Kernel partial least squares for stationary data

Tatyana Krivobokova, Marco Singer, Axel Munk

Georg-August-Universität Göttingen

Bert de Groot

Max Planck Institute for Biophysical Chemistry

Van Dantzig Seminar,
06 April 2017
Motivating example

Proteins

- are large biological molecules
- function often requires dynamics
- configuration space is high-dimensional

Group of Bert de Groot seeks to identify a relationship between

collective atomic motions of a protein

and

some specific protein’s (biological) function.
Motivating example

The data from the Molecular Dynamics (MD) simulations:

- $Y_t \in \mathbb{R}$ is a functional quantity of interest at time $t$, $t = 1, \ldots, n$
- $X_t \in \mathbb{R}^{3N}$ are Euclidean coordinates of $N$ atoms at time $t$

Stylized facts

- $d = 3N$ is typically high, but $d \ll n$
- $\{X_t\}_t$, $\{Y_t\}_t$ are (non-)stationary time series
- some (large) atom movements might be unrelated to $Y_t$

Functional quantity $Y_t$ to be modelled a function of $X_t$. 
Yeast aquaporin (AQY1)

- Gated water channel
- $Y_t$ is the opening diameter (red line)
- 783 backbone atoms
- $n = 20,000$ observations on 100 ns timeframe
AQY1 time series

Movements of the first atom and the diameter of channel opening

![Coordinate vs Time](image1)

![Diameter vs Time](image2)
Model

Assume

\[ Y_t = f(X_t) + \epsilon_t, \quad t = 1, \ldots, n, \]

where

- \( \{X_t\}_t \) is a \( d \)-dimensional stationary time series
- \( \{\epsilon_t\}_t \) i.i.d. zero mean sequence independent of \( \{X_t\}_t \)
- \( f \in \mathcal{L}^2(P\tilde{X}), \tilde{X} \) is independent of \( \{X_t\}_t \) and \( \{\epsilon_t\}_t \) and \( P\tilde{X} = PX_1 \)

The closeness of an estimator \( \hat{f} \) of \( f \) is measured by

\[
\| \hat{f} - f \|_2^2 = E_{\tilde{X}} \left\{ \hat{f}(\tilde{X}) - f(\tilde{X}) \right\}^2.
\]
Hub, J.S. and de Groot, B. L. (2009) assumed a linear model

$$Y_i = X_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n,$$

$$X_i \in \mathbb{R}^d,$$ or in matrix form $Y = X\beta + \epsilon$, ignored dependence in the data and tried to regularise the estimator by using PCA.
Motivating example

PC regression with 50 components

![Plot of time in ns vs number of components and correlation]
Motivating example

Partial Least Squares (PLS) leads to superior results

![Graph showing the relationship between the number of components and correlation for PLS and PCR. The PLS line is consistently higher than the PCR line, indicating superior results.]
Regularisation with PCR and PLS

Consider a linear regression model with fixed design

\[ Y = X\beta + \epsilon. \]

In the following let \( A = X^T X \) and \( b = X^T Y \).

PCR and PLS regularise \( \beta \) with a transformation \( H \in \mathbb{R}^{d \times s} \) s.t.

\[
\hat{\beta}_s = H \arg \min_{\alpha \in \mathbb{R}^s} \frac{1}{n} \| Y - XH\alpha \|^2 = H(H^T A H)^{-1} H^T b,
\]

where \( s \leq d \) plays the role of a regularisation parameter.

In PCR matrix \( H \) consists of the first \( s \) eigenvectors of \( A = X^T X \).
Regularisation with PLS

In PLS one derives $H = (h_1, \ldots, h_s)$, $h_i \in \mathbb{R}^d$ as follows

1. Find

$$h_1 = \arg \max_{h \in \mathbb{R}^d, \|h\|=1} \text{cov}(Xh, Y)^2 \propto X^T Y = b$$

2. Project $Y$ orthogonally:

$$Xh_1(h_1^T A h_1)^{-1} h_1^T X^T Y = X \hat{\beta}_1$$

3. Iterate the procedure according to

$$h_i = \arg \max_{h \in \mathbb{R}^d, \|h\|=1} \text{cov}(Xh, Y - X \hat{\beta}_{i-1})^2, \quad i = 2, \ldots, s$$

Apparently, $\hat{\beta}_s$ is highly non-linear in $Y$. 

Regularisation with PLS

For PLS is known that \( h_i \in \mathcal{K}_i(A, b), \ i = 1, \ldots, s, \) where

\[
\mathcal{K}_i(A, b) = \text{span}\{b, Ab, \ldots, A^{i-1}b\}
\]

is a Krylov space of order \( i \).

With this the alternative definition of PLS is

\[
\hat{\beta}_s = \arg \min_{\beta \in \mathcal{K}_s(A, b)} \| Y - X\beta \|^2.
\]

Note that any \( \beta_s \in \mathcal{K}_s(A, b) \) can be represented as

\[
\beta_s = P_s(A)b = P_s(X^TX)X^TY = X^TP_s(XX^T)Y,
\]

where \( P_s \) is a polynomial of degree at most \( s - 1 \).
Regularisation with PLS

For the implementation and proofs the residual polynomials

\[ R_s(x) = 1 - xP_s(x) \]

are of interest. Polynomials \( R_s \)

- are orthogonal w.r.t. an appropriate inner product
- satisfy a recurrence relation

\[ R_{s+1}(x) = a_s x R_s(x) + b_s R_s(x) + c_s R_{s-1}(x) \]

- are convex on \([0, r_s]\), where \( r_s \) is the first root of \( R_s(x) \) and \( R_s(0) = 1 \).
PLS and conjugate gradient

PLS is closely related to the conjugate gradient (CG) algorithm for

\[ A\beta = X^T X\beta = X^T Y = b. \]

The solution of this linear equation by CG is defined by

\[ \hat{\beta}_s^{CG} = \arg \min_{\beta \in K_s(A,b)} \| b - A\beta \|^2 = \arg \min_{\beta \in K_s(A,b)} \| X^T (Y - X\beta) \|^2. \]
CG in deterministic setting

CG algorithm has been studied in Nemirovskii (1986) as follows:

- Consider $\bar{A}\beta = \bar{b}$ for a linear bounded $\bar{A} : \mathcal{H} \rightarrow \mathcal{H}$
- Assume that only approximation $A$ of $\bar{A}$ and $b$ of $\bar{b}$ are given
- Set $\hat{\beta}_s^{CG} = \arg\min_{\beta \in \mathcal{K}_s(A,b)} \| b - A\beta \|_H^2$. 
CG in deterministic setting

Assume

(A1) $\max\{\|\bar{A}\|_{op}, \|A\|_{op}\} \leq L$, $\|\bar{A} - A\|_{op} \leq \epsilon$ and $\|\bar{b} - b\|_{\mathcal{H}}^2 \leq \delta$

(A2) The stopping index $\hat{s}$ satisfies the discrepancy principle

$$\hat{s} = \min\{s > 0 : \|b - A\hat{\beta}_s\|_{\mathcal{H}} < \tau(\delta\|\hat{\beta}_s\|_{\mathcal{H}} + \epsilon)\}, \quad \tau > 0$$

(A3) $\beta = \bar{A}^\mu u$ for $\|u\|_{\mathcal{H}} \leq R$, $\mu, R > 0$ (source condition).

Theorem (Nemirovskii, 1986)

Let (A1) – (A3) hold and $\hat{s} < \infty$. Then for any $\theta \in [0, 1]$

$$\|\bar{A}^\theta (\hat{\beta}_{\hat{s}} - \beta)\|_{\mathcal{H}}^2 \leq C(\mu, \tau) R^{\frac{2(1-\theta)}{1+\mu}} (\epsilon + \delta RL^\mu)^{\frac{2(\theta+\mu)}{1+\mu}}.$$
Kernel regression

A nonparametric model

\[ Y_i = f(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad X_i \in \mathbb{R}^d \]

is handled in the reproducing kernel Hilbert space (RKHS) framework.

Let \( \mathcal{H} \) be a RKHS, that is

- \((\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})\) is a Hilbert space of functions \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) with
- a kernel function \( k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \), s.t. \( k(\cdot, x) \in \mathcal{H} \) and

\[ f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}, \quad x \in \mathbb{R}^d, \quad f \in \mathcal{H}. \]

Unknown \( f \) is estimated by \( \hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(\cdot, X_i) \).
Kernel regression

Define operators

- Sample evaluation operator (analogue of $X$):
  \[ T_n : f \in \mathcal{H} \mapsto \{ f(X_1), \ldots, f(X_n) \}^T \in \mathbb{R}^n \]

- Sample kernel integral operator (analogue of $X^T/n$):
  \[ T_n^* : u \in \mathbb{R}^n \mapsto n^{-1} \sum_{i=1}^n k(\cdot, X_i)u_i \in \mathcal{H} \]

- Sample kernel covariance operator (analogue of $XX^T/n$):
  \[ S_n = T_n^* T_n : f \in \mathcal{H} \mapsto n^{-1} \sum_{i=1}^n f(X_i)k(\cdot, X_i) \in \mathcal{H} \]

- Sample kernel (analogue of $XX^T/n$):
  \[ K_n = T_n T_n^* = n^{-1}\{ k(X_i, X_j) \}_{i,j=1}^n \]
Now we can define the kernel PLS estimator as

\[
\hat{\alpha}_s = \arg \min_{\alpha \in \mathcal{K}_s(K_n, Y)} \| Y - K_n \alpha \|^2 = \arg \min_{\alpha \in \mathcal{K}_s(T_n T^*_n, Y)} \| Y - T_n T^*_n \alpha \|^2,
\]

or, equivalently, for \( f = T^*_n \alpha \)

\[
\hat{f}_s = \arg \min_{f \in \mathcal{K}_s(S_n, T^*_n Y)} \| Y - T_n f \|^2, \quad s = 1, \ldots, n.
\]

The kernel CG estimator is then defined as

\[
\hat{f}^{CG}_s = \arg \min_{f \in \mathcal{K}_s(S_n, T^*_n Y)} \| T^*_n (Y - T_n f) \|^2_{\mathcal{H}}.
\]
Results for Kernel CG and PLS

Blanchard and Krämer (2010)

- used stochastic setting with i.i.d. data \((Y_i, X_i)\)
- argued that the proofs for kernel CG can not be directly transferred to kernel PLS

In this work we

- use stochastic setting with dependent data
- prove convergence rates for kernel PLS

building up on Hanke (1995) and Blanchard and Krämer (2010).
Kernel PLS: assumptions

Consider now the model specified for the protein data

\[ Y_t = f(X_t) + \epsilon_t, \quad t = 1, \ldots, n. \]

Let \( \mathcal{H} \) be a RKHS with kernel \( k \) and assume

(C1) \( \mathcal{H} \) is separable;
(C2) \( \exists \ k > 0 \text{ s.t. } |k(x, y)| \leq k, \forall x, y \in \mathbb{R}^d \) and \( k \) is measurable;

Under (C1) the Hilbert-Schmidt norm of operators from \( \mathcal{H} \) to \( \mathcal{H} \) is well-defined and (C2) implies that all functions in \( \mathcal{H} \) are bounded.
Kernel PLS: assumptions

Let $T$ and $T^*$ be population versions of $T_n$ and $T_n^*$:

\[
T : f \in \mathcal{H} \mapsto f \in L^2(P\tilde{X}) \\
T^* : f \in L^2(P\tilde{X}) \mapsto \int f(x)k(\cdot, x)dP\tilde{X}(x) \in \mathcal{H}.
\]

It implies population versions of $S_n$ and $K_n$:

\[
S = T^* T \quad \text{and} \quad K = TT^*.
\]

Operators $T$ and $T^*$ are adjoint and $S$, $K$ are self-adjoint.
Kernel PLS: assumptions

As in Nemirovskii (1986) we use the source condition as an assumption on regularity of $f$:

(SC) $\exists \ r \geq 0, \ R > 0$ and $u \in L^2(P\hat{X})$ s.t. $f = K^r u$ and $\|u\|_2 \leq R$

If $r \geq 1/2$, then $f \in L^2(P\hat{X})$ coincides a.s. with $f_\mathcal{H} \in \mathcal{H}$ ($f = T f_\mathcal{H}$). The setting with $r < 1/2$ is referred to as the outer case.
Kernel PLS: assumptions

Under suitable regularity conditions due to Mercer’ theorem

\[ K(x, y) = \sum_i \eta_i \phi_i(x) \phi_i(y) \]

for an orthonormal basis \( \{\phi_i\}_{i=1}^{\infty} \) for \( L^2(P\tilde{X}) \) and \( \eta_1 \geq \eta_2 \geq \ldots \).

Hence,

\[ \mathcal{H} = \left\{ f : f = \sum_i \theta_i \phi_i(x) \in L^2(P\tilde{X}) \text{ and } \sum_i \frac{\theta_i^2}{\eta_i} < \infty \right\} . \]

The source condition corresponds to \( f \in \mathcal{H}_r \), where

\[ \mathcal{H}_r = \left\{ f : f = \sum_i \theta_i \phi_i(x) \in L^2(P\tilde{X}) \text{ and } \sum_i \frac{\theta_i^2}{\eta_i^{2r}} \leq R^2 \right\} . \]
Kernel PLS: first result

Theorem (Singer, K., Munk, 2017)

Assume (C1), (C2) and (SC) hold with $r \geq 3/2$, as well as

\[
P(\| S_n - S \|_{HS} \leq C_\delta \gamma_n) \geq 1 - \nu/2
\]
\[
P(\| T_n^* Y - Sf \|_{\mathcal{H}} \leq C_\epsilon \gamma_n) \geq 1 - \nu/2,
\]

for constants $C_\epsilon, C_\delta > 0$, $\nu \in (0, 1]$ and a sequence $\{\gamma_n\}_{n \in [0, \infty)}$, $\gamma_n \to 0$. Define the stopping index with $C = C(\nu, C_\epsilon, C_\delta, r, \kappa, R)$

\[
\hat{s} = \min \left\{ 1 \leq s \leq n : \sum_{i=0}^{s} \| S_n \hat{f}_i - T_n^* Y \|_{\mathcal{H}}^{-2} \geq (C \gamma_n)^{-2} \right\}.
\]

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

\[
\| \hat{f}_{\hat{s}} - f \|_2 = O \left\{ \gamma_n^{2r/(2r+1)} \right\}.
\]
Kernel PLS: first result

• The rate of convergence is driven by $\gamma_n$, which enters the concentration inequalities.

• $\gamma_n = O(n^{-1/2})$ results in the same convergence rates as in Blanchard & Krämer (2010) for independent data.

• The rate is adaptive: $\hat{s}$ is independent of $r$.

• The stopping rule for the kernel CG has the form $\|S_n\hat{f}_s^{CG} - T_n^* Y\|_{\mathcal{H}} \leq C\gamma_n$. 
Kernel PLS: assumptions

The optimal rates depend both on the regularity of the function and on the structure of $\mathcal{H}$ described e.g. via $\text{tr} \left\{ K(K + \lambda I)^{-1} \right\}$.

Zhang (2005) suggested the concept of effective dimensionality (ED) $\exists \, \zeta \in (0, 1], \, D > 0 \text{ s.t. } \text{tr} \left\{ K(K + \lambda I)^{-1} \right\} \leq D \lambda^{-\zeta}, \forall \lambda > 0.$ and found the optimal convergence rates that depends on $r$ and $\zeta$.

For example, if $\eta_i \leq c \, i^{-1/\zeta}$, then

$$\text{tr} \left\{ K(K + \lambda I)^{-1} \right\} = \sum_{i} \frac{\eta_i}{\eta_i + \lambda} \leq \tilde{c}(\alpha, c) \lambda^{-\zeta}.$$
Kernel PLS: second result

To adapt the results of Caponnetto & De Vito (2007) to our setting, the following concentration inequalities (CI) need to hold:

\[
P(\|S_n - S\|_{HS} \leq C_\delta \gamma_n) \geq 1 - \nu/3
\]
\[
P(\|(S + \lambda)^{-1/2}(T^*_n Y - Sf)\|_\mathcal{H} \leq C_\epsilon \lambda^r) \geq 1 - \nu/3
\]
\[
P(\|(S + \lambda)(S_n + \lambda)^{-1}\|_{HS} \leq C_\psi^2) \geq 1 - \nu/3
\]

for \(C_\epsilon, C_\delta, C_\psi > 0\), \(\lambda > 0\), \(\nu \in (0, 1]\) and a sequence \(\{\gamma_n\}_n, \gamma_n \to 0\).
Theorem (Singer, K., Munk, 2017)

Let \((C1), (C2), (SC), (ED)\) hold with \(r \geq 1/2\) and \(\zeta \in (0, 1]\), as well as \((CI)\) with \(\lambda \propto \gamma_n^{2/(2r+\zeta)}\). Define the stopping index \(\hat{s}\) by

\[
\hat{s} = \min \left\{ 1 \leq s \leq n : \sum_{i=0}^{s} \| S_n \hat{f}_i - T_n^* Y \|_{\mathcal{H}}^{-2} \geq (C \gamma_n)^{-2r/(2r+\zeta+1)} \right\}
\]

for \(C = C(\nu, C_{\epsilon}, C_{\delta}, C_{\psi}, \kappa, r, R, D)\). Then it holds with probability at least \(1 - \nu\)

\[
\| \hat{f}_{\hat{s}} - f \|_2 = O \left\{ \gamma_n^{2r/(2r+\zeta)} \right\}.
\]
Kernel PLS: second result

Similar to Blanchard & Krämer (2010):

• Rates obtained in the theorem without (ED) correspond to the worst case $\zeta = 1$, but are adaptive.

• Rates obtained in the theorem with (ED) are optimal if $\gamma_n = O(n^{-1/2})$, but require the knowledge of $r$ and $\zeta$ for $\hat{s}$.

• For the outer case $f \notin \mathcal{H}$ additional assumptions are needed to obtain the optimal rate, see e.g. Mendelson & Neeman (2009).
Kernel PLS: Concentration inequalities

Under (C1) and (C2) it holds with probability at least $1 - \nu$ that

$$\|S_n - S\|_{HS}^2 \leq \frac{\delta_n}{\nu} \quad \text{and} \quad \|T_n^* Y - Sf\|_H^2 \leq \frac{\epsilon_n}{\nu},$$

where

$$\delta_n = \frac{C_1}{n} + \frac{2}{n^2} \sum_{h=2}^{n} (n - h) \int_{\mathbb{R}^{2d}} k^2(x, y) d\mu_h(x, y)$$

and

$$\epsilon_n = \frac{C_2}{n} + \frac{2}{n^2} \sum_{h=2}^{n} (n - h) \int_{\mathbb{R}^{2d}} k(x, y) f(x) f(y) d\mu_h(x, y)$$

for $d\mu_h(x, y) = dP^{X_h, X_1}(x, y) - dP^{X_1}(x) dP^{X_1}(y)$. 
Kernel PLS: Concentration inequalities

Hence, $\gamma_n \propto (\delta_n + \epsilon_n)$ converges to zero iff the sums in $\delta_n$ and $\epsilon_n$ are of order not larger than $n^{2-\varepsilon}$, $\varepsilon > 0$.

We make additional assumptions on $\{X_t\}_t$:

(D1) $X_1 \sim \mathcal{N}_d(0, \sigma_1 \Sigma)$, $(X_h, X_1)^T \sim \mathcal{N}_{2d}(0, \Sigma_h)$, $h = 2, \ldots, n$ with

$$
\Sigma_h = \begin{pmatrix}
\sigma_1 & \sigma_h \\
\sigma_h & \sigma_1
\end{pmatrix} \otimes \Sigma,
$$

where $\Sigma$ is a positive definite symmetric matrix.

(D2) For $\rho_h = \sigma_1^{-1} \sigma_h$ there exists $q > 0$ and $0 < c_1 < c_2$ such that

$$
c_1 h^{-q} \leq |\rho_h| \leq c_2 h^{-q}, \quad h = 1, \ldots, n.
$$
Kernel PLS: Concentration inequalities

If additionally to (C1) and (C2), also (D1) and (D2) hold, then

\[ \delta_n \leq C_1\{\phi_n(q) + n^{-1}\} \quad \text{and} \quad \epsilon_n \leq C_2\{\phi_n(q) + n^{-1}\}, \]

for suitable \( C_1, C_2 > 0 \) and

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

for the Riemann-zeta function \( \tilde{\zeta}(q) \).
Kernel PLS with Gaussian data

Under assumptions of Theorem 1 and (D1), (D2) we get

$$\| \hat{f} - f \|_2 \leq \begin{cases} O\{n^{-r/(2r+1)}\}, & q > 1, \\ O\{n^{-qr/(2r+1)}\}, & q \in (0, 1) \end{cases}$$

Under assumptions of Theorem 2 and (D1), (D2) we get

$$\| \hat{f} - f \|_2 \leq \begin{cases} O\{n^{-r/(2r+\zeta)}\}, & q > 1, \\ O\{n^{-qr/(2r+\zeta)}\}, & q \in (0, 1) \end{cases}$$

Stationary data with $q > 1$ do not alter the convergence rate, in contrast to the long-range dependent data with $q \in (0, 1)$.
Let $\mathcal{H}$ be the RKHS corresponding to $K(x, y) = \exp(-l\|x - y\|^2)$, $l > 0$ and take $f \in \mathcal{H}$:
Simulations

$L_2$ errors of KPLS and KCG for different sample sizes and dependence

Independent

Autoregressive

Long-range
Simulations

Stopping times (CV) of KPLS and KCG for different sample sizes and i.i.d. data

$n = 200$

$n = 1000$
Aquaporin data are well-described by a linear model; CPLS is a linear PLS that takes into account dependence in the data:
Another protein: T4 Lysozyme of the bacteriophage T4; \( n = 4601, \ d = 3 \cdot 486 \) estimated by KPLS, KPCR and PLS.