The prediction error in functional linear regression

Christophe Crambes, Alois Kneip and Pascal Sarda

Université Paul Sabatier, Universität Bonn,
and Université Toulouse-le-Mirail

December 5, 2008
Introduction

In a number of important applications the outcome of a response variable \( Y \) depends on the variation of an explanatory variable \( X \) over time (age, etc.).

- Outcome \( Y \) of a chemical reaction depending on the varying temperatures \( X(t) \) in a reactor over some time interval \( I \).

- End price of an EBAY online auction (duration 7 days) in dependence of the price changes \( X(t) \) in the first 5 days of the auction.

- Maximum of ozone measured during a day in dependence of Curves representing repeated measurements of concentration in ozone measured the previous day.

Linear regression model based on \( p \) repeated measures of an explanatory variable (often \( p \gg n \)):

\[
Y_{ij} = \beta_0 + \sum_{j=1}^{p} \beta_j X_i(t_j) + \epsilon_i, \quad i = 1, \ldots, n, \ j = 1, \ldots, p
\]
Functional Regression:

- Scalar response variable $Y_i$, $i = 1, \ldots, n$
- The variation of $Y_i$ is modelled in dependence of a functional explanatory variable $X_i$; $X_i$ is a square integrable function defined on a compact interval $I$ of $\mathbb{R}$.
- $X_1, \ldots, X_n$ is a sequence of identically distributed random functions with the same distribution as a generic $X$. $X$ it is a second order variable, $\mathbb{E}(\int_I X^2(t)dt) < +\infty$.

Functional linear regression model:

$$Y_i = \beta_0 + \int_I \beta(t)X_i(t)dt + \epsilon_i, \quad i = 1, \ldots, n,$$

- $\epsilon_i$ - i.i.d. centered random errors, $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{E}(\epsilon_i^2) = \sigma^2_\epsilon$, and $\epsilon_i$ is independent of $X_i$.
- $\beta$ is a square integrable functional parameter defined on $I$ that must be estimated from the pairs $(X_i, Y_i)$, $i = 1, \ldots, n$. 
Sparseness and functional regression

“Sparseness literature”: High dimensional regression problem

\( p \) explanatory variables (usually not functional); \( p \) large compared to \( n \)

\[
Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad i = 1, \ldots, n
\]

Structural hypotheses:

- **Sparseness**: There is only a small number \( q \ll p \) of coefficients \( \beta_j \) which are nonzero.

- **Sufficiently weak correlations between explanatory variables.**

  Candes and Tao (2007): ”every set of columns with cardinality less than \( q \) approximately behaves like an orthonormal system”

Methods: LASSO, Dantzig selector

Some literature: For example Meinshausen and Bühlmann (2006), Candes and Tao (2007), Bickel, Ritov and Tsybakov (2007)
The setup of functional regression:
Discrete approximation ($p$ large); Equidistant design

\[ t_j - t_{j-1} = \frac{b-a}{n}, \ I = [a, b] \]

\[ Y_i = \sum_{j=1}^{p} \beta_j X_i(t_j) + \epsilon_i = \frac{1}{p} \sum_{j=1}^{p} \beta(t_j) X_{ij} \]

Structural setup:

- **No Sparseness of coefficients**: \( \beta_j \equiv \frac{\beta(t_j)}{p} \); \( \beta(t_j) \) discretized values of a *continuous* slope function \( \beta(t) \); in general, all coefficients \( \beta_j \) of “comparable” order of magnitude

- **Very strong correlations between \( X_i(t_j) \) and \( X_i(t_k) \), \( j \neq k \)**: \( X_1, X_2, \ldots \) i.i.d. sample of continuous random functions, equidistant design \( I = [a, b] \), \( t_j = a + (j-1)\frac{b-a}{p} \)

\[ \Rightarrow \text{for any fixed } j, \ m \in \mathbb{N} \]

\[ \text{corr}(X_i(t_j), X_i(t_{j+m})) \rightarrow 1 \quad \text{as } p \rightarrow \infty \]
But: Dimension reduction by (functional) principal components

**Discretized case:** Let $X_i = (X_i(t_1), \ldots, X_i(t_p))^T$. Assume that variables are centered and possess zero mean.

\[ \text{Covariance matrix } \Sigma = E(X_iX_i^T) \]

Let $l_1 > l_2 > \ldots$ and $\zeta_1, \zeta_2$ denote eigenvalues and a corresponding system of orthonormal eigenvectors (principal components) of $\Sigma$.

**Continuous case:** Covariance operator $\Gamma$ defined by

\[ \Gamma(\beta) := E \left( \int \beta(t)(X(t) - E(X)(t))dt \right) \]

Let $\lambda_1 > \lambda_2 > \ldots$ and $\gamma_1, \gamma_2, \ldots$ denote eigenvalues and a corresponding orthonormal system of eigenfunctions of $\Gamma$; necessarily $\sum_{r=1}^{\infty} \lambda_r < \infty$

Equidistant design, $X$ smooth: For fixed $r \in \mathbb{N}$

\[ \frac{l_r}{p} \rightarrow \lambda_r, \quad \frac{1}{p} \sum_{j} (\sqrt{p} \zeta_{rj} - \gamma_r(t_j))^2 \rightarrow 0 \]
Remark: Equidistant design, $X$ smooth: As $n,p \to \infty$

$$\frac{\hat{l}_r}{p} \to_p \lambda_r, \quad \frac{1}{p} \sum_j (\sqrt{p} \hat{\zeta}_{rj} - \gamma_r(t_j))^2 \to_p 0,$$

where $\hat{l}_1 > \hat{l}_2 > \ldots$ and $\hat{\zeta}_1, \hat{\zeta}_2$ denote eigenvalues and eigenvectors (principal components) of the empirical covariance matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_i X_i X_i^T$$

Insight: $l_1 \approx p \lambda_1, \ l_2 \approx p \lambda_2 \Rightarrow l_1 - l_2 \approx p(\lambda_1 - \lambda_2)$

For any $k = 1, 2, \ldots$ the first $k$ principal components (asymptotically) explain a fixed percentage of the variability of $X_i$. Let

- $\mathcal{E}_{p,k} \subset \mathbb{R}^p$ - $k$-dimensional eigenspace spanned by $\zeta_1, dots, \zeta_k$
- $\mathcal{E}_k \subset L^2([a, b])$ - $k$-dimensional eigenspace spanned by $\gamma_1, dots, \gamma_k$
- $\mathbb{L}$ - set of all $k$-dimensional linear subspaces of $L^2([a, b])$

As $p \to \infty$

$$
\mathbb{E} \left( \inf_{f \in \mathcal{E}_{p,k}} \frac{1}{p} \sum_{j=1}^{p} \left| X_i(t_j) - f(t_j) \right|^2 \right) \to \mathbb{E} \left( \inf_{f \in \mathcal{E}_k} \int_I \left| X_i(t) - f(t) \right|^2 \right)
$$

$$
= \inf_{\mathcal{L}_k \in \mathbb{L}} \mathbb{E} \left( \inf_{g \in \mathcal{L}_k} \int_I \left| X_i(t) - g(t) \right|^2 \right)
$$

$$
= \sum_{j \geq k+1} \lambda_j
$$

$X_i$ a.s. $\nu$-times continuously differentiable: $\sum_{j \geq k+1} \lambda_j = O(k^{-2\nu})$
Regression on functional principal components (continuous case):
$L^2(I)$ will be endowed with the inner product $\langle f, g \rangle = \int_I f(t)g(t)dt$ and its associated norm $\| . \|$.

With $\delta_{ri} = \langle X_i, \gamma_r \rangle$ and $\beta_r := \langle \beta, \gamma_r \rangle$ the functional regression model $Y_i = \langle \beta, X_i \rangle + \epsilon_i$ implies

$$Y_i = \sum_{r=1}^{\infty} \beta_r \delta_{ri} + \epsilon_i, \quad \beta(t) = \sum_{r=1}^{\infty} \beta_r \gamma_r(t)$$

Note: $var(\delta_{ri}) = \lambda_r, \lambda_r \rightarrow 0$ as $r \rightarrow \infty$

$$\sum_{r=1}^{\infty} \beta_r \delta_{ri} = \sum_{r=1}^{\infty} \beta^*_r \delta^*_r, \quad \beta^*_r := \lambda_r \beta_r, \quad \delta^*_r := \frac{\delta_{ri}}{\lambda_r}$$

“Sparseness”: If $|\beta_r| \leq D, D < \infty$ for all $r = 1, 2, \ldots$, for any $\epsilon > 0$ there exists a $k_\epsilon \in \mathbb{N}$ such that $\sum_{r>k_\epsilon} |\beta^*_r| \leq \epsilon$
Standard approach: Truncation

For a prespecified $k \in \mathbb{N}$ estimates $\hat{\beta}_r$ are determined from

$$Y_i \approx \sum_{r=1}^{k} \beta_r \delta_{ri} + \epsilon_i,$$

where $\hat{\delta}_{ri} = \langle X_i, \hat{\gamma}_r \rangle$ are determined from estimated functional principal components; $k$ - smoothing parameter

Then $\hat{\beta} = \sum_{r=1}^{k} \hat{\beta}_r \hat{\gamma}_r$


- Cai and Hall (2006) derive optimal rates of convergence of $|\int_I \beta(t)x(t) - \int_I \hat{\beta}(t)x(t)dt|$ for a pre-specified, fixed function $x$.

- Hall and Horowitz (2007) provide optimal convergence rates for the $L^2$-distance $\|\beta - \hat{\beta}\|$. 
The smoothing spline approach

We assume that the functions $X_i$ are observed at $p$ equidistant points $t_1, \ldots, t_p \in I$, $I = [0, 1]$. Let $\tilde{Y}_i := Y_i - \bar{Y}$, $\tilde{X}_i = X_i - \bar{X}$

Smoothing splines estimator of $\beta$: For some $m = 1, 2, \ldots$ and a smoothing parameter $\rho > 0$ an estimate $\hat{\beta}$ of $\beta$ is determined by minimizing

$$
\frac{1}{n} \sum_{i=1}^{n} \left( \tilde{Y}_i - \frac{1}{p} \sum_{j=1}^{p} b(t_j) \tilde{X}_i(t_j) \right)^2 + \rho \left( \frac{1}{p} \sum_{j=1}^{p} \pi_b(t_j)^2 + \int_{0}^{1} b^{(m)}(t)^2 dt \right)
$$

over all functions $b$ in the Sobolev space $W^{m, 2}(I) \subset L^2(I)$, where $\pi_b(t) = \sum_{l=1}^{m} \psi_{b,l} t^{l-1}$ with

$$
\sum_{j=1}^{p} (b(t_j) - \pi_b(t_j))^2 = \min_{\psi_1, \ldots, \psi_m} \sum_{j=1}^{p} (b(t_j) - \sum_{l=1}^{m} \psi_l t^{l-1})^2.
$$

$\pi_b$ denotes the best possible approximation of $(b(t_1), \ldots, b(t_p))$ by a polynomial of degree $m - 1$. 

Any solution \( \hat{\beta} \) has to be an element of the space \( NS^m(t_1, \ldots, t_p) \) of *natural splines* of order \( 2m \) with knots at \( t_1, \ldots, t_p \).

\( NS^m(t_1, \ldots, t_p) \) is a \( p \)-dimensional linear space of functions with \( v^{(m)} \in L^2(I) \).

There exists a canonical one-to-one mapping between \( \mathbb{R}^p \) and the space \( NS^m(t_1, \ldots, t_p) \) in the following way. For any vector \( \mathbf{w} = (w_1, \ldots, w_p)^\top \in \mathbb{R}^p \), there exists a unique natural spline interpolant \( s_{\mathbf{w}} \) with \( s_{\mathbf{w}}(t_j) = w_j, j = 1, \ldots, p \). With \( \mathbf{A} \) denoting the \( p \times p \) matrix with elements \( a_i(t_j) \), \( s_{\mathbf{w}} \) is given by

\[
s_{\mathbf{w}}(t) = \mathbf{a}(t)^\top (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{w}.
\]

Important property:

\[
\int_0^1 s_{\mathbf{w}}^{(m)}(t)^2 \, dt \leq \int_0^1 f^{(m)}(t)^2 \, dt \text{ for any other function } f \in \mathcal{W}^{m,2}(I)
\]

with \( f(t_j) = w_j, j = 1, \ldots, p \).
\[ \Rightarrow \hat{\beta} = (\hat{\beta}(t_1), \ldots, \hat{\beta}(t_p))^\top \in \mathbb{R}^p \text{ minimizes} \]

\[ \frac{1}{n} \sum_{i=1}^{n} \left( \tilde{Y}_i - \frac{1}{p} \sum_{j=1}^{p} b(t_j) \tilde{X}_i(t_j) \right)^2 + \rho \left( \frac{1}{p} \sum_{j=1}^{p} \pi_b(t_j)^2 + \int_{0}^{1} s_b^{(m)}(t)^2 dt \right), \]

with respect to all vectors \( b = (b(t_1), \ldots, b(t_p))^\top \in \mathbb{R}^p \).

**Matrix Notation:**

- \( Y = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^\top, \ X_i = (\tilde{X}_i(t_1), \ldots, \tilde{X}_i(t_p))^\top \) for all \( i = 1, \ldots, n \), \( \beta = (\beta(t_1), \ldots, \beta(t_p))^\top \) and \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top \)
- \( X: n \times p \) matrix with general term \( \tilde{X}_i(t_j) \) for all \( i = 1, \ldots, n, j = 1, \ldots, p \).
- \( P_m: p \times p \) projection matrix projecting into the \( m \)-dimensional linear space \( E_m := \{ w = (w_1, \ldots, w_p)^\top \in \mathbb{R}^p | w_j = \sum_{l=1}^{m} \theta_l t_j^{l-1}, j = 1, \ldots, p \} \) of all (discretized) polynomials of degree \( m - 1 \).
we have \( \int_0^1 s_b^{(m)}(t)^2 dt = b^\top B_m^* b \)

\[ B_m^* = A (A^\top A)^{-1} \left[ \int_0^1 a^{(m)}(t)a^{(m)}(t)^\top dt \right] (A^\top A)^{-1} A^\top \] is a \( p \times p \) matrix.

With \( B_m := P_m + \rho B_m^* \), spline minimization is equivalent to solving

\[
\min_{B \in \mathbb{R}^p} \left\{ \frac{1}{n} \left\| Y - \frac{1}{p} X b \right\|^2 + \frac{\rho}{p} b^\top B_m b \right\},
\]

where \( \| . \| \) stands here for the usual Euclidean norm, and

\[
\hat{\beta} = \frac{1}{np} \left( \frac{1}{np^2} X^\top X + \frac{\rho}{p} B_m \right)^{-1} X^\top Y = \frac{1}{n} \left( \frac{1}{np} X^\top X + \rho B_m \right)^{-1} X^\top Y.
\]

\[ \hat{\beta} = s\hat{\beta} \] constitutes our final estimator of \( \beta \),

\[ \hat{\beta}_0 = \overline{Y} - \langle \hat{\beta}, \overline{X} \rangle \] estimates the intercept \( \beta_0 \).

All eigenvalues of the matrix \( B_m \) are strictly positive which ensures existence as well as uniqueness of a solution.
Asymptotic Theory

$L^2(I)$ will be endowed with the inner product
\[ \langle f, g \rangle = \int_I f(t)g(t)\,dt \]
and its associated norm $\|\cdot\|$.

Asymptotic performance of the estimator is evaluated with respect to its prediction error.

We consider $\|\hat{\beta} - \beta\|_\Gamma$, where $\|\cdot\|_\Gamma$ is the $L^2$ semi-norm defined by
\[ \|u\|_\Gamma^2 = \langle \Gamma u, u \rangle \quad u \in L^2(I), \]
Here, $\Gamma$ is the covariance operator of $X$ given by
\[ \Gamma u = \mathbb{E} \left( \langle X - \mathbb{E}(X), u \rangle X - \mathbb{E}(X) \right), \quad u \in L^2(I). \]

Note: If $X_{n+1}$ is a random function possessing the same distribution as $X_i$, but independent of the sample $X_1, \ldots, X_n$, then
\[
\mathbb{E} \left( \left( \hat{\beta}_0 + \int_I \hat{\beta}(t)X_{n+1}(t)\,dt - \beta_0 - \int_I \beta(t)X_{n+1}(t)\,dt \right)^2 \Bigg| \hat{\beta}_0, \hat{\beta} \right) \\
= \|\hat{\beta} - \beta\|_\Gamma^2 + O_P(n^{-1}).
\]
Assumptions

(A.1) $\beta$ is $m$ times differentiable and $\beta^{(m)}$ belongs to $L^2(I)$.

(A.2) For every $\delta > 0$, there exist constants $C_3, C_4 < +\infty$ such that, $\mathbb{P}(|X_i(t)|^2 \leq C_3) \geq 1 - \delta$ and

$$
\mathbb{P} \left( |X_i(t) - X_i(s)| \leq C_4 |t - s|^\kappa, \ t, s \in I \right) \geq 1 - \delta
$$

(A.3) For some constant $C_5 < \infty$ and all $k \in \mathbb{N}^\star$ there is a $k$-dimensional linear subspace $\mathcal{L}_k$ of $L^2(I)$ with

$$
\mathbb{E} \left( \inf_{f \in \mathcal{L}_k} \sup_t |X(t) - f(t)|^2 \right) \leq C_5 k^{-2q}.
$$

(A.4) There exists a constant $C_7 < \infty$ such that

$$
\text{Var} \left( \frac{1}{n} \sum_i \langle X_i - \mathbb{E}(X), \zeta_r \rangle \langle X_i - \mathbb{E}(X), \zeta_s \rangle \right)
\leq \frac{C_7}{n} \mathbb{E}(\langle X - \mathbb{E}(X), \zeta_r \rangle^2)\mathbb{E}(\langle X - \mathbb{E}(X), \zeta_s \rangle^2)
$$

holds for all $n$ and all $r, s = 1, 2, \ldots$, $\|\overline{X} - \mathbb{E}(X)\|^2 = O_P(n^{-1})$. 
Lemma

For some $q_1 = 0, 1, 2, \ldots$ and $0 \leq q_2 \leq 1$ assume that $X_i$ is almost surely $q_1$-times continuously differentiable and that there exists some $C_6 < \infty$ such that

$$\mathbb{E} \left( \sup_{|t-s| \leq d} \left| X_i^{(q_1)}(t) - X_i^{(q_1)}(s) \right|^2 \right) \leq C_6 d^{-2q_2}$$

holds for all $d > 0$. There then exists a constant $C_7 < \infty$, only depending on $q_1, q_2$, such that for all $k = 1, 2, \ldots$

$$\mathbb{E} \left( \inf_{f \in \mathcal{P}_k} \sup_t |X_i(t) - f(t)|^2 \right) \leq C_7 C_6 k^{-2(q_1+q_2)},$$

where $\mathcal{P}_k$ denotes the space of all polynomials of order $k$ on $I$. 
Theorem

Under (A.1) - (A.4) as well as $np^{-2\kappa} = O(1)$, $\rho \to 0$, $1/(n\rho) \to 0$ as $n, p \to \infty$:

$$\left\| \hat{\beta} - \beta \right\|_\Gamma^2 = O_P \left( \rho + (n\rho^{2m+2q+1})^{-1} + n^{-\left(2q+1\right)/2} \right),$$

Out-of-sample predictions:

If $q \geq 1/2$ and $\rho \sim n^{-\left(2m+2q\right)/\left(2m+2q+1\right)}$ then generally

$$\mathbb{E} \left[ \left( \hat{\beta}_0 + \int\hat{\beta}(t)X_{n+1}(t)dt - \beta_0 - \int_0^1 \beta(t)X_{n+1}(t)dt \right)^2 \mid \hat{\beta}_0, \hat{\beta} \right] = \left\| \hat{\beta} - \beta \right\|_\Gamma^2 + O_P(n^{-1}) = O_P(n^{-\left(2m+2q+1\right)/\left(2m+2q+2\right)}).$$
Optimality of the convergence rate

- $\mathcal{P}_{q,C}$ - space of all probability distributions on $L^2([0, 1])$ with $\mathbb{E}(X_i) = 0$, $\sup_t |X_i(t)| \leq C$, and $\sum_{j=k+1}^{\infty} \lambda_j \leq Ck^{-2q}$ for all sufficiently large $k$.
- $\mathcal{C}_{m,D}$ - space of all $m$-times continuously differentiable functions $\beta$ with $\int_0^1 (\beta^{(j)}(t))^2 dt \leq C$ for all $j = 0, 1, \ldots, m$.
- For given $\beta \in \mathcal{C}_{m,D}$, probability distribution $P \in \mathcal{P}_{q,C}$ and i.i.d. random functions $X_1, \ldots, X_n$, $X_i \sim P$, let $\hat{a}(\beta, P)$ denote an arbitrary estimator of $\beta$ based on corresponding data $(Y_i, X_i)$, $i = 1, \ldots, n$.

Proposition

Let $c_n$ denote an arbitrary sequence of positive numbers with $c_n \to 0$ as $n \to \infty$, and let $2q = 1, 3, 5, \ldots$. Then

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}_{q,C}} \sup_{\beta \in \mathcal{C}_{m,D}} \inf_{\hat{a}(\beta, P)} \mathbb{P}\left( \|\beta - \hat{a}(\beta, P)\|_\Gamma \geq c_n \cdot n^{-(2m+2q+1)/(2m+2q+2)} \right) = 1$$
Prediction for fixed function $x$

**Theoretical Results:** Cai and Hall (2006)

- Let $\lambda_1 \geq \lambda_2 \geq \ldots$ denote the eigenvalues of $\Gamma$, and let $\gamma_1, \gamma_2, \ldots$ be a orthonormal basis of corresponding eigenfunctions.

- $X_i(t) = \sum_r x_{i,r} \gamma_r(t)$, $E(x_{i,r}^2) = \lambda_r$, $\text{cov}(x_{ir}, x_{is}) = 0$, $r \neq s$

- Decompositions: $\beta = \sum_r \beta_r \gamma_r(t)$, $x = \sum_r x_r \gamma_r(t)$

- Estimator $\hat{\beta} = \sum_{r=1}^{L} \hat{\beta}_r\hat{\gamma}_r$, $L$ - smoothing parameter

**Rates of convergence (Cai and Hall):** It is assumed that there exist some $\nu \in \mathbb{R}$ and $0 < D_0 < \infty$ such that for all $r = 1, 2, \ldots$

\[
D_0^{-1} r^\nu \leq \frac{x_r^2}{\lambda_r} \leq D_0 r^\nu
\]

- Parametric rates of convergence if $\nu \leq -1$:
  \[
  |\langle \hat{\beta}, x \rangle - \langle \beta, x \rangle|^2 = O(n^{-1}) \quad \text{[or } O(n^{-1} \log n)\text{]}
  \]

- Nonparametric rates of convergence for $\nu > -1$ which also depend on the speed of decrease of $|\beta_r|$ as $r \to \infty$
Prediction for a random function $X_{n+1}$

Gaussian distribution:

$$X_{n+1} = \sum_{r} x_{n+1,r} \gamma_r(t), \ x_{n+1,r} \sim N(0, \lambda_r), \ x_{n+1,r} \text{ independent of } x_{n+1,s}$$

$$\Rightarrow \mathbb{P} \left( D_0^{-1} r^\nu \leq \frac{x_{n+1,r}^2}{\lambda_r} \leq D_0 r^\nu \text{ for all } r = 1, 2, \ldots \right) = 0 \quad \text{if } \nu \leq 0$$

- Assumption (A.3): $\sum_{r=k+1}^\infty \lambda_r = O(k^{-2q})$
  $$\Rightarrow \text{One may assume } \lambda_r = O(r^{-2q-1})$$

- Assume that $\gamma_1, \gamma_2, \ldots$ define an appropriate basis for approximating smooth functions and that
  $$\inf_{f \in \text{span}\{\gamma_1, \ldots, \gamma_k\}} \|\beta - f\|^2 = \sum_{r=k+1}^\infty \beta_r^2 = O(k^{-2m}) \text{ as well as } \beta_r^2 = O(r^{-2m-1}).$$

For any $\nu > 0$ the results of Cai and Hall (2006) then imply

$$\langle \hat{\beta} - \beta, X_{n+1} \rangle^2 = O_P\left(n^{-\left(2m+2q+1-2\nu\right)/(2m+2q+2)}\right),$$
$L^2$-distances $\|\hat{\beta} - \beta\|$

A statistically very different problem consists in an optimal estimation of $\beta$ by $\hat{\beta}$ with respect to the usual $L^2$-norm.

- Under (A.1)-(A.4): Components of $\beta$ which are in $\ker(\Gamma)$ are not identifiable; generally $\|\hat{\beta} - \beta\|^2 = O_P(1)$

- Any results on rates of convergence of $\hat{\beta}$ will heavily depend on how well $X$ and $\beta$ “fit” together.

- Oversmoothing: an estimator minimizing $\|\hat{\beta} - \beta\|^2$ will have to rely on $\rho \gg n^{-(2m+2q+1)/(2m+2q+2)}$

$$\rho \sim n^{-(2m+2q+1)/(2m+2q+2)} \Rightarrow \|\hat{\beta} - \mathbb{E}_e(\hat{\beta})\|^2 = O_P(n^{-1/(2m+2q+2)})$$

Hall and Horowitz (2007): $\hat{\beta} = \sum_{r=1}^{k} \hat{\beta}_r \hat{\gamma}_r$

If $\lambda_r - \lambda_{r+1} \geq C \cdot r^{-2q-2}$, $\lambda_r \geq C^* r^{-2q-1}$, $|\beta_r| \leq C^{**} r^{-\alpha}$, $k \sim n^{1/(2q+2\alpha+1)}$, then under some additional regularity conditions

$$\|\hat{\beta} - \beta\|^2 = O_P \left(n^{-\frac{2\alpha-1}{2q+2\alpha+1}}\right)$$

More general results: Cardot and Johannes (2008), Johannes (2008)
Choice of the smoothing parameter

Similar to ordinary nonparametric regression an optimal smoothing parameter $\rho$ may be estimated by

- leave-one-out cross-validation
- Generalized cross-validation (GCV)

The GCV criterion takes the form

$$GCV_m(\rho) = \frac{1}{n} \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \frac{1}{(1 - n^{-1} Tr(H(\rho)))^2},$$

where

$$H(\rho) := (np)^{-1}X \left( \frac{1}{np^2} X^\tau X + \frac{\rho}{p} A_m \right)^{-1} X^\tau$$
Theoretical justification of GCV

\[
ASE_m(\rho) := \frac{1}{n} \sum_i \left[ \langle X_i - \bar{X}, \beta \rangle - \frac{1}{p} \sum_j (X_i(t_j) - \bar{X}(t_j)) \hat{\beta}_{\rho,m}(t_j) \right]^2
\]

**Proposition 2:** Assume (A.1)-(A.3), \( np^{-2\kappa} = O(1) \), 
\( \mathbb{E}(\exp(\delta \epsilon_i^2)) < \infty \) for some \( \delta > 0 \).

- \( m \) fixed; \( \hat{\rho} \) minimizer of GCV(\( \rho \)) over \( \rho \in \left[ n^{-2m+\delta}, \infty \right), \delta > 0 \)

\[
|ASE_m(\hat{\rho}) - ASE_m(\rho_{opt})| = O_P(n^{-\frac{1}{2}} ASE_m(\rho_{opt})^{\frac{1}{2}})
\]

- \( \hat{m}, \hat{\rho} \) minimizers of GCV over \( \rho \in \left[ n^{-2m+\delta}, \infty \right), \delta > 0 \), and \( m = 1, \ldots, M_n, M_n \leq n/2 \)

\[
|ASE_{\hat{m}}(\hat{\rho}) - ASE_{m_{opt}}(\rho_{opt})| = O_P(n^{-\frac{1}{2}} ASE_{m_{opt}}(\rho_{opt})^{\frac{1}{2}} \log M_n),
\]
The case of a noisy covariate

- Errors-in-variable problem: In practice the true functional values $X_i(t_j)$ are often not directly observable. There only exist discrete observations contaminated with several kind of errors.

(More realistic) observational model:

$$W_i(t_j) = X_i(t_j) + \delta_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, p,$$

where $(\delta_{ij})_{i=1,\ldots,n,j=1,\ldots,p}$ is a sequence of independent real random variables with $\mathbb{E}_\epsilon(\delta_{ij}) = 0$, $\mathbb{E}_\epsilon(\delta_{ij}^2) = \sigma_\delta^2$, as well as $\mathbb{E}_\epsilon(\delta_{ij}^4) \leq C_9 < \infty$.

The goal is then to introduce some corrections in the estimator $\hat{\beta}$ in order to account for this additional error.
Construction of an estimator

Under our assumptions we obtain the representation

\[ \frac{1}{np^2} W^\tau W = \frac{1}{np^2} X^\tau X + \frac{\sigma_\delta^2}{p^2} I_p + R, \]

where \( R \) is a \( p \times p \) matrix such that its largest singular value is of order \( O_P \left( \frac{1}{n^{1/2}p} \right) \)

\[ \Rightarrow \frac{1}{np^2} W^\tau W - \frac{\sigma_\delta^2}{p^2} I_p \text{ is used to approximate } \frac{1}{np^2} X^\tau X \]

Estimation of the error variance \( \sigma_\delta^2 \):

\[ \hat{\sigma}_\delta^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{6(p-2)} \sum_{j=2}^{p-1} [W_i(t_{j-1}) - W_i(t_j) + W_i(t_{j+1}) - W_i(t_j)]^2. \]

Corrected estimator:

\[ \hat{\beta}_W = \frac{1}{np} \left( \frac{1}{np^2} W^\tau W - \frac{\hat{\sigma}_\delta^2}{p^2} I_p + \frac{\rho}{p} A_m \right)^{-1} W^\tau Y \]
Theoretical results

Additional assumption:

\[(A.5)\] For every \(\delta > 0\) there exists a constant \(C_\beta < \infty\) such that

\[
\frac{1}{\rho^{1/2}} \left\| \frac{1}{np} \mathbf{X}^\tau \mathbf{X} \beta \right\| > C_\beta,
\]

holds with probability larger or equal to \(1 - \delta\).

Assumption (A.5) essentially means that \(\beta\) does not belong to the kernel of the covariance operator \(\Gamma\).

Theorem

Assume (A.1)- (A.5) as well as \(np^{-2\kappa} = O(1), \rho \to 0, 1/(n\rho) \to 0\) as \(n, p \to \infty\). Then

\[
\left\| \hat{\beta}_W - \hat{\beta} \right\|^2_{\Gamma} = O_P \left( \frac{1}{np\rho} + \frac{1}{n} + n^{-(2q+1)/2} \right).
\]
Rates of convergence

Upper bound for the rate of convergence of $\hat{\beta}_W$ for $q \geq 1/2$:

$$\left\| \hat{\beta}_W - \beta \right\|_\Gamma^2 = O_P \left( \rho + \left( n\rho^{\frac{1}{2m+2q+1}} \right)^{-1} + \frac{1}{np\rho} \right).$$

- the use of $\hat{\beta}_W$ results in the addition of the extra term $1/(np\rho)$ in the rate of convergence.
- For a choice of $\rho \sim n^{-\frac{(2m+2q+1)}{(2m+2q+2)}}$ we have $1/(np\rho) \sim n^{-1/(2m+2q+2)}/p$. This term is of order $n^{-\frac{(2m+2q+1)}{(2m+2q+2)}}$ for $p \sim n^{\frac{(2m+2q-1)}{(2m+2q+2)}}$.
- In the errors-in-variables context we reach the same rate of convergence as for correctly observed $X_i$, provided that $p$ is large enough compared to $n$, i.e. $p \geq C_p \max(n^{1/2\kappa}, n^{\frac{(2m+2q-1)}{(2m+2q+2)}})$ for some positive constant $C_p$. 
Application to ozone pollution forecasting

- **Data:** Observatoire Régional de l’Air en Midi-Pyrénées; measures of specific pollutants, as well as meteorological measures, are made each hour.

- Explanatory variable $X_i$: Curves representing repeated measurements of concentration in ozone during a day

- Response variable $Y_i$: Maximum of ozone to be measured the following day

- Each ozone concentration curve $X_i$ is measured in $p = 24$ discretized (equispaced) points corresponding to hourly measurements and the size of the sample is $n = 474$.

- After centering the $X_i$'s, we modelize the link between $Y_i$ and $X_i$ by a functional linear regression model.

**Goal:** For a curve $X_{n+1}$ predict $Y_{n+1}$ via

$$\hat{Y}_{n+1} = \int \hat{\beta}(t)X_{n+1}(t)dt,$$
Interval of prediction

- Assumption: $\epsilon_1, \ldots, \epsilon_{n+1}$ are i.i.d. with $\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)$.
- $\sigma^2_\epsilon$ is consistently estimated by the empirical variance

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \frac{1}{p} \sum_{j=1}^{p} \hat{\beta}(t_j) X_i(t_j) \right)^2.$$

- Our theoretical results imply asymptotic normality of $\frac{Y_{n+1} - \hat{Y}_{n+1}}{\hat{\sigma}_\epsilon}$.

Given $\tau \in ]0, 1[$, an asymptotic $(1 - \tau)$-prediction interval for $Y_{n+1}$ is given by

$$\left[ \hat{Y}_{n+1} - z_{1-\tau/2} \hat{\sigma}_\epsilon, \hat{Y}_{n+1} + z_{1-\tau/2} \hat{\sigma}_\epsilon \right],$$

where $z_{1-\tau/2}$ is the quantile of order $1 - \tau/2$ of the $\mathcal{N}(0,1)$ distribution.
Empirical results

We split the initial sample into two sub-samples:

- a learning sample \((X_i, Y_i)_{i=1,\ldots,n_l}, (n_l = 300)\), used to determine the estimator \(\hat{\beta}\),
- a test sample \((X_i, Y_i)_{i=n_l+1,\ldots,n_l+n_t}, (n_t = 174)\) used to evaluate the quality of the estimation.

- The procedure is applied using \(m = 2\) (cubic smoothing splines)
- The two estimators \(\hat{\beta}\) and \(\hat{\beta}_W\) are considered
- Smoothing parameters \(\rho\) are selected by generalized cross-validation

Resulting prediction errors \(EQM\left(\hat{\beta}\right) = \frac{1}{n_t} \sum_{i=n_l+1}^{n_l+n_t} (Y_i - \hat{Y}_i)^2:\n
- \(EQM\left(\hat{\beta}\right) = 281.97\) for \(\hat{\beta}\)
- \(EQM\left(\hat{\beta}_W\right) = 270.13\) for \(\hat{\beta}_W\)