Partial least squares for dependent data

Tatyana Krivobokova, Marco Singer, Axel Munk

Georg-August-Universität Göttingen

Bert de Groot

Max Planck Institute for Biophysical Chemistry

Tsinghua University,

18 September 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:

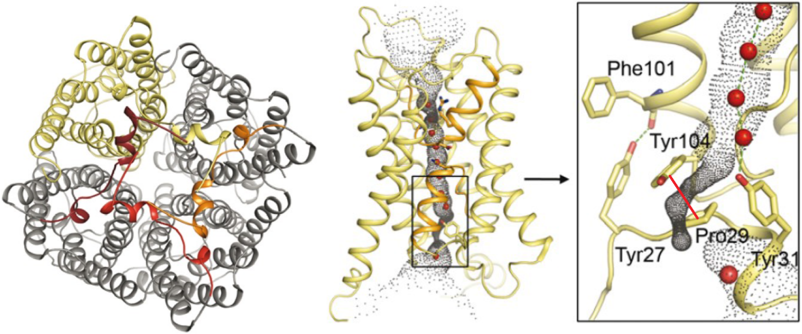
- $Y_t \in \mathbb{R}$ is a functional quantity of interest at time t, $t = 1, \ldots, n$
- $X_t \in \mathbb{R}^{3N}$ are Euclidean coordinates of N atoms at time t

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 is to be modelled as a function of X_t.

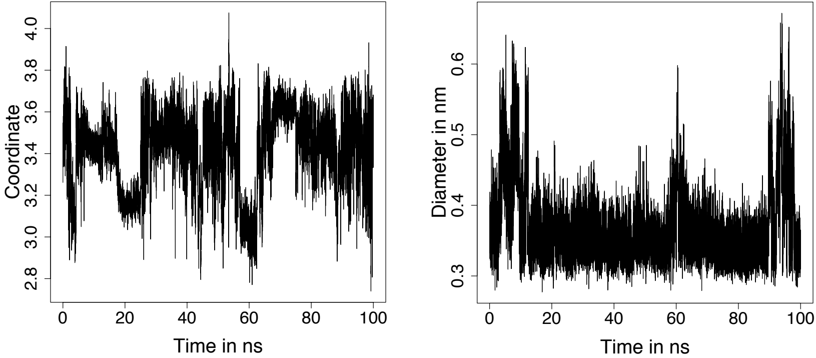
Yeast aquaporin (AQY1)



- 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 channel opening diameter



Simple linear case

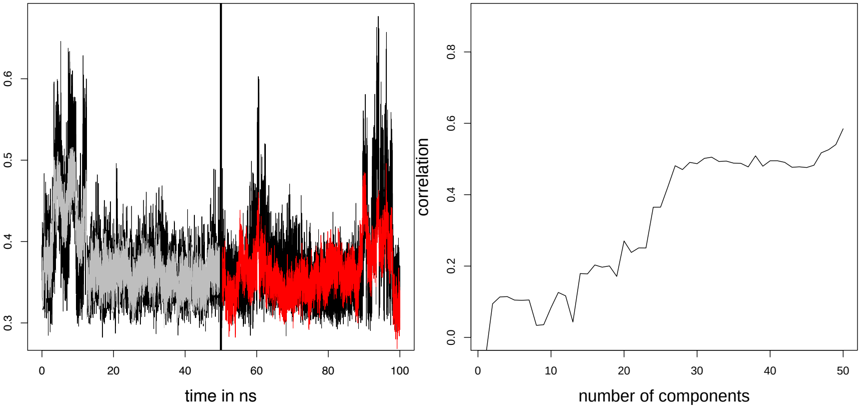
Hub, J.S. and de Groot, B. L. (2009) assumed a linear model

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

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

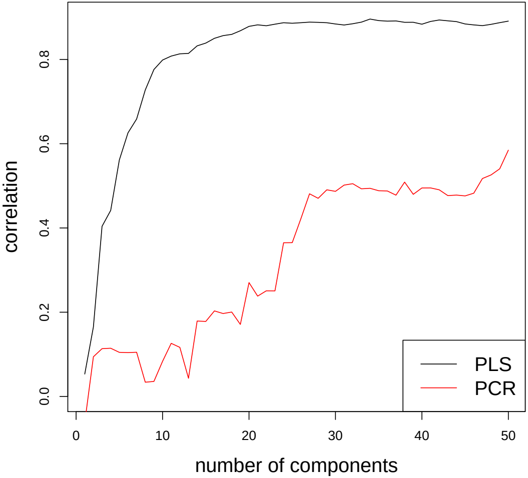
Motivating example

PC regression with 50 components



Motivating example

Partial Least Squares (PLS) lead to 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 β with a transformation $H \in \mathbb{R}^{d\times s}$ s.t.

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

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

Regularisation with PCR

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

$$\begin{aligned} h_1 &= \arg \max_{\substack{h \in \mathbb{R}^p \\ \|h\|=1}} \widehat{\text{cov}}(h^t x) \\ h_i &= \arg \max_{\substack{h \in \mathbb{R}^p \\ \|h\|=1}} \widehat{\text{cov}}(h^t x), \text{ s.t. } h_1 \perp \dots \perp h_i, \quad i = 2, \dots, k \end{aligned}$$

Since $\widehat{\text{cov}}(h^t x) = h^t X^t X h / n$, h_i is the ith eigenvector of $X^t X / n$.

Regularisation with PLS

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

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

2. Project Y orthogonally: $Xh_1(h_1^T A\, h_1)^{-1} h_1^T X^T Y = X\widehat{\beta}_1$

3. Iterate the procedure according to
$$h_i = \arg \max_{\substack{h \in \mathbb{R}^d \\ \|h\|=1}} \widehat{\operatorname{cov}}(Xh, Y - X\widehat{\beta}_{i-1})^2, \quad i = 2, \ldots, s$$

Theoretical properties of PLS

- PLS is highly non-linear in the response Y
- Little is known on statistical properties
- Influence of dependence in the data on PLS is unclear
- PLS is closely related to the conjugate gradient

PLS and Krylov spaces

For PLS is known that $h_i \in \mathcal{K}_i(A, b)$ ($A = X^t X$, $b = X^t Y$), where $\mathcal{K}_i(A, b) = \text{span}\{b, Ab, \dots, A^{i-1}b\}$ is a Krylov space of order i

With this the alternative definition of the PLS estimator is given by

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

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

$$\beta_s = P_s(A)b = P_s(X^T X)X^T Y = X^T P_s(XX^T)Y,$$

where P_s is a polynomial of degree at most $s - 1$.

Regularisation with PLS

For the implementation and proofs the residual polynomials

$$R_s(x) = 1 - x\, P_s(x)$$

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

$$\widehat{\beta}_s^{CG} = \arg \min_{\beta \in \mathcal{K}_s(A,b)} \|b - A\beta\|^2 = \arg \min_{\beta \in \mathcal{K}_s(A,b)} \|X^T(Y - X\beta)\|^2.$$

CG in deterministic setting

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

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

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 s satisfies the discrepancy principle

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

(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 $\theta \in [0,1]$

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

Results for CG and PLS

Blanchard and Krämer (2010)

- used stochastic setting with i.i.d. data (Y_i, X_i)
- proved convergence rates for kernel CG using ideas in Nemirovskii (1986), Hanke (1995), Caponnetto & de Vito (2007)
- argued that the proofs for kernel CG can not be directly transferred to kernel PLS

In two recent papers we

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

building upon Blanchard and Krämer (2010) and Hanke (1995).

First paper

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

Latent variable model

Standard linear model $Y = X\beta + \epsilon$ is extended by assuming

$$\begin{aligned} X &= T(NP^t + \eta F) \\ Y &= T(Nq + \varepsilon f) \end{aligned}$$

where N and F are random matrix ($n \times l$, $n \times d$), f is a random vector
N, F, f are independent with i.i.d. entries, mean 0 and variance 1;
$T \in \mathbb{R}^{n\times n}$, $P \in \mathbb{R}^{d\times l}$ and $q \in \mathbb{R}^l$ are deterministic, $\eta, \varepsilon \geq 0$.

If T^2 is a covariance matrix, then one can interpret X as
a matrix form of a time series $\{X_t\}_{t=1}^n$, $X_t = (X_{t,1}, \ldots, X_{t,d})$
and Y as a vector of a real-valued time series $\{Y_t\}_{t=1}^n$.

In this latent model (compare setting of Nemirovskii)

$$\bar{A} = PP^t + \eta^2 I_d \quad \text{is estimated by} \quad A = X^t X/n$$
$$\bar{b} = Pq \quad \text{is estimated by} \quad b = X^t Y/n$$

and the PLS estimators are obtained as before

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

Note that the true parameter is $\beta(\eta) = (PP^t + \eta^2 I_d)^{-1} Pq$.

First result

Theorem (Singer, K., Munk, de Groot, 2016)

If under the latent variable model the fourth moments of N_{11}, F_{11} exist, then for $A = X^t X/\|T\|^2$, $b = X^t Y/\|T\|^2$

$$
\begin{aligned}
E\|\bar{A} - A\|^2 &= \frac{\|T^2\|^2}{\|T\|^4}\left(c_1 + \sum_{t=1}^{n} \frac{\|T_t\|^4}{\|T^2\|^2} c_2\right) \\
E\|\bar{b} - b\|^2 &= \frac{\|T^2\|^2}{\|T\|^4}\left(c_3 + \sum_{t=1}^{n} \frac{\|T_t\|^4}{\|T^2\|^2} c_4\right),
\end{aligned}
$$

where c_i, $i = 1, 2, 3, 4$ are known and independent of n.

Second result

For the standard PLS algorithm we find

Theorem (Singer, K., Munk, de Groot, 2016)

Let the latent variable model with $\eta > 0$ hold and the fourth moments of N_{11}, F_{11} exist. Let also $\hat{s}$ be the first index $0 < \hat{s} \leq d$ such that

$$\|X^t(X\hat{\beta}_{\hat{s}} - Y)\|^2 \leq \rho_1\|\hat{\beta}_{\hat{s}}\| + \rho_2$$

for $\rho_1, \rho_2 \to 0$. Then it holds with probability at least $1-\gamma$, $\gamma \in (0,1]$

$$\|\hat{\beta}_{\hat{s}} - \beta(\eta)\| \quad \leq \quad \frac{\|T^2\|}{\|T\|^2}\left\{c_5(\gamma) + \frac{\|T^2\|}{\|T\|^2}c_6(\gamma)\right\}$$

where $c_5(\gamma)$ and $c_6(\gamma)$ are known and independent of n.

Convergence term

	$\|T^2\|$	$\|T\|^2$	$\|T^2\|\|T\|^{-2}$
Independence	$\sqrt{n}$	n	$n^{-1/2}$
AR	$\sim \sqrt{n}$	n	$\sim n^{-1/2}$
ARIMA	$\sim n^2$	$\sim n^2$	$\sim c,\ c > 0$

Population Krylov space elements $\bar{A}$ and $\bar{b}$ can not be estimated consistently for non-stationary processes; what about $\hat{\beta}_i$?

Third result

Since $\hat{\beta}_i$ is highly non-linear in Y, only $\hat{\beta}_1$ is feasible for the analysis

For the standard PLS algorithm we get

Theorem (Singer, K., Munk, de Groot, 2016)

Let the latent variable model hold and eights moments of N_{11}, F_{11} and f_1 exist.

If $\|T^2\|\|T\|^{-2} \nrightarrow 0$, then $\hat{\beta}_1(\eta)$ is an inconsistent estimator for $\beta_1(\eta)$

Corrected PLS

It seems natural to standardise the data before running PLS,

or, equivalently, to use $A_T = X^t \widehat{T}^{-2} X/n$ and $b_T = X^t \widehat{T}^{-2} Y/n$

Theorem (Singer, K., Munk, de Groot, 2016)

Let $\widehat{T}^2$ be a consistent estimator for T^2 s.t.

$\|T \widehat{T}^{-2} T - I_n\|_2 = O_p(r_n)$ *for some positive sequence* $r_n \to 0$, $n \to \infty$.

Then

$$\|\bar{A} - A_T\|_2 = O_p(r_n), \quad \|\bar{b} - b_T\| = O_p(r_n).$$

Moreover, with probability at least $1 - \nu$, $\nu \in (0,1)$

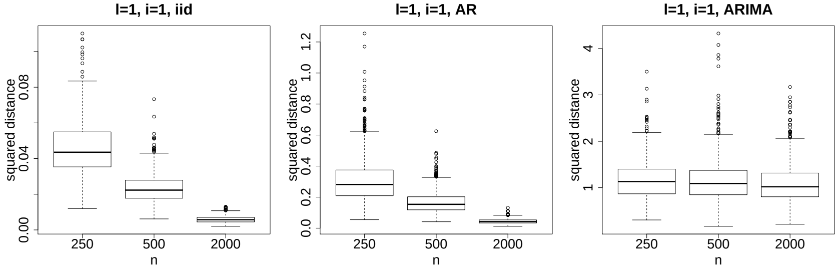
$$\|\hat{\beta}_{\hat{s}}(\widehat{T}) - \beta(\eta)\| = O(r_n).$$

Simulation setting

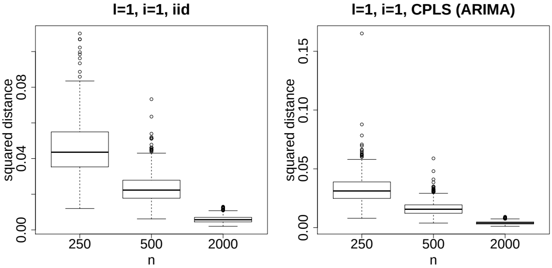
Latent variable model with $X = T(NP^t + \eta F)$, $Y = T(Nq + \varepsilon f)$

- N_{11}, F_{11}, f_1 are $\mathcal{N}(0,1)$, $d = 20$, $n \in \{250, 500, 2000\}$
- P_{ij} are i.i.d. $B(1, 0.5)$; $q_i = 1/i$
- η and ε chosen so that the signal-to-noise ratio is 2
- number of latent components $l = 1$
- T^2: identity, AR(1), ARIMA(1,1,1)
- $M = 1000$ Monte Carlo replications for $\hat{\beta}_1$

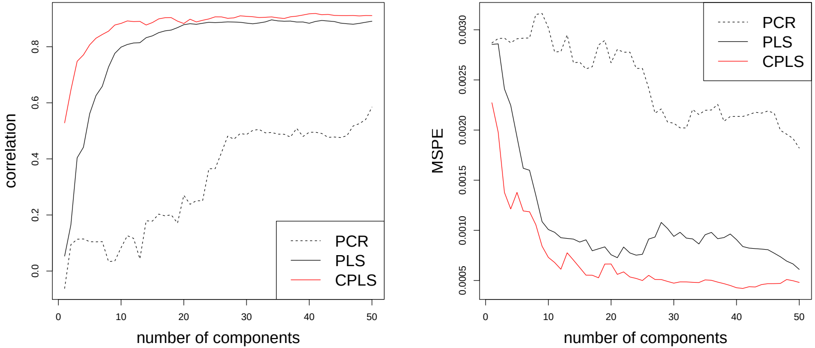
Simulation results



Simulation results



Protein data



Second paper

Singer, M., Krivobokova, T., Munk, A. (2017)
Kernel partial least squares for stationary data.
Conditionally accepted in the Journal of Machine Learning Research

Kernel regression

A nonparametric model

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

where

- $\{X_t\}_t$ is a d-dimensional stationary time series
- $\{\epsilon_t\}_t$ i.i.d. zero mean sequence independent of $\{X_t\}_t$
- $f \in \mathcal{L}^2(\rho_X)$, X is independent of $\{X_t\}_t$ and $\{\epsilon_t\}_t$ and $\rho_X = P^{X_1}$

A nonparametric regression model is treated 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 $f : \mathbb{R}^d \to \mathbb{R}$ with
- a kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, s.t. $k(\cdot, x) \in \mathcal{H}$ and

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

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

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):

  $T_n^* : u \in \mathbb{R}^n \mapsto n^{-1} \sum_{i=1}^n k(\cdot, X_i) u_i \in \mathcal{H}$

Kernel PLS and kernel CG

Now we can define the kernel PLS estimator as

$$\widehat{f_s} = \arg \min_{f \in \mathcal{K}_s(T_n^* T_n, T_n^* Y)} \|Y - T_n f\|^2.$$

The kernel CG estimator is defined as

$$\widehat{f}_s^{CG} = \arg \min_{f \in \mathcal{K}_s(T_n^* T_n, T_n^* Y)} \|T_n^*(Y - T_n f)\|_{\mathcal{H}}^2.$$

Kernel PLS: assumptions

Two standard (not restrictive) assumptions on $\mathcal{H}$

(C1) $\mathcal{H}$ is separable;

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

To obtain optimal convergence rates we need also assumptions on

- regularity of the true function f
- complexity of $\mathcal{H}$ (w.r.t. ρ_X)

which can be expressed in terms of the eigenvalues of

$$k(x, y) = \sum_{i=1}^{\infty} \eta_i \phi_i(x) \phi_i(y).$$

Source condition

Regularity of f is described by the source condition

(SC) $f \in \mathcal{H}_r$, $r \geq 1/2$, where

$$\mathcal{H}_r = \left\{ f : \; f = \sum_i \theta_i \phi_i(x) \in \mathcal{L}^2(\rho_X) \text{ and } \sum_i \frac{\theta_i^2}{\eta_i^{2(r+1/2)}} \leq R^2 \right\}.$$

Effective dimensionality condition

Complexity of $\mathcal{H}$ is described by the effective dimensionality

$$d_\lambda = \sum_{i=1}^{\infty} \frac{\eta_i}{\eta_i + \lambda}, \quad \lambda > 0$$

(ED1) $d_\lambda \leq C\lambda^{-\zeta}$, $\zeta \in (0, 1]$

(ED2) $d_\lambda \leq C \log(1 + \xi/\lambda)$, $\xi > 0$

Assumptions on the data

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

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

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

where Σ is a positive definite symmetric matrix.

(D2) For $\rho_h = \sigma^{-1}\sigma_h$ 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 with Gaussian data

With appropriate concentration inequalities and optimal stopping times we get under (C1), (C2), (D1), (D2), (SC) and (ED1)

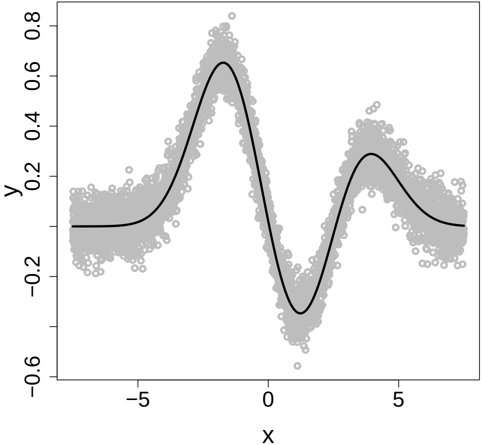
$$\|\widehat{f_{\hat{s}}} - f\|_2 = \begin{cases} O\{n^{-r/(2r+\zeta)}\}, & q > 1, \\ O\{n^{-qr/(2r+\zeta)}\}, & q \in (0,1). \end{cases}$$

while under (C1), (C2), (D1), (D2), (SC) and (ED2)

$$\|\widehat{f_{\hat{s}}} - f\|_2 = \begin{cases} O\{n^{-1/2}\log(n/2\}, & q > 1, \\ O\{n^{-q/2}\log(n^q/2)\}, & q \in (0,1). \end{cases}$$

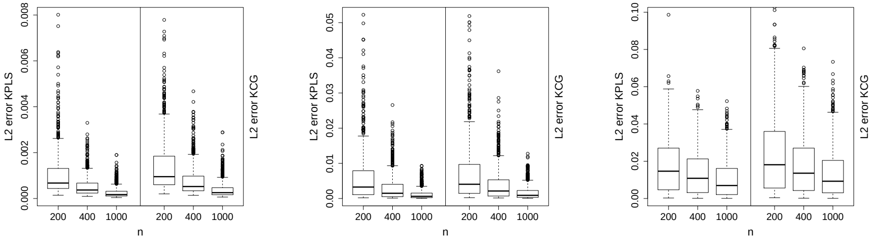
Stationary data with $q > 1$ do not alter the convergence rate, in contrast to the long-range dependent data with $q \in (0,1)$.

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

Protein data

Another protein: T4 Lysozyme of the bacteriophafe T4;
$n = 4601$, $d = 3 \cdot 486$ estimated by KPLS, KPCR and PLS.