Defense of the PhD thesis:

On reproducing kernel methods in functional statistics

Beatriz Bueno-Larraz

Advisors:

José R. Berrendero and Antonio Cuevas

departamento de matemáticas

UNIVERSIDAD AUTONOMA DE MADRID

1. Introduction: FDA and RKHS's

2. Functional Mahalanobis distance

   Berrendero, J. R., Bueno-Larraz, B., Cuevas, A. (2018). On Mahalanobis distance in functional settings. *arXiv:1803.06550*.

3. Variable selection in functional regression

   Berrendero, J. R., Bueno-Larraz, B., Cuevas, A. (2018). An RKHS model for variable selection in functional linear regression. *Journal of Multivariate Analysis (to appear)*

   Bueno-Larraz, B. and Klepsch, J. (2018). Variable selection for the prediction of C[0,1]-valued autoregressive processes using RKHS. *Technometrics (to appear)*

4. Functional logistic regression

Introduction: FDA and RKHS's

Historical development of statistical inference

Summary table from Cuevas (2014):
(n sample size, d number of variables)

Statistical theory	**Data**	**Parameters**	**Start**
Classical inference	$\mathbb{R}$	$\mathbb{R}$	1920s
Multivariate analysis	$\mathbb{R}^d$ $(n \gg d)$	$\mathbb{R}^k$ $(n \gg k)$	1940s
Nonparametrics	$\mathbb{R}^d$ $(n \gg d)$	$\mathcal{F}$	1960s
Functional data analysis	$\mathcal{F}$	$\mathbb{R}^k$ or $\mathcal{F}$	1990s
High dimension	$\mathbb{R}^d$ $(n < d)$	$\mathbb{R}^k$	2000s
Object oriented DA	Structure	$\mathbb{R}^k$ or $\mathcal{F}$	2010s

Our setting

Our sample is made of trajectories $x \in L^2[0,1]$, drawn from a second order process $X(s)$ with

- mean function $m(s) = \mathbb{E}[X(s)]$,
- covariance function $K(s,t) = \mathrm{cov}(X(s), X(t))$.

Some of the main difficulties:

- high correlation in high-dimensional vectors,
- lack of a natural translation-invariant measure,
- no obvious complete order among curves, ...

Definitions of RKHS

Def. 1: $r : [0,1] \times [0,1] \to \mathbb{R}$ and $\mathcal{H} \subset L^2[0,1]$ a Hilbert space.

1. $r(s,\cdot) \in \mathcal{H}, \forall s \in [0,1]$.
2. Reproducing property: $\langle f, r(s,.)\rangle = f(s), \forall f \in \mathcal{H}, \forall s \in [0,1]$.

Def. 2: $\mathcal{H}(K) = \{\mathcal{K}^{1/2} g,\ g \in L^2[0,1]\}$, where $\mathcal{K}g(t) = \int_0^1 K(t,s)g(s)\mathrm{d}s$.

$$\|f\|_K^2 \ = \ \|\mathcal{K}^{-1/2} f\|_2^2 \ = \ \sum_{j=1}^{\infty} \frac{\langle f, e_i\rangle_2^2}{\lambda_i}$$

Def. 3: Completion of:

$$\mathcal{H}_0(K) \ \equiv \ \left\{ f \ : f(s) = \sum_{j=1}^{n} a_j K(s, t_j),\ a_j \in \mathbb{R},\ t_j \in [0,1],\ n \in \mathbb{N} \right\}$$

Functional Mahalanobis distance

Direct functional extension

Let Σ be the (non-singular) covariance matrix of the d-dimensional random vector X, the (square) Mahalanobis distance between $x \in \mathbb{R}^d$ and $m \in \mathbb{R}^d$ is

$$M^2(x, m) = (x - m)'\Sigma^{-1}(x - m).$$

Its most popular applications are:

- Supervised classification
- Outlier detection
- Multivariate depth measures, with depth function $(1+M(x, m))^{-1}$.
- Hypothesis testing,...

The counterpart of Σ^{-1} would be $\mathcal{K}^{-1}$, but $\mathcal{K}$ is not invertible (there is no linear continuous $\mathcal{K}^{-1}$ s.t. $\mathcal{K}^{-1}\mathcal{K} = \mathcal{K}^{-1}\mathcal{K} = \mathrm{Id}$).

Direct functional extension

Let Σ be the (non-singular) covariance matrix of the d-dimensional random vector X, the (square) Mahalanobis distance between $x \in \mathbb{R}^d$ and $m \in \mathbb{R}^d$ is

$$M^2(x,m) = (x-m)'\Sigma^{-1}(x-m).$$

Its most popular applications are:

- Supervised classification
- Outlier detection
- Multivariate depth measures, with depth function $(1+M(x,m))^{-1}$.
- Hypothesis testing,...

The counterpart of Σ^{-1} would be $\mathcal{K}^{-1}$, but $\mathcal{K}$ is not invertible (there is no linear continuous $\mathcal{K}^{-1}$ s.t. $\mathcal{K}^{-1}\mathcal{K} = \mathcal{K}^{-1}\mathcal{K} = \mathrm{Id}$).

Main contributions

- RKHS's give a natural framework to extend Mahalanobis distance: the new distance is fully mathematically grounded.
- It is a true metric defined for any pair of functions in $L^2[0, 1]$.
- It depends on a single, real, easy to interpret smoothing parameter whose choice is not critical.
- It shares some interesting properties with the original distance.

Spectral decomposition

Given λ_i, e_i the eigenvalues and eigenvectors of Σ,

$$M^2(x,m) = \sum_{i=1}^{\mathbf{d}} \frac{((x-m)'e_i)^2}{\lambda_i}.$$

Then the naive functional extension would be

$$M^2(x,m) = \sum_{i=1}^{\infty} \frac{\langle x-m, e_i\rangle_2^2}{\lambda_i},$$

for λ_i, e_i the eigenvalues and eigenfunctions of $\mathcal{K}$.

Problem: when x is the trajectory of a Gaussian process, this series diverges with probability 1.

Existing proposals Galeano et al. (2015), Ghiglietti et al. (2017).

Spectral decomposition

Given λ_i, e_i the eigenvalues and eigenvectors of Σ,

$$M^2(x,m) = \sum_{i=1}^{\mathbf{d}} \frac{((x-m)'e_i)^2}{\lambda_i}.$$

Then the naive functional extension would be

$$M^2(x,m) = \sum_{i=1}^{\infty} \frac{\langle x-m, e_i\rangle_2^2}{\lambda_i},$$

for λ_i, e_i the eigenvalues and eigenfunctions of $\mathcal{K}$.

Problem: when x is the trajectory of a Gaussian process, this series diverges with probability 1.

Existing proposals Galeano et al. (2015), Ghiglietti et al. (2017).

Finite dimensional RKHS

If $T = \{1, 2, \ldots, d\}$ then $K : T \times T \to \mathbb{R}$ is summarized by the covariance matrix $\Sigma_{i,j} = K(i,j)$, and then

$$\mathcal{H}(\Sigma) = \{x = \Sigma^{1/2}a,\ a \in \mathbb{R}^d\}.$$

In this case, for $x \in \mathcal{H}(\Sigma)$,

$$\|x\|_K^2 \ = \ (\Sigma^{-1/2}x)'(\Sigma^{-1/2}x) \ = \ x'\Sigma^{-1}x \ = \ M^2(x,0).$$

That is, the square norm of x in $\mathcal{H}(K)$ coincides with the Mahalanobis distance.

Trouble with the extension to functional data

Then, one would like to define $M(x, m) = \|x - m\|_K$.

But it is not possible since, when x is the trajectory of a Gaussian process, $x \notin \mathcal{H}(K)$ with probability 1. (Lukić and Beder (2001))

Idea and definition

It seems natural to define the Mahalanobis distance replacing the trajectory $x(t)$ by the "closest" function in $\mathcal{H}(K)$.

But since $\mathcal{H}(K)$ is not closed in $L^2[0,1]$, we need to penalize,

$$x_\alpha = \underset{f \in \mathcal{H}(K)}{\arg\min} \left(\|x - f\|_2^2 + \alpha \|f\|_K^2 \right).$$

Then we define,

$$M_\alpha(x, m) = \|x_\alpha - m_\alpha\|_K$$

Explicit expression

The minimization problem for x_α has an explicit solution,

$$x_\alpha = (\mathcal{K} + \alpha \mathbb{I})^{-1} \mathcal{K} x.$$

Then the distance can be rewritten as

$$\|x_\alpha - m_\alpha\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)^2} \langle x - m, e_j \rangle_2^2$$

which is a true metric.

Properties

This distance M_α is invariant for isometries in $L^2[0,1]$. (Th. 2.4)

We know its distribution for Gaussian processes: (Th. 2.5)

$$M_\alpha(x,m)^2 \sim \sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j+\alpha)^2} Y_j,$$

where Y_j are independent χ_1^2.

It is continuous with respect to α: (Prop. 2.8)

$$\|x_{\alpha_j}\|_K \to \|x_\alpha\|_K \text{ a.s. as } \alpha_j \to \alpha \text{ (all positive).}$$

Properties

This distance M_α is invariant for isometries in $L^2[0,1]$. (Th. 2.4)

We know its distribution for Gaussian processes: (Th. 2.5)

$$M_\alpha(x,m)^2 \sim \sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j+\alpha)^2} Y_j,$$

where Y_j are independent χ_1^2.

It is continuous with respect to α: (Prop. 2.8)

$$\|x_{\alpha_j}\|_K \to \|x_\alpha\|_K \text{ a.s. as } \alpha_j \to \alpha \text{ (all positive).}$$

Properties

This distance M_α is invariant for isometries in $L^2[0,1]$. (Th. 2.4)

We know its distribution for Gaussian processes: (Th. 2.5)

$$M_\alpha(x,m)^2 \sim \sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j+\alpha)^2} Y_j,$$

where Y_j are independent χ_1^2.

It is continuous with respect to α: (Prop. 2.8)

$$\|x_{\alpha_j}\|_K \to \|x_\alpha\|_K \text{ a.s. as } \alpha_j \to \alpha \text{ (all positive).}$$

Asymptotic properties

If $\mathbb{E}\|X\|_2^2 < \infty$, the estimator $\widehat{M}_\alpha(x, \bar{x})$ (using $\widehat{\lambda}_j, \widehat{e}_j$ from the data) converges a.s. to $M_\alpha(x, m)$. (Th. 2.10)

Moreover, given a sample of size n, (Th. 2.13a)

$$\sqrt{n}\,\widehat{M}_\alpha(\bar{x}, m) \xrightarrow{d} \left(\sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j + \alpha)^2} Y_j \right)^{1/2},$$

where Y_j are independent χ_1^2.

Asymptotic properties

If $\mathbb{E}\|X\|_2^2 < \infty$, the estimator $\widehat{M}_\alpha(x, \bar{x})$ (using $\widehat{\lambda}_j, \widehat{e}_j$ from the data) converges a.s. to $M_\alpha(x, m)$. (Th. 2.10)

Moreover, given a sample of size n, (Th. 2.13a)

$$\sqrt{n}\, \widehat{M}_\alpha(\bar{x}, m) \xrightarrow{d} \left(\sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j + \alpha)^2} Y_j \right)^{1/2},$$

where Y_j are independent χ_1^2.

Variable selection in regression

Functional regression problem

A typical functional regression model has the form:

$$y_i = g(x_i) + \varepsilon_i,$$

where x_i are trajectories of a process X, y_i are the responses and ε_i are errors, which are often assumed independent of X.

Different models can be constructed depending on g.
The most typical model for scalar response is

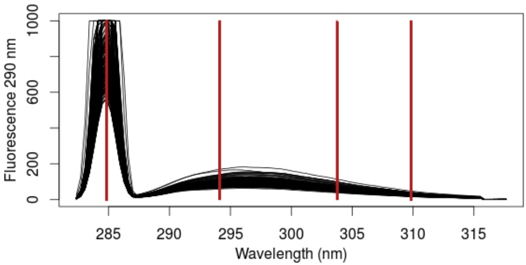
$$y_i = \alpha_0 + \langle x_i, \beta \rangle_2 + \varepsilon_i, \qquad i = 1, \ldots, n$$

Variable selection

It is a particular case of dimension reduction (PCA, PLS, LASSO,...)

In this setting of FDA,

$$\{X(t), t \in [0,1]\} \to (X(t_1), \ldots, X(t_p))\,.$$



Sugar example (Aneiros and Vieu (2014))

Main contributions

A functional model to integrate variable selection in functional regression via RKHS's: asymptotic consistency results.

We derive a consistent estimator for the number of variables.

We arrive to a well-known estimator:

- Multivariate Forward Selection.
- Sample kriging (Cambanis (1985)).
- Functional least squares (Ji and Müller (2016), McKeague and Sen (2010)),...

Regression model

One may be tempted to define, for $\beta \in \mathcal{H}(K)$:

$$y_i = \langle \beta, x_i \rangle_K + \varepsilon_i, \quad i = 1, \ldots, n.$$

Since then

$$y_i = \sum_{j=1}^{p} \beta_j \, x_i(t_j) + \varepsilon_i,$$

would be a particular case.

Problem: In general, $x_i \notin \mathcal{H}(K)$ with probability one.

Loève's isometry

$\mathcal{L}(X)$ is the closed linear span of $X(s) - m(s)$. (denoting $m(t) = \mathbb{E}[X(t)]$)

Loéve's Theorem

The RKHS $\mathcal{H}(K)$ is an isometric copy of $\mathcal{L}(X)$, since

$$\Psi_X \left(\sum_{i=1}^{p} a_i \big(X(t_i) - m(t_i) \big) \right) = \sum_{i=1}^{p} a_i K(t_i, \cdot), \qquad \forall a_i \in \mathbb{R}$$

defines a congruence (Loéve's Isometry).

We identify $\langle \beta, X \rangle_K \equiv \Psi_X^{-1}(\beta)$.

Then we propose the model:

$$Y = \Psi_X^{-1}(\beta) + \varepsilon$$

Minimizing the response error

When the slope function is of the form

$$\beta(\cdot) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot),$$

the regression model reduces to

$$y_i = \sum_{j=1}^{p^*} \beta_j \big(x_i(t_j^*) - m(t_j^*)\big) + \varepsilon_i.$$

A sensible approach is to choose $T_p = (t_1, \ldots, t_p)$ that minimizes

$$Q(T_p) \equiv \min_{(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p} \left\| Y - \sum_{j=1}^{p} \beta_j (X(t_j) - m(t_j)) \right\|_2^2.$$

We prove that $\arg\min_{T_p} Q(T_p) = \arg\max_{T_p} c_{T_p}' \Sigma_{T_p}^{-1} c_{T_p}$. (Prop. 3.1)

Minimizing the response error

When the slope function is of the form

$$\beta(\cdot) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot),$$

the regression model reduces to

$$y_i = \sum_{j=1}^{p^*} \beta_j \big(x_i(t_j^*) - m(t_j^*)\big) + \varepsilon_i.$$

A sensible approach is to choose $T_p = (t_1, \dots, t_p)$ that minimizes

$$Q(T_p) \;\equiv\; \min_{(\beta_1,\dots,\beta_p)\in\mathbb{R}^p} \Big\| Y - \sum_{j=1}^{p} \beta_j (X(t_j) - m(t_j)) \Big\|_2^2.$$

We prove that $\arg\min_{T_p} Q(T_p) = \arg\max_{T_p} c'_{T_p} \Sigma_{T_p}^{-1} c_{T_p}$. (Prop. 3.1)

Asymptotic results

Natural estimators: $\widehat{Q}_n(T_p) = \widehat{c}'_{T_p} \widehat{\Sigma}_{T_p}^{-1} \widehat{c}_{T_p}$ and $\widehat{T}_{p^*} = \arg\max_{T_{p^*}} \widehat{Q}_n(T_{p^*})$.

Whenever $\mathbb{E}[\sup_t X(t)^2] < \infty$, $\Sigma_{T_{p^*}}$ invertible and p^* known:

- $\widehat{T}_{p^*} \xrightarrow{a.s} T^*$ (only global maximum) and in quadratic mean. (Th. 3.8)
- $\widehat{Y}_{\widehat{T}_{p^*}} \xrightarrow{a.s} Y_{T^*}$, and if in addition $\mathbb{E}\left[\sup_t |X(t)|^{2+\eta}\right] < \infty$, then also in quadratic mean. (Th. 3.9)

If p^* is unknown:

We derive an estimator $\widehat{p} \xrightarrow{a.s} p^*$. (Th. 3.11)

For $p > p^*$: T^* are very close to p^* points of $\widehat{T}_p$. (Th. 3.12)

Asymptotic results

Natural estimators: $\widehat{Q}_n(T_p) = \widehat{c}'_{T_p}\widehat{\Sigma}^{-1}_{T_p}\widehat{c}_{T_p}$ and $\widehat{T}_{p^*} = \arg\max_{T_{p^*}} \widehat{Q}_n(T_{p^*})$.

Whenever $\mathbb{E}[\sup_t X(t)^2] < \infty$, $\Sigma_{T_{p^*}}$ invertible and p^* known:

- $\widehat{T}_{p^*} \overset{a.s}{\to} T^*$ (only global maximum) and in quadratic mean. (Th. 3.8)
- $\widehat{Y}_{\widehat{T}_{p^*}} \overset{a.s}{\to} Y_{T^*}$, and if in addition $\mathbb{E}\big[\sup_t |X(t)|^{2+\eta}\big] < \infty$, then also in quadratic mean. (Th. 3.9)

If p^* is unknown:

- We derive an estimator $\widehat{p} \overset{a.s}{\to} p^*$. (Th. 3.11)
- For $p > p^*$: T^* are very close to p^* points of $\widehat{T}_p$. (Th. 3.12)

Extension to functional response

The extended model is

$$Y(s) = \Psi_X^{-1}\big(\beta(s,\cdot)\big) + \varepsilon(s), \ \text{ for } \beta(s,\cdot) \in \mathcal{H}(K).$$

The sparse representation of β is extended as

$$\beta(s,\cdot) = \sum_{j=1}^{p} \beta_j(s) K(t_j,\cdot).$$

Then, the sparse model, pointwisely for $s \in [0,1]$ is

$$Y(s) = \sum_{j=1}^{p} \beta_j(s) X(t_j) + \varepsilon(s).$$

Main difficulties

- The optimization cannot be simply performed for every $Y(s)$ using

$$q_1(T_p\ ;\ \alpha_1,\ldots,\alpha_p)(s) \ = \ \left\| Y(s) - \sum_{j=1}^{p} \alpha_j(s)\big(X(t_j) - m(t_j)\big) \right\|^2.$$

Then, we minimize with respect to T_p in Θ_p,

$$Q_1(T_p) \ := \ \int_0^1 \min_{\alpha_j(s)\in\mathbb{R}} q_1(T_p\ ;\ \alpha_1,\ldots,\alpha_p)(s)\, \mathrm{d}s$$

To get uniform a.s. convergence of $\mathrm{cov}\big(X(s), Y(t)\big)$ it is required that $U(s) = \Psi_X^{-1}\big(\beta(s,\cdot)\big)$ fulfills $\mathbb{E}[\sup_s |U(s)|^2] < \infty$.

Main difficulties

- The optimization cannot be simply performed for every $Y(s)$ using

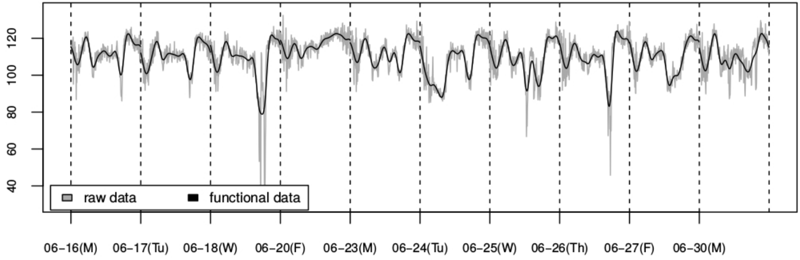
$$q_1(T_p\ ;\ \alpha_1,\ldots,\alpha_p)(s)\ =\ \Big\|Y(s)-\sum_{j=1}^{p}\alpha_j(s)\big(X(t_j)-m(t_j)\big)\Big\|^2.$$

Then, we minimize with respect to T_p in Θ_p,

$$Q_1(T_p)\ :=\ \int_0^1 \min_{\alpha_j(s)\in\mathbb{R}} q_1(T_p\ ;\ \alpha_1,\ldots,\alpha_p)(s)\ \mathrm{d}s$$

- To get uniform a.s. convergence of $\operatorname{cov}\big(X(s),Y(t)\big)$ it is required that $U(s)=\Psi_X^{-1}\big(\beta(s,\cdot)\big)$ fulfills $\mathbb{E}[\sup_s|U(s)|^2]<\infty$.

Main difficulties: convergence



In the functional time series context:

H1. The process $X_n(t)X_n(s)$ for $t, s \in [0, 1]$ is uniformly geometrically strong mixing.

H2. Cramer conditions (bounds on the moments of $X_n(t)X_n(s)$).

Functional logistic regression

Introduction to logistic regression

Logistic regression arises from biomedical problems, where one requires probabilities of being “healthy or sick”.

Then it is used for problems with categorical response ($Y \in \{0, 1\}$).

A classifier that gives the probability of belonging to the assigned class.

The theory for finite dimensional regressors is well developed, but the functional one is still quite new.

Main properties of logistic regression

- It is assumed that $\log\big(p(x)/(1-p(x))\big)$ is linear in $x \in \mathbb{R}^d$, where $p(x) = \mathbb{P}(Y{=}1|\,x)$, which leads to:

$$\mathbb{P}(Y{=}1|\,x) = \frac{1}{1 + \exp\{-\alpha_0 - \alpha' x\}}, \qquad \alpha_0 \in \mathbb{R} \text{ and } \alpha \in \mathbb{R}^d.$$

- It holds when X_0, X_1 are Gaussian with common covariance matrix.

- The standard functional extension for $x \in \mathcal{L}^2[0,1]$ is

$$\mathbb{P}(Y{=}1|\,x) = \frac{1}{1 + \exp\{-\beta_0 - \langle \beta, x\rangle_2\}}, \qquad \beta_0 \in \mathbb{R} \text{ and } \beta \in \mathcal{L}^2[0,1].$$

Main contributions

- The proposed model holds whenever X_0, X_1 are homoscedastic Gaussian processes:
$$\mathbb{P}(Y{=}1 \,|\, X{=}x) \;=\; \big(1 + \exp\{-\beta_0 - \Psi_x^{-1}(\beta)\}\big)^{-1}.$$
- We clarify the relationship with the standard L^2 approach,
$$\mathbb{P}(Y{=}1 \,|\, X{=}x) \;=\; \big(1 + \exp\{-\beta_0 - \langle \beta, x\rangle_2\}\big)^{-1}.$$
- We carefully analyze whether the ML estimator exists.

Conditional Gaussian distributions

Let $X_0(s), X_1(s)$ be Gaussian processes with continuous trajectories, continuous mean functions m_0, m_1 and continuous covariance function K (equal for both classes).

Let P_0 and P_1 be the probability measures on $C[0,1]$ (or $L^2[0,1]$) induced by the processes X_0, X_1 respectively.

(Theorem 4.1)

(a) if $m_0, m_1 \in \mathcal{H}(K)$, then $P_0 \sim P_1$ and the RKHS model holds with $\beta := m_1 - m_0$ and $\beta_0 := (\|m_0\|_K^2 - \|m_1\|_K^2)/2 - \log((1-p)/p)$.

(b) if $m_1 - m_0 \in \mathcal{K}(L^2) = \{\mathcal{K}(f) : f \in L^2[0,1]\}$, then $P_0 \sim P_1$ and L^2 model holds.

(c) if $m_1 - m_0 \notin \mathcal{K}(L^2)$, L^2 model is never recovered, but different situations are possible.

Conditional Gaussian distributions

Let $X_0(s), X_1(s)$ be Gaussian processes with continuous trajectories, continuous mean functions m_0, m_1 and continuous covariance function K (equal for both classes).

Let P_0 and P_1 be the probability measures on $C[0,1]$ (or $L^2[0,1]$) induced by the processes X_0, X_1 respectively.

(Theorem 4.1)

(a) if $m_0, m_1 \in \mathcal{H}(K)$, then $P_0 \sim P_1$ and the RKHS model holds with $\beta := m_1 - m_0$ and $\beta_0 := (\|m_0\|_K^2 - \|m_1\|_K^2)/2 - \log((1-p)/p)$.

(b) if $m_1 - m_0 \in \mathcal{K}(L^2) = \{\mathcal{K}(f) : f \in L^2[0,1]\}$, then $P_0 \sim P_1$ and L^2 model holds.

(c) if $m_1 - m_0 \notin \mathcal{K}(L^2)$, L^2 model is never recovered, but different situations are possible.

Conditional Gaussian distributions

Let $X_0(s), X_1(s)$ be Gaussian processes with continuous trajectories, continuous mean functions m_0, m_1 and continuous covariance function K (equal for both classes).

Let P_0 and P_1 be the probability measures on $C[0,1]$ (or $L^2[0,1]$) induced by the processes X_0, X_1 respectively.

(Theorem 4.1)

(a) if $m_0, m_1 \in \mathcal{H}(K)$, then $P_0 \sim P_1$ and the RKHS model holds with $\beta := m_1 - m_0$ and $\beta_0 := (\|m_0\|_K^2 - \|m_1\|_K^2)/2 - \log((1-p)/p)$.

(b) if $m_1 - m_0 \in \mathcal{K}(L^2) = \{\mathcal{K}(f) : f \in L^2[0,1]\}$, then $P_0 \sim P_1$ and L^2 model holds.

(c) if $m_1 - m_0 \notin \mathcal{K}(L^2)$, L^2 model is never recovered, but different situations are possible.

Conditional Gaussian distributions

Let $X_0(s), X_1(s)$ be Gaussian processes with continuous trajectories, continuous mean functions m_0, m_1 and continuous covariance function K (equal for both classes).

Let P_0 and P_1 be the probability measures on $C[0,1]$ (or $L^2[0,1]$) induced by the processes X_0, X_1 respectively.

(Theorem 4.1)

(a) if $m_0, m_1 \in \mathcal{H}(K)$, then $P_0 \sim P_1$ and the RKHS model holds with $\beta := m_1 - m_0$ and $\beta_0 := (\|m_0\|_K^2 - \|m_1\|_K^2)/2 - \log((1-p)/p)$.

(b) if $m_1 - m_0 \in \mathcal{K}(L^2) = \{\mathcal{K}(f) : f \in L^2[0,1]\}$, then $P_0 \sim P_1$ and L^2 model holds.

(c) if $m_1 - m_0 \notin \mathcal{K}(L^2)$, L^2 model is never recovered, but different situations are possible.

Variable selection

Whenever the slope function β has the form

$$\beta(\cdot) = \sum_{i=1}^{p} \beta_j K(t_j, \cdot),$$

the RKHS model is reduced to,

$$\mathbb{P}(Y = 1 | X) = \left(1 + \exp\left\{-\beta_0 - \sum_{j=1}^{p} \beta_j X(t_j)\right\}\right)^{-1}.$$

Maximum Likelihood estimator

Given a sample $(x_1^0, y_1^0), \ldots, (x_{n_0}^0, y_{n_0}^0)$ and $(x_1^1, y_1^1), \ldots, (x_{n_1}^1, y_{n_1}^1)$, the ML function is

$$L_n(\beta, \beta_0) = \frac{1}{n_0} \sum_{i=1}^{n_0} \log \frac{e^{-\beta_0 - \Psi^{-1}_{x_i^0}(\beta)}}{1 + e^{-\beta_0 - \Psi^{-1}_{x_i^0}(\beta)}} + \frac{1}{n_1} \sum_{i=1}^{n_1} \log \underbrace{\frac{1}{1 + e^{-\beta_0 - \Psi^{-1}_{x_i^1}(\beta)}}}_{p_{\beta,\beta_0}}.$$

The expected log-likelihood function,

$$L(\beta, \beta_0) = \mathbb{E}\left[\log\left(p_{\beta,\beta_0}(X)^Y \left(1 - p_{\beta,\beta_0}(X)\right)^{1-Y}\right)\right],$$

has a unique maximum in $\mathcal{H}(K) \times \mathbb{R}$. (Prop. 4.2)

Maximum Likelihood estimator

Given a sample $(x_1^0, y_1^0), \ldots, (x_{n_0}^0, y_{n_0}^0)$ and $(x_1^1, y_1^1), \ldots, (x_{n_1}^1, y_{n_1}^1)$, the ML function is

$$L_n(\beta, \beta_0) = \frac{1}{n_0} \sum_{i=1}^{n_0} \log \frac{e^{-\beta_0 - \Psi_{x_i^0}^{-1}(\beta)}}{1 + e^{-\beta_0 - \Psi_{x_i^0}^{-1}(\beta)}} + \frac{1}{n_1} \sum_{i=1}^{n_1} \log \underbrace{\frac{1}{1 + e^{-\beta_0 - \Psi_{x_i^1}^{-1}(\beta)}}}_{p_{\beta,\beta_0}}.$$

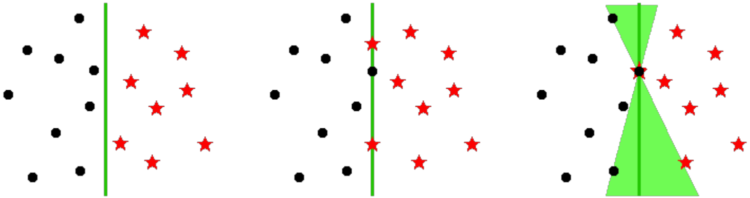
The expected log-likelihood function,

$$L(\beta, \beta_0) = \mathbb{E}\left[\log\left(p_{\beta,\beta_0}(X)^Y \left(1 - p_{\beta,\beta_0}(X)\right)^{1-Y}\right)\right],$$

has a unique maximum in $\mathcal{H}(K) \times \mathbb{R}$. (Prop. 4.2)

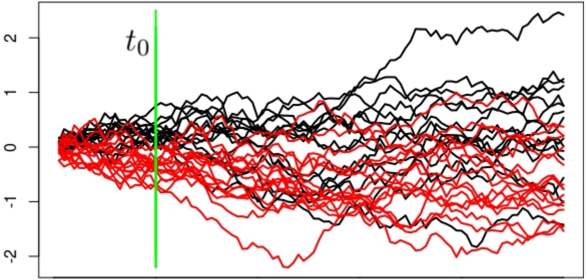
Non-existence of the finite MLE

The maximum of L_n is not attainable whenever the sample is (quasi)-linearly separable (Albert and Anderson (1984)).



Sign-choice property

$Z(t) = (X_1(t), \ldots, X_n(t))$, $t \in [0,1]$ satisfies the **Sign Choice (SC)** property when $\forall\,(s_1, \ldots, s_n) \in \{+,-\}^n$, with probability one, $\exists\, t_0 \in [0,1]$ such that $\text{sign}(X_i(t_0)) = s_i$.



Non-existence of the functional MLE

Let $X(s)$, $s \in [0, 1]$, be a L^2 stochastic process with $\mathbb{E}[X(s)] = 0$ and let $X_1, \ldots, X_n$ be independent copies of X.

If $Z_n(s) = (X_1(s), \ldots, X_n(s))$ fulfills the SC property, with probability one, the MLE estimator does not exist for any sample size. (Th. 4.4)

Prop. 4.5: The n-dimensional Brownian motion fulfills the SC property (via Blumental's 0-1 law).

Asymptotic non-existence

Let (x_i, y_i) be a independent sample satisfying the RKHS model, X Gaussian with K continuous and Σ_T invertible for any finite set T. Then, (Th. 4.6)

$$\lim_{n\to\infty} \mathbb{P}(\text{MLE exists}) = 0.$$

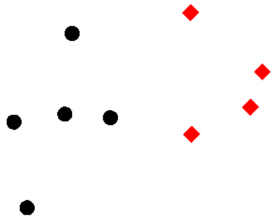
The estimation of β in practice:
Taking a finite number of points p and using Firth's estimator.

Asymptotic non-existence

Let (x_i, y_i) be a independent sample satisfying the RKHS model, X Gaussian with K continuous and Σ_T invertible for any finite set T. Then, (Th. 4.6)

$$\lim_{n\to\infty} \mathbb{P}(\text{MLE exists}) = 0.$$

The estimation of β in practice:
Taking a finite number of points p and using Firth's estimator.

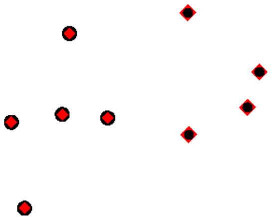


Asymptotic non-existence

Let (x_i, y_i) be a independent sample satisfying the RKHS model, X Gaussian with K continuous and Σ_T invertible for any finite set T. Then, (Th. 4.6)

$$\lim_{n\to\infty} \mathbb{P}(\text{MLE exists}) = 0.$$

The estimation of β in practice:
Taking a finite number of points p and using Firth's estimator.

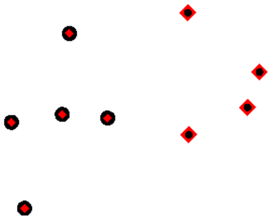


Asymptotic non-existence

Let (x_i, y_i) be a independent sample satisfying the RKHS model, X Gaussian with K continuous and Σ_T invertible for any finite set T. Then, (Th. 4.6)

$$\lim_{n\to\infty} \mathbb{P}(\text{MLE exists}) = 0.$$

The estimation of β in practice:
Taking a finite number of points p and using Firth's estimator.

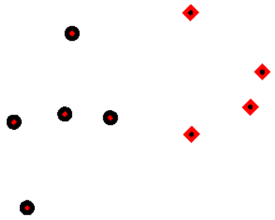


Asymptotic non-existence

Let (x_i, y_i) be a independent sample satisfying the RKHS model, X Gaussian with K continuous and Σ_T invertible for any finite set T. Then, (Th. 4.6)

$$\lim_{n\to\infty} \mathbb{P}(\text{MLE exists}) = 0.$$

The estimation of β in practice:
Taking a finite number of points p and using Firth's estimator.

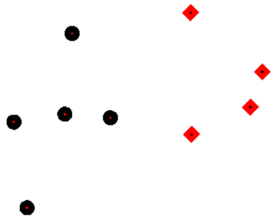


Asymptotic non-existence

Let (x_i, y_i) be a independent sample satisfying the RKHS model, X Gaussian with K continuous and Σ_T invertible for any finite set T. Then, (Th. 4.6)

$$\lim_{n\to\infty} \mathbb{P}(\text{MLE exists}) = 0.$$

The estimation of β in practice:
Taking a finite number of points p and using Firth's estimator.



Bibliography I

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Aneiros, G. and Vieu, P. (2014). Variable selection in infinite-dimensional problems. *Statistics and Probability Letters*, 94:12–20.

Cambanis, S. (1985). Sampling designs for time series. *Handbook of Statistics*, 5:337–362.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Galeano, P., Joseph, E., and Lillo, R. E. (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 57(2):281–291.

Ghiglietti, A., Ieva, F., and Paganoni, A. M. (2017). Statistical inference for stochastic processes: two-sample hypothesis tests. *Journal of Statistical Planning and Inference*, 180:49–68.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

Ji, H. and Müller, H.-G. (2016). Optimal designs for longitudinal and functional data. *J. Roy. Statist. Soc., B, to appear*, 79, Part 3:859–876.

Bibliography II

Lukić, M. N. and Beder, J. H. (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969.

McKeague, I. W. and Sen, B. (2010). Fractals with point impact in functional linear regression. *Ann. Statist.*, 38(4):2559–2586.

Peszat, S. and Zabczyk, J. (2007). *Stochastic partial differential equations with Lévy noise: An evolution equation approach*, volume 113. Cambridge University Press.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis*. Springer Series in Statistics.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics.