Lobaugh et al. (2001) and Nguyen & Rocke (2002). However, the statistical properties of the algorithm have received little attention, perhaps because of the nonlinearity of partial least squares estimators in the response variable. Some attempts to understand the properties of partial least squares estimators can be found in Höskuldsson (1988), Phatak & de Hoog (2002) and Krämer (2007). The almost sure convergence of the method was established by Naik & Tsai (2000). For kernel partial least squares, Blanchard & Krämer (2010a) obtained results on convergence in probability by early stopping. For the closely linked kernel conjugate gradient algorithm, Blanchard & Krämer (2010b) established order-optimal convergence rates dependent on the regularity of the target function. Delaigle & 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 condition 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 let $\|\cdot\|_{\mathcal{L}}$ denote the spectral norm and $\|\cdot\|$ the Frobenius norm for matrices; $\|\cdot\|$ also denotes the Euclidean norm for vectors.

2. Partial least squares under dependence

2·1. *Latent variable model*

In many applications the standard linear model (1) is too restrictive. For example, if a covariate that is relevant to the response cannot be observed or measured directly, so-called latent variable or structural equation models are used (Skrondal & Rabe-Hesketh, 2006), where it is assumed that X and y are linked by $l \leqslant k$ latent vectors and the remaining vectors in the k-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 useful in the 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:

$$X = V(NP^{\mathrm{T}} + \eta_1 F), \qquad y = V(Nq + \eta_2 f), \tag{3}$$

where N and F are $n \times l$ and $n \times k$ random matrices, respectively, and f is an n-dimensional random vector. The random elements N, F and f can have different distributions, but they are independent of each other, with all entries being independent and identically distributed with zero expectation and unit variance. The matrix $P \in \mathbb{R}^{k\times l}$ and vector $q \in \mathbb{R}^{l}$ are deterministic and unknown, along with the real-valued parameters $\eta_1, \eta_2 \geqslant 0$. We assume that $n \geqslant k \geqslant l$ and that $\operatorname{rank}(N) = \operatorname{rank}(P) = l$ and $\operatorname{rank}(F) = k$ 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 (3) can be viewed as the matrix form of a k-dimensional time series $\{X_t\}_{t=1}^{n}$ ($X_t \in \mathbb{R}^k$), and y can be viewed as a real-valued time series $\{y_t\}_{t=1}^{n}$. The covariance matrix V^2 determines the dependence between observations, which could be nonstationary. We call V^2 the temporal covariance matrix of X and define $\Sigma^2 = PP^{\mathrm{T}} + \eta_1^2 I_k$. Setting $l = k$ and $\eta_1 = 0$ reduces model (3) to the standard linear regression model 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 (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 with 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 § 1 with matrix X and vector y defined as in (3). If $\eta_1 = 0$, then model (1) is correctly specified with $q = P^{\mathrm{T}}\beta$, and the partial least squares estimator (2) estimates β. If $\eta_1 > 0$, then model (1) is misspecified and $\beta(\eta_1) = \Sigma^{-2}Pq$ is estimated instead. 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. *Population and sample partial least squares*

The population partial least squares algorithm for independent observations was first introduced by Helland (1990). Under model (3), we modify the definition of the population partial least squares basis vectors to

$$w_i = \underset{\substack{w\in\mathbb{R}^k\\ \|w\|=1}}{\arg\max}\ \frac{1}{n^2}\sum_{t,s=1}^{n} \operatorname{cov}(y_t - X_t^{\mathrm{T}}\beta_{i-1}, X_s^{\mathrm{T}}w), \qquad \beta_0 = 0,$$

where $\beta_i \in \mathbb{R}^k$ 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 the Supplementary Material. In particular, under model (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

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

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

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

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

To obtain the sample partial least squares estimators $\hat{\beta}_i$, we replace Σ^2 and Pq 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 L_2-inconsistent estimation, as the following theorem shows.

Theorem 1. *Suppose that model* (3) *holds and that the fourth moments of* $N_{1,1}$ *and* $F_{1,1}$ *exist. Define* $A = \|V\|^{-2}X^{\mathrm{T}}X$ *and* $b = \|V\|^{-2}X^{\mathrm{T}}y$*. Then*

$$E\left(\|\Sigma^2 - A\|^2\right) = \frac{\|V^2\|^2}{\|V\|^4}\left(C_A + \sum_{t=1}^{n} \frac{\|V_t\|^4}{\|V^2\|^2} c_A\right),$$

$$E\left(\|Pq - b\|^2\right) = \frac{\|V^2\|^2}{\|V\|^4}\left(C_b + \sum_{t=1}^{n} \frac{\|V_t\|^4}{\|V^2\|^2} c_b\right),$$

where

$$C_A = \|P\|^4 + \|P^{\mathrm{T}}P\|^2 + 4\eta_1^2\|P\|^2 + \eta_1^4 k(1+k),$$

$$c_A = \{E(N_{1,1}^4) - 3\}\sum_{i=1}^{l}\|P_i\|^4 + \{E(F_{1,1}^4) - 3\}\eta_1^4 k,$$

$$C_b = \|Pq\|^2 + \|P\|^2\|q\|^2 + \eta_1^2 k\|q\|^2 + \eta_1^2\eta_2^2 k + \eta_2^2\|P\|^2,$$

$$c_b = \{E(N_{1,1}^4) - 3\}\sum_{i=1}^{l}\|P_i\|^2 q_i^2$$

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), so 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 $E(N_{1,1}^4) = E(F_{1,1}^4) = 3$, then the constants c_A and c_b vanish, simplifying the expressions for the mean squared errors of A and b. This condition is satisfied, for example, by the standard normal distribution. Thus these terms can be interpreted as a penalization for nonnormality.

Finally, $\sum_{t=1}^{n}\|V_t\|^4 \leqslant \sum_{t,s=1}^{n}(V_t^{\mathrm{T}}V_s)^2 = \|V^2\|^2$ 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, then ignoring the temporal dependence V^2 may lead to inconsistent estimation.

3. Properties of partial least squares estimators under dependence

3·1. *Concentration inequality for partial least squares estimators*

In this subsection we apply the techniques of Blanchard & Krämer (2010b), who derived convergence rates for the kernel conjugate gradient algorithm, which is closely related to kernel partial least squares. Both algorithms approximate the solution on Krylov subspaces, but they employ different norms. In particular, Blanchard & Krämer (2010b) showed that if the conjugate gradient algorithm is stopped early, then 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 and identically distributed observations.

We avoid the nonparametric setting of Blanchard & Krämer (2010b) and study a standard linear partial least squares algorithm with a fixed dimension k 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.

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 $0 < a_0 \leqslant a$ such that

$$\left\| A^{1/2}\hat{\beta}_{a_0} - A^{-1/2}b \right\| \leqslant \tau(\delta\|\hat{\beta}_{a_0}\| + \epsilon), \tag{4}$$

for $\delta, \epsilon > 0$ defined in Theorem 2 and some $\tau \geqslant 1$. Here a denotes the maximal dimension of the sample Krylov space $\mathcal{K}_i(A, b)$ and almost surely equals $l + (k - l)\mathbb{I}(\eta_1 > 0)$ where $\mathbb{I}(\cdot)$ denotes the indicator function. For technical reasons we stop at $a^* = a_0 - 1$ if $p_{a_0}(0) \geqslant \zeta\delta^{-1}$, where p_i is a polynomial of degree $i - 1$ with $p_i(A)b = \hat{\beta}_i$ and $\zeta < \tau^{-1}$. The existence of such polynomials was proved by Phatak & de Hoog (2002). If (4) never holds, we take $a^* = a$. With this stopping index we get the following concentration inequality.

Theorem 2. *Assume that model* (3) *with* $\eta_1 > 0$ *holds and that the fourth moments of* $N_{1,1}$ *and* $F_{1,1}$ *exist. Furthermore, let* a^* *satisfy* (4) *with* $\tau \geqslant 1$ *and* $\zeta < \tau^{-1}$. *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*

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

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

$$\left\|\hat{\beta}_{a^*} - \beta(\eta_1)\right\| \leqslant \frac{\|V^2\|}{\|V\|^2}\left\{c_1(\nu) + \frac{\|V^2\|}{\|V\|^2}c_2(\nu)\right\}, \tag{5}$$

where

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

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

If $N_{1,1}, F_{1,1}, f_1 \sim N(0, 1)$, 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 states that the convergence rate of the optimally stopped partial least squares estimator $\hat{\beta}_{a^*}$ to the true parameter $\beta(\eta_1)$ is driven by the ratio of the Frobenius norms of V^2 and V, similar to the assertions of Theorem 1. In particular, if the data are independent with $V = I_n$, then $\hat{\beta}_{a^*}$ is square-root consistent. In this case $c_2(\nu)$ is asymptotically negligible. The theorem excludes the case where $\|V\|^{-2}\|V^2\|$ does not converge to zero.

3·2. *Properties of $\hat{\beta}_1$ under dependence*

Nonlinearity in the response variable of $\hat{\beta}_i$ hinders its standard statistical analysis, as no closed-form expression for the mean squared error of $\hat{\beta}_i$ is available and concentration inequalities similar to (5) are, to the best of our knowledge, the only existing results on convergence rates of partial least squares estimators. However, if the ratio of $\|V^2\|$ and $\|V\|^2$ does not converge to zero, Theorem 2 does not hold.

In this subsection we focus on the first partial least squares estimator $\hat{\beta}_1$, for several reasons. First, an explicit expression for its mean squared error can be derived. Second, if there is only one latent component that links X and y, i.e., $l = 1$ in (3), then consistent estimation of β_1 is crucial. Third, $\hat{\beta}_1$ is collinear to the direction of the maximal covariance between X and y given by $\hat{w}_1$, 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 $\hat{\beta}_1$ is an inconsistent estimator of β_1.

THEOREM 3. *Assume that model* (3) *holds, $k > 1$, and the eighth moments of $N_{1,1}$, $F_{1,1}$ and f_1 exist. Furthermore, suppose that the ratio $\|V\|^{-2}\|V^2\|$ does not converge to zero as $n \to \infty$. Then, for either $l > 1$ and $\eta_1 \geqslant 0$ or $l = 1$ and $\eta_1 > 0$, $\hat{\beta}_1$ is an inconsistent estimator for β_1.*

The case of $l = 1$ and $\eta_1 = 0$, not included in Theorem 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.

3·3. *Examples of dependence structures*

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

$$|\gamma(t)| \leqslant \gamma(0)c\exp(-\rho t) \quad (t \in \mathbb{N}), \tag{6}$$

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

In what follows, $f(n) \sim g(n)$ means $c_1 \leqslant f(n)/g(n) \leqslant c_2$ for n large, $0 < c_1 < c_2$ and $f, g : \mathbb{N} \to \mathbb{R}$.

THEOREM 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* (6). *Then $\|V^2\| \sim n^{1/2}$ and $\|V\|^2 \sim n$.*

Hence, if V^2 in model (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 could affect the constants. Examples of processes with exponentially decaying autocovariances are stationary autoregressive moving average processes.

As an example of a nonstationary process, we consider first-order integrated processes. If $\{X_t\}_{t\in\mathbb{Z}}$ is stationary with autocovariance function γ satisfying (6), then $\sum_{i=1}^{t} X_i$ is an integrated process of order 1.

THEOREM 5. *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary process with autocovariance function γ satisfying* (6)*. If $\gamma(t) < 0$ for some t, we assume additionally that $\rho > \log(2c+1)$. 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 can become negative, so that V^2 is a valid covariance matrix.

Theorem 5 implies that the ratio $\|V\|^{-2}\|V^2\|$ does not converge to zero for certain integrated processes. In particular, combining this result with Theorems 1 and 3 shows that the elements of the sample Krylov space, A and b, as well as $\hat{\beta}_1$, are inconsistent if the dependence structure of the data can be described by an integrated process satisfying the conditions of Theorem 5, such as an integrated autoregressive moving average process of order $(1, 1, 1)$.

4. PRACTICAL ISSUES

4·1. *Corrected partial least squares estimator*

So far we have been considering the standard partial least squares algorithm, showing that if certain dependences in the data are ignored, then estimation will be 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 $k_i(S) = A(S)^{i-1}b(S)$, $K_i(S) = [k_1(S), \ldots, k_i(S)] \in \mathbb{R}^{k\times i}$ and $\hat{\beta}_i(S) = K_i(S)\{K_i(S)^{\mathrm{T}}A(S)K_i(S)\}^{-1}K_i(S)^{\mathrm{T}}b(S)$ $(i = 1, \ldots, k)$.

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 mean squared error independent of V, according to Theorem 1. Hence, the resulting $\hat{\beta}_i(V)$ is also a consistent estimator of β_i, and Theorem 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 usually unknown and is replaced by a consistent estimator $\hat{V}^2$. We call the estimator $\hat{\beta}_i(\hat{V})$ 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 6. *Let $\hat{V}^2$ be an estimator for V^2 that is invertible for all $n \in \mathbb{N}$ and satisfies $\|V\hat{V}^{-2}V - I_n\|_{\mathcal{L}} = O_{\mathrm{p}}(r_n)$, where r_n is some sequence of positive numbers such that $r_n \to 0$ as $n \to \infty$. Then*

$$\|A(\hat{V}) - \Sigma^2\|_{\mathcal{L}} = O_{\mathrm{p}}(r_n), \quad \|b(\hat{V}) - Pq\| = O_{\mathrm{p}}(r_n).$$

Moreover, with probability at least $1 - \nu$ (where $0 < \nu \leqslant 1$),

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

where the definition of a^* in (4) *is updated by replacing A, b and $\hat{\beta}_i$ with $A(\hat{V})$, $b(\hat{V})$ and $\hat{\beta}_i(\hat{V})$, respectively.*

Theorem 6 states that if a consistent estimator of the covariance matrix V^2 is available, then the elements of the population Krylov space, A and b, as well as the coefficient $\beta(\eta_1)$, can be consistently estimated by $A(\hat{V})$, $b(\hat{V})$ and $\hat{\beta}_{a^*}(\hat{V})$, respectively. The convergence rate of these estimators is no faster than that of $\hat{V}^2$. For example, if the temporal dependence in the data follows some parametric model, then parametric rates of $n^{-1/2}$ are also achieved for $A(\hat{V})$, $b(\hat{V})$ and $\hat{\beta}_{a^*}(\hat{V})$. Estimation of V^2 by some nonparametric methods, e.g., with a banding or tapering approach, leads to slower convergence rates; see Bickel & Levina (2008) or Wu & Xiao (2012). Similar results are well known in the context of linear regression. For example, Theorem 5.7.1 of 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.

4·2. *Estimation of covariance matrices*

To obtain the corrected partial least squares estimator, some consistent estimator of V^2 based on a single realization of the process is needed. In model (3), the dependence structure over the observations of X is the same as that of y and so 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 & Davis (1991, § 4.4). Parameters of autoregressive moving average processes can be estimated by either Yule–Walker or maximum likelihood estimators, both of which attain parametric rates. Another approach is to band or taper the empirical autocovariance function of y (Bickel & Levina, 2008; Wu & Pourahmadi, 2009; Wu & Xiao, 2012). These nonparametric estimators are very flexible, but are computationally intensive and have slower convergence rates.

If y is an integrated process of order 1, then V^2 can easily be derived from the covariance matrix estimator of the corresponding stationary process.

5. Simulations

To evaluate the 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 (3) we set $k = 20$ and $l = 1, 5$ and take the elements of P to be independent and 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 obtain a signal-to-noise ratio of 2 in y, 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 a first-order autoregressive process 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. For this, the covariance matrix of the autoregressive moving average process was estimated parametrically, as discussed in § 4·2. A nonparametric estimation of this covariance matrix led to qualitatively similar results.

The boxplots in Fig. 1 show the squared distance between $\hat{\beta}_i$ and $\beta(\eta_1)$ in 500 Monte Carlo replications. Two cases are displayed in each panel: one where the model has just one latent

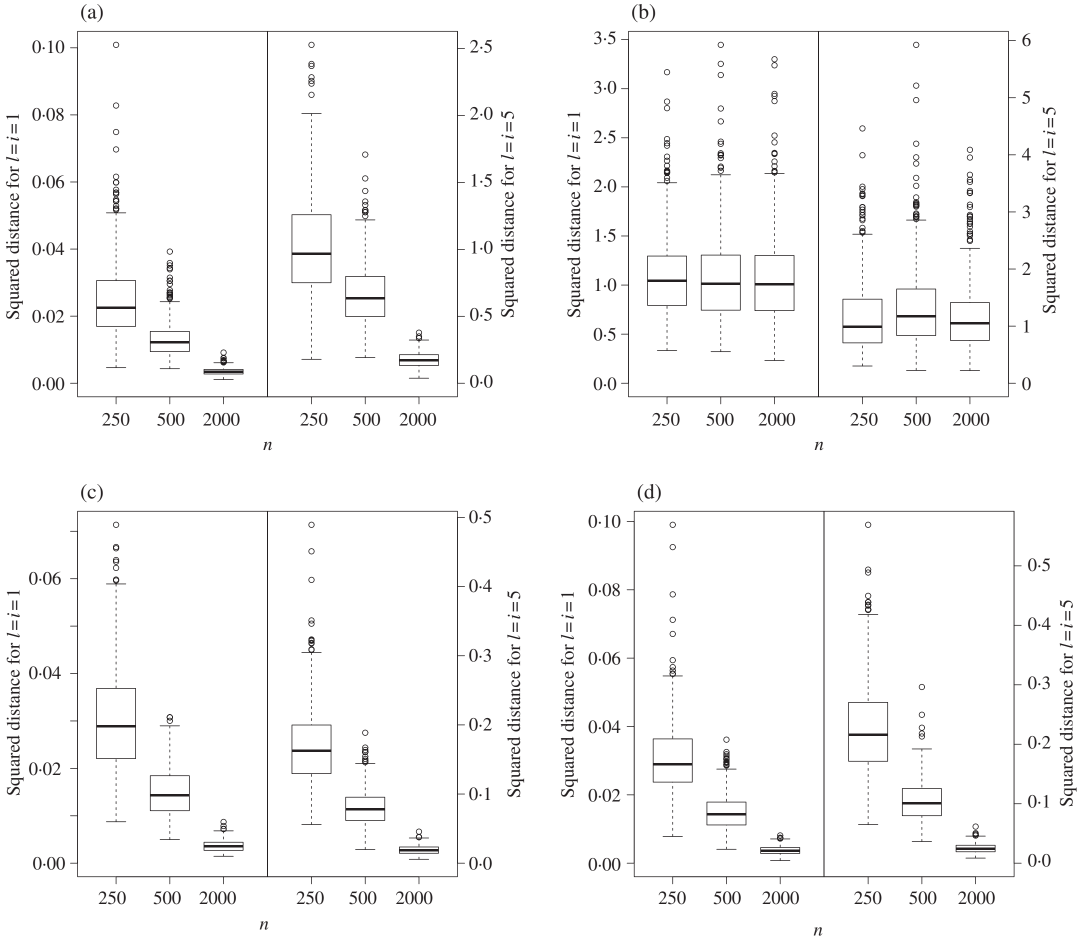

Fig. 1. Boxplots of the squared distance of the partial least squares estimators $\hat{\beta}_i$ to $\beta(\eta_1)$ in 500 Monte Carlo samples. In each panel, the three boxplots on the left correspond to $l = i = 1$, and the three on the right correspond to $l = i = 5$. The dependence structures are: (a) first-order autoregressive; (b), (d) autoregressive integrated moving average of order (1, 1, 1); (c) independent and identically distributed. Standard partial least squares was employed in (a) and (c) and corrected partial least squares in (d).

component and $\hat{\beta}_1$ is considered, i.e., $l = i = 1$, and another where the model has five latent components and the squared distance of $\hat{\beta}_5$ to $\beta(\eta_1)$ is considered, i.e., $l = i = 5$.

We observe that the mean squared error of $\hat{\beta}_i$ obtained from the standard partial least squares method converges to zero with increasing sample size for autoregressive and independent data. However, an autoregressive dependence in the data leads to a somewhat higher mean squared error; compare panels (a) and (d) in Fig. 1. If the data follow an autoregressive integrated moving average process and this structure is ignored in the partial least squares algorithm, then the mean squared error of $\hat{\beta}_i$ converges to some positive constant; see Fig. 1(b). By taking into account these nonstationary dependencies, the corrected partial least squares algorithm yields consistent estimation, similar to the independent data case; compare panels (c) and (d) in Fig. 1.

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

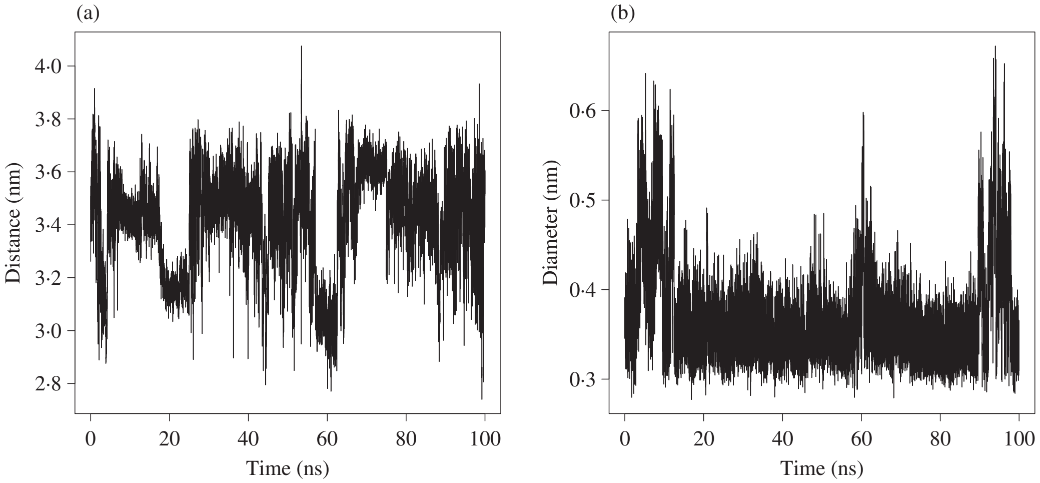

Fig. 2. (a) Distance between the first backbone atom and the first centre of mass of aquaporine; (b) the opening diameter over time.

6. Application to protein dynamics

Proteins perform their biological functions through particular movements (see, e.g., Henzler-Wildman & Kern, 2007), so a key step in understanding protein function is to gain 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-level detail on time-scales of nanoseconds to microseconds. Although in principle such studies allow us to directly examine function-dynamics relationships, the analysis is frequently hampered by the large dimensionality of the protein configuration space, which makes it nontrivial to identify collective modes of motion that are directly related to a functional property of interest.

Krivobokova et al. (2012) have shown that partial least squares helps to identify a hidden relationship between the atom coordinates of a protein and a functional parameter of interest, yielding robust and parsimonious solutions that are superior to those obtained from principal component regression. In this section we look at a particular protein studied in Krivobokova et al. (2012), the water channel aquaporine found in the yeast *Pichia pastoris*. This is a gated channel, i.e., the diameter of its opening can change, which controls the flow of water into the cell. Our goal is to study which collective motions of protein atoms influence the diameter y_t of the channel at time t, as measured by the distance between two centres of mass of the residues of the protein that 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 between each backbone atom of the protein and a set of four fixed base points. We denote the $p = 739$ atoms by $A_{t,1}, \ldots, A_{t,p} \in \mathbb{R}^3$ and the fixed base points by $B_1, \ldots, B_4 \in \mathbb{R}^3$, and we let

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

The available time-frame has a length of 100 ns, with $n = 20\,000$ equidistant points of observation. Krivobokova et al. (2012) found that a linear relationship between X and y can be assumed. Additionally, these data seem not to contradict model (3). 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 Fig. 2. For the calculation of $\hat{V}^2$ we used the banding approach mentioned in § 4·2 and found the results to be very similar

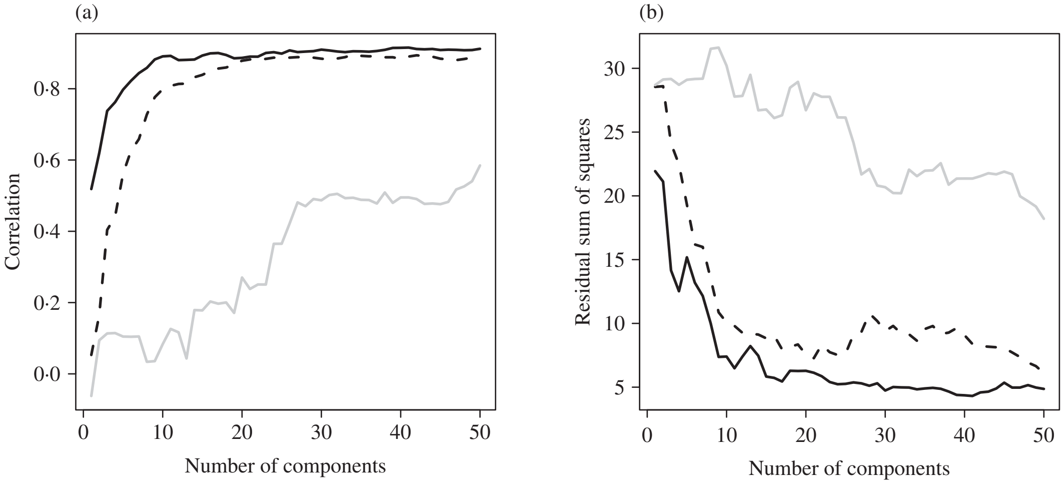

Fig. 3. (a) Correlation coefficient and (b) residual sum of squares for the predicted opening diameter and the real data on the test set. The compared methods are principal component regression (grey), corrected partial least squares (solid black) and partial least squares (dashed black).

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 used previously 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 built on the first half. Then the prediction was done on the test set consisting of the second half of the data and the results were 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 of these measures are plotted Fig. 3. The partial least squares estimator clearly outperforms principal component regression. The corrected partial least squares algorithm, which takes temporal dependence into account, delivers better prediction than standard partial least squares. The improvement is most noticeable in the first components.

High predictive power of the first corrected partial least squares components is particularly relevant to the interpretation of the underlying protein dynamics. Krivobokova et al. (2012) established that the first partial least squares regression coefficient $\hat{\beta}_1$ corresponds to the so-called ensemble-weighted maximally correlated mode of motion, which contributes most to the fluctuation in the response y. Overall, because of the low dimensionality, corrected partial least squares greatly facilitates interpretation of the underlying relevant dynamics, compared with partial least squares and principal component regression, where many more components are required to attain the same predictive power.

Acknowledgement

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

Supplementary material

Supplementary Material available at *Biometrika* online includes all the technical details.

References

Alakent, B., Doruker, P. & Camurdan, M. (2004). Time series analysis of collective motions in proteins. *J. Chem. Phys.* **120**, 1072–88.

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

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

Blanchard, G. & Krämer, N. (2010a). Kernel partial least squares is universally consistent. In *Proc. 13th Int. Conf. Artif. Intel. Statist.*, Y. W. Teh, ed., vol. 9 of *JMLR Workshop and Conference Proceedings*. Cambridge, Massachusetts: Journal of Machine Learning Research, pp. 57–64.

Blanchard, G. & Krämer, N. (2010b). Optimal learning rates for kernel conjugate gradient regression. *Adv. Neural Info. Proces. Syst.* **23**, 226–34.

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

de Groot, B. L., Hayward, S., van Aalten, D. M. F., Amadei, A. & Berendsen, H. J. C. (1998). Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data. *Proteins* **31**, 116–27.

Delaigle, A. & Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40**, 322–52.

Farkas, O. & 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. Info. Mod.* **45**, 339–46.

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

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

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

Hahn, C., Johnson, M., Herrmann, A. & Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Bus. Rev.* **54**, 243–69.

Helland, I. S. (1988). On the structure of partial least squares regression. *Commun. Statist.* B **17**, 581–607.

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

Henzler-Wildman, K. & Kern, D. (2007). Dynamic personalities of proteins. *Nature* **450**, 964–72.

Höskuldsson, A. (1988). PLS regression methods. *J. Chemomet.* **2**, 211–28.

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

Krämer, N. (2007). An overview on the shrinkage properties of partial least squares regression. *Comp. Statist.* **22**, 249–73.

Krivobokova, T., Briones, R., Hub, J., Munk, A. & de Groot, B. (2012). Partial least squares functional mode analysis: Application to the membrane proteins AQP1, Aqy1 and CLC-ec1. *Biophys. J.* **103**, 786–96.

Lobaugh, N., West, R. & McIntosh, A. (2001). Spatiotemporal analysis of experimental differences in event-related potential data with partial least squares. *Psychophysiol.* **38**, 517–30.

Naik, P. & Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *J. R. Statist. Soc.* B **62**, 763–71.

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

Phatak, A. & de Hoog, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: Alternative proofs of some properties of PLS. *J. Chemomet.* **16**, 361–7.

Skrondal, A. & Rabe-Hesketh, S. (2006). Latent variable modelling: A survey. *Scand. J. Statist.* **34**, 712–45.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedure. In *Research Papers in Statistics: Festschrift for J. Neyman*, F. N. David, ed. London: Wiley, pp. 411–44.

Wold, S., Sjøstrøma, M. & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemomet. Intel. Lab.* **58**, 109–30.

Wu, W. & Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sinica* **19**, 1755–68.

Wu, W. & Xiao, H. (2012). Covariance matrix estimation in time series. *Ann. Statist.* **40**, 466–93.

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

[*Received August* 2015. *Revised February* 2016]