



HAL
open science

Classification and regression based on derivatives : a consistency result

Nathalie Villa-Vialaneix, Fabrice Rossi

► **To cite this version:**

Nathalie Villa-Vialaneix, Fabrice Rossi. Classification and regression based on derivatives : a consistency result. II Simposio sobre Modelamiento Estadístico, Dec 2010, Valparaiso, Chile. hal-00668212

HAL Id: hal-00668212

<https://hal.science/hal-00668212>

Submitted on 9 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Classification and regression based on derivatives: a consistency result

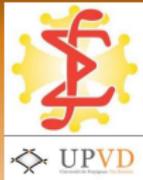
Nathalie Villa-Vialaneix (Joint work with Fabrice Rossi)

<http://www.nathalievilla.org>



II Simposio sobre Modelamiento Estadístico

Valparaiso, December, 3rd



Outline

- 1 Introduction and motivations**
- 2 A general consistency result
- 3 Examples



Regression and classification from an infinite dimensional predictor

Settings

(X, Y) is a random pair of variables where

- $Y \in \{-1, 1\}$ (binary classification problem) or $Y \in \mathbb{R}$



Regression and classification from an infinite dimensional predictor

Settings

(X, Y) is a random pair of variables where

- $Y \in \{-1, 1\}$ (binary classification problem) or $Y \in \mathbb{R}$
- $X \in (\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$, an infinite dimensional Hilbert space.



Regression and classification from an infinite dimensional predictor

Settings

(X, Y) is a random pair of variables where

- $Y \in \{-1, 1\}$ (binary classification problem) or $Y \in \mathbb{R}$
- $X \in (\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$, an infinite dimensional Hilbert space.

We are given a **learning set** $S_n = \{(X_i, Y_i)\}_{i=1}^n$ of n i.i.d. copies of (X, Y) .



Regression and classification from an infinite dimensional predictor

Settings

(X, Y) is a random pair of variables where

- $Y \in \{-1, 1\}$ (binary classification problem) or $Y \in \mathbb{R}$
- $X \in (\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$, an infinite dimensional Hilbert space.

We are given a **learning set** $S_n = \{(X_i, Y_i)\}_{i=1}^n$ of n i.i.d. copies of (X, Y) .

Purpose: Find $\phi_n : \mathcal{X} \rightarrow \{-1, 1\}$ or \mathbb{R} , that is universally consistent:

Classification case: $\lim_{n \rightarrow +\infty} \mathbb{P}(\phi_n(X) \neq Y) = L^*$ where $L^* = \inf_{\phi: \mathcal{X} \rightarrow \{-1, 1\}} \mathbb{P}(\phi(X) \neq Y)$ is the **Bayes risk**.



Regression and classification from an infinite dimensional predictor

Settings

(X, Y) is a random pair of variables where

- $Y \in \{-1, 1\}$ (binary classification problem) or $Y \in \mathbb{R}$
- $X \in (\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$, an infinite dimensional Hilbert space.

We are given a **learning set** $S_n = \{(X_i, Y_i)\}_{i=1}^n$ of n i.i.d. copies of (X, Y) .

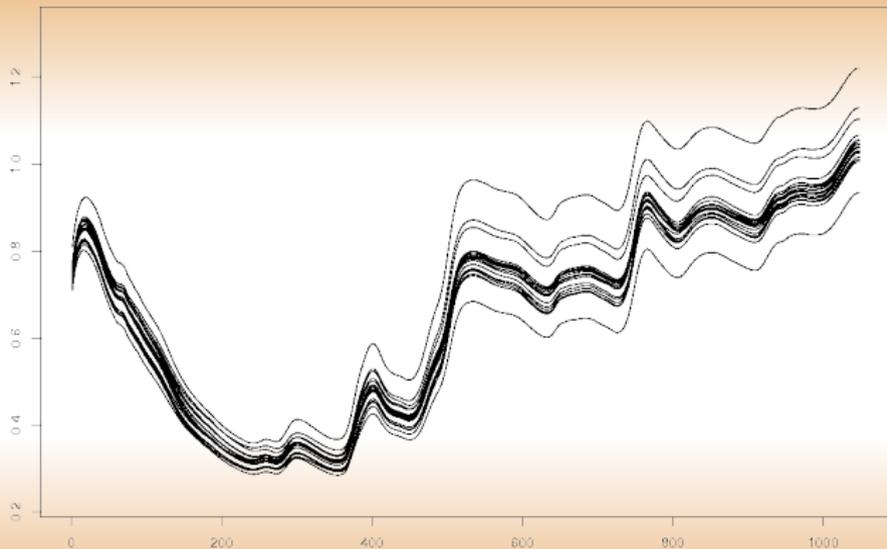
Purpose: Find $\phi_n : \mathcal{X} \rightarrow \{-1, 1\}$ or \mathbb{R} , that is universally consistent:

Regression case: $\lim_{n \rightarrow +\infty} \mathbb{E}([\phi_n(X) - Y]^2) = L^*$ where

$L^* = \inf_{\phi: \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}([\phi(X) - Y]^2)$ will also be called the Bayes risk.



An example

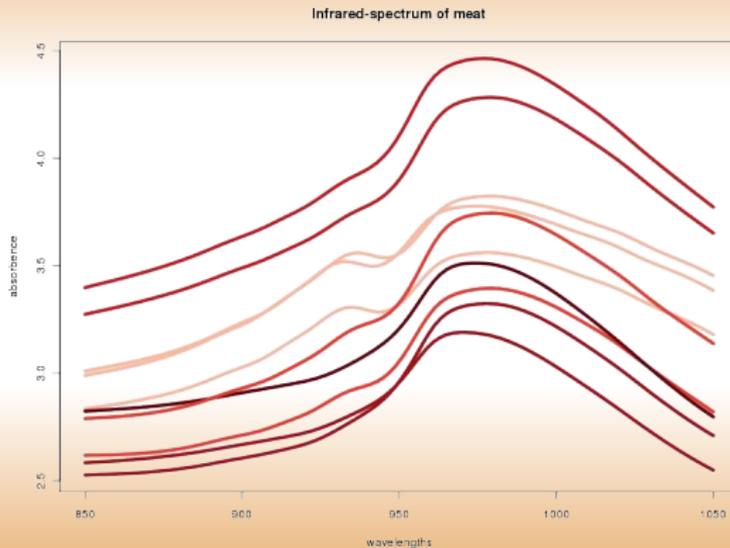


Predicting the **rate of yellow berry in durum wheat** from its **NIR spectrum**.



Using derivatives

Practically, $X^{(m)}$ is often more relevant than X for the prediction.

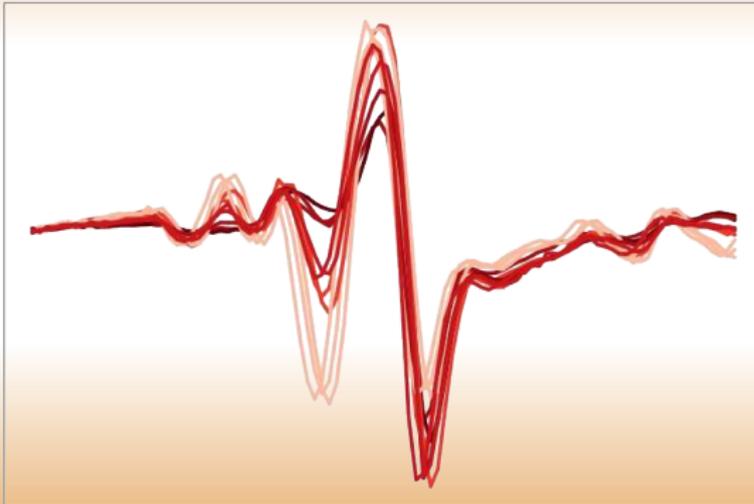




Using derivatives

Practically, $X^{(m)}$ is often more relevant than X for the prediction.

Second derivative: infrared-spectrum of meat





Using derivatives

Practically, $X^{(m)}$ is often more relevant than X for the prediction.
But $X \rightarrow X^{(m)}$ induces **information loss** and

$$\inf_{\phi: D^m \mathcal{X} \rightarrow \{-1,1\}} \mathbb{P}(\phi(X^{(m)}) \neq Y) \geq \inf_{\phi: \mathcal{X} \rightarrow \{-1,1\}} \mathbb{P}(\phi(X) \neq Y) = L^*$$

and

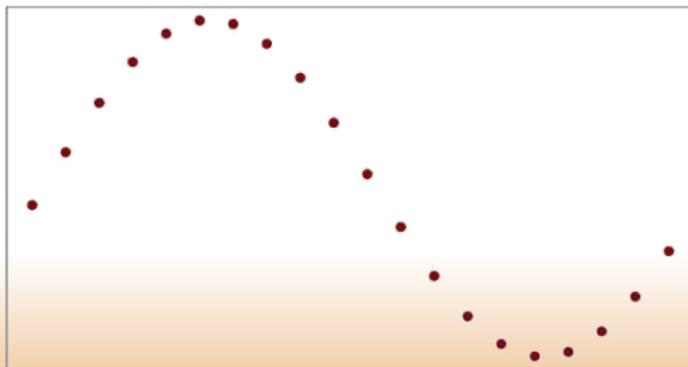
$$\inf_{\phi: D^m \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}([\phi(X^{(m)}) - Y]^2) \geq \inf_{\phi: \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}([\phi(X) - Y]^2) = L^*.$$



Sampled functions

Practically, $(X_i)_i$ are not perfectly known; only a discrete sampling is given: $\mathbf{X}_i^{\tau_d} = (X_i(t))_{t \in \tau_d}$ where $\tau_d = \{t_1^{\tau_d}, \dots, t_{|\tau_d|}^{\tau_d}\}$.

Uniform sampling,
non noisy data





Sampled functions

Practically, $(X_i)_i$ are not perfectly known; only a discrete sampling is given: $\mathbf{X}_i^{\tau_d} = (X_i(t))_{t \in \tau_d}$ where $\tau_d = \{t_1^{\tau_d}, \dots, t_{|\tau_d|}^{\tau_d}\}$.

Non uniform sampling,
non noisy data



The sampling can be non uniform...



Sampled functions

Practically, $(X_i)_i$ are not perfectly known; only a discrete sampling is given: $\mathbf{X}_i^{\tau_d} = (X_i(t))_{t \in \tau_d}$ where $\tau_d = \{t_1^{\tau_d}, \dots, t_{|\tau_d|}^{\tau_d}\}$.

Non uniform sampling,
noisy data



... and the data can be corrupted by noise.



Sampled functions

Practically, $(X_i)_i$ are not perfectly known; only a discrete sampling is given: $\mathbf{X}_i^{\tau_d} = (X_i(t))_{t \in \tau_d}$ where $\tau_d = \{t_1^{\tau_d}, \dots, t_{|\tau_d|}^{\tau_d}\}$.

Then, $X_i^{(m)}$ is **estimated** from $\mathbf{X}_i^{\tau_d}$, by $\widehat{X}_{\tau_d}^{(m)}$, which also induces **information loss**:

$$\inf_{\phi: D^m \mathcal{X} \rightarrow \{-1,1\}} \mathbb{P} \left(\phi(\widehat{X}_{\tau_d}^{(m)}) \neq Y \right) \geq \inf_{\phi: D^m \mathcal{X} \rightarrow \{-1,1\}} \mathbb{P} \left(\phi(X^{(m)}) \neq Y \right) \geq L^*$$

and

$$\inf_{\phi: D^m \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E} \left(\left[\phi(\widehat{X}_{\tau_d}^{(m)}) - Y \right]^2 \right) \geq \inf_{\phi: D^m \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E} \left(\left[\phi(X^{(m)}) - Y \right]^2 \right) \geq L^*.$$



Purpose of the presentation

Find a classifier or a regression function ϕ_{n,τ_d} built from $\widehat{X}_{\tau_d}^{(m)}$ such that the risk of ϕ_{n,τ_d} **asymptotically reaches** the Bayes risk L^* :

$$\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{P} \left(\phi_{n,\tau_d}(\widehat{X}_{\tau_d}^{(m)}) \neq Y \right) = L^*$$

or

$$\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{E} \left(\left[\phi_{n,\tau_d}(\widehat{X}_{\tau_d}^{(m)}) - Y \right]^2 \right) = L^*$$



Purpose of the presentation

Find a classifier or a regression function ϕ_{n,τ_d} built from $\widehat{X}_{\tau_d}^{(m)}$ such that the risk of ϕ_{n,τ_d} **asymptotically reaches** the Bayes risk L^* :

$$\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{P} \left(\phi_{n,\tau_d}(\widehat{X}_{\tau_d}^{(m)}) \neq Y \right) = L^*$$

or

$$\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{E} \left(\left[\phi_{n,\tau_d}(\widehat{X}_{\tau_d}^{(m)}) - Y \right]^2 \right) = L^*$$

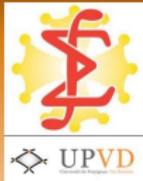
Main idea: Use a relevant way to estimate $X^{(m)}$ from \mathbf{X}^{τ_d} (by smoothing splines) and combine the consistency of splines with the consistency of a $\mathbb{R}^{|\tau_d|}$ -classifier or regression function.



A general consistency result

Outline

- 1 Introduction and motivations
- 2 A general consistency result**
- 3 Examples



A general consistency result

Basics about smoothing splines I

Suppose that \mathcal{X} is the Sobolev space

$$\mathcal{H}^m = \left\{ h \in L^2_{[0,1]} \mid \forall j = 1, \dots, m, D^j h \text{ exists (weak sense) and } D^m h \in L^2 \right\}$$



Basics about smoothing splines I

Suppose that \mathcal{X} is the Sobolev space

$\mathcal{H}^m = \{h \in L^2_{[0,1]} \mid \forall j = 1, \dots, m, D^j h \text{ exists (weak sense) and } D^m h \in L^2\}$
equipped with the scalar product

$$\langle u, v \rangle_{\mathcal{H}^m} = \langle D^m u, D^m v \rangle_{L^2} + \sum_{j=1}^m B^j u B^j v$$

where B are m boundary conditions such that $\text{Ker} B \cap \mathbb{P}^{m-1} = \{0\}$.



Basics about smoothing splines I

Suppose that \mathcal{X} is the Sobolev space

$\mathcal{H}^m = \{h \in L^2_{[0,1]} | \forall j = 1, \dots, m, D^j h \text{ exists (weak sense) and } D^m h \in L^2\}$
equipped with the scalar product

$$\langle u, v \rangle_{\mathcal{H}^m} = \langle D^m u, D^m v \rangle_{L^2} + \sum_{j=1}^m B^j u B^j v$$

where B are m boundary conditions such that $\text{Ker} B \cap \mathbb{P}^{m-1} = \{0\}$.

$(\mathcal{H}^m, \langle \cdot, \cdot \rangle_{\mathcal{H}^m})$ is a RKHS: $\exists k_0 : \mathbb{P}^{m-1} \times \mathbb{P}^{m-1} \rightarrow \mathbb{R}$ and $k_1 : \text{Ker} B \times \text{Ker} B \rightarrow \mathbb{R}$ such that

$$\forall u \in \mathbb{P}^{m-1}, t \in [0, 1], \langle u, k_0(t, \cdot) \rangle_{\mathcal{H}^m} = u(t)$$

and

$$\forall u \in \text{Ker} B, t \in [0, 1], \langle u, k_1(t, \cdot) \rangle_{\mathcal{H}^m} = u(t)$$

See **[Berlinet and Thomas-Agnan, 2004]** for further details.



Basics about smoothing splines II

A simple example of boundary conditions:

$$h(0) = h^{(1)}(0) = \dots = h^{(m-1)}(0) = 0.$$

Then,

$$k_0(s, t) = \sum_{k=0}^{m-1} \frac{t^k s^k}{(k!)^2}$$

and

$$k_1(s, t) = \int_0^1 \frac{(t-w)_+^{m-1} (s-w)_+^{m-1}}{(m-1)!} dw.$$



Estimating the predictors with smoothing splines I

Assumption (A1)

- $|\tau_d| \geq m - 1$
- sampling points are distinct in $[0, 1]$
- B^j are linearly independent from $h \rightarrow h(t)$ for all $t \in \tau_d$



Estimating the predictors with smoothing splines I

Assumption (A1)

- $|\tau_d| \geq m - 1$
- sampling points are distinct in $[0, 1]$
- B^j are linearly independent from $h \rightarrow h(t)$ for all $t \in \tau_d$

[Kimeldorf and Wahba, 1971]: for \mathbf{x}^{τ_d} in $\mathbb{R}^{|\tau_d|}$, $\exists ! \hat{\chi}_{\lambda, \tau_d} \in \mathcal{H}^m$
solution of

$$\arg \min_{h \in \mathcal{H}^m} \frac{1}{|\tau_d|} \sum_{l=1}^{|\tau_d|} (h(t_l) - \mathbf{x}^{\tau_d})^2 + \lambda \int_{[0,1]} (h^{(m)}(t))^2 dt.$$

and $\hat{\chi}_{\lambda, \tau_d} = \mathcal{S}_{\lambda, \tau_d} \mathbf{x}^{\tau_d}$ where $\mathcal{S}_{\lambda, \tau_d} : \mathbb{R}^{|\tau_d|} \rightarrow \mathcal{H}^m$.



Estimating the predictors with smoothing splines I

Assumption (A1)

- $|\tau_d| \geq m - 1$
- sampling points are distinct in $[0, 1]$
- B^j are linearly independent from $h \rightarrow h(t)$ for all $t \in \tau_d$

[Kimeldorf and Wahba, 1971]: for \mathbf{x}^{τ_d} in $\mathbb{R}^{|\tau_d|}$, $\exists ! \hat{\chi}_{\lambda, \tau_d} \in \mathcal{H}^m$
solution of

$$\arg \min_{h \in \mathcal{H}^m} \frac{1}{|\tau_d|} \sum_{l=1}^{|\tau_d|} (h(t_l) - \mathbf{x}^{\tau_d})^2 + \lambda \int_{[0,1]} (h^{(m)}(t))^2 dt.$$

and $\hat{\chi}_{\lambda, \tau_d} = \mathcal{S}_{\lambda, \tau_d} \mathbf{x}^{\tau_d}$ where $\mathcal{S}_{\lambda, \tau_d} : \mathbb{R}^{|\tau_d|} \rightarrow \mathcal{H}^m$.

These assumptions are fulfilled by the previous simple example as long as $0 \notin \tau_d$.



Estimating the predictors with smoothing splines II

$\mathcal{S}_{\lambda, \tau_d}$ is given by:

$$\begin{aligned}\mathcal{S}_{\lambda, \tau_d} &= \omega^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|}) U^T)^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} \\ &\quad + \eta^T (K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} (\mathbb{I}_{|\tau_d|} - U^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|}) U^T)^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1}) \\ &= \omega^T M_0 + \eta^T M_1\end{aligned}$$

with

- $\{\omega_1, \dots, \omega_m\}$ is a basis of \mathbb{P}^{m-1} , $\omega = (\omega_1, \dots, \omega_m)^T$ and $U = (\omega_i(t))_{i=1, \dots, m, t \in \tau_d}$;
- $\eta = (k_1(t, \cdot))_{t \in \tau_d}^T$ and $K_1 = (k_1(t, t'))_{t, t' \in \tau_d}$.



Estimating the predictors with smoothing splines II

$\mathcal{S}_{\lambda, \tau_d}$ is given by:

$$\begin{aligned}\mathcal{S}_{\lambda, \tau_d} &= \omega^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|}) U^T)^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} \\ &\quad + \eta^T (K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} (\mathbb{I}_{|\tau_d|} - U^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|}) U^T)^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1}) \\ &= \omega^T M_0 + \eta^T M_1\end{aligned}$$

with

- $\{\omega_1, \dots, \omega_m\}$ is a basis of \mathbb{P}^{m-1} , $\omega = (\omega_1, \dots, \omega_m)^T$ and $U = (\omega_i(t))_{i=1, \dots, m, t \in \tau_d}$;
- $\eta = (k_1(t, \cdot))_{t \in \tau_d}^T$ and $K_1 = (k_1(t, t'))_{t, t' \in \tau_d}$.

The observations of the **predictor X (NIR spectra)** are then **estimated** from their sampling \mathbf{X}^{τ_d} by $\widehat{X}_{\lambda, \tau_d}$.



Two important consequences

1 No information loss

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$$

and

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{E}([\phi(\widehat{X}_{\lambda, \tau_d}) - Y]^2) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{E}([\phi(\mathbf{X}^{\tau_d}) - Y]^2)$$



Two important consequences

1 No information loss

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$$

and

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{E}([\phi(\widehat{X}_{\lambda, \tau_d}) - Y]^2) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{E}([\phi(\mathbf{X}^{\tau_d}) - Y]^2)$$

2 Easy way to use derivatives:

$$\langle \mathcal{S}_{\lambda, \tau_d} \mathbf{u}^{\tau_d}, \mathcal{S}_{\lambda, \tau_d} \mathbf{v}^{\tau_d} \rangle_{\mathcal{H}^m} = \langle \widehat{\mathbf{u}}_{\lambda, \tau_d}, \widehat{\mathbf{v}}_{\lambda, \tau_d} \rangle_{\mathcal{H}^m}$$



Two important consequences

1 No information loss

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{\tau_d} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$$

and

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{E}([\phi(\widehat{X}_{\lambda, \tau_d}) - Y]^2) = \inf_{\phi: \mathbb{R}^{\tau_d} \rightarrow \{-1,1\}} \mathbb{P}([\phi(\mathbf{X}^{\tau_d}) - Y]^2)$$

2 Easy way to use derivatives:

$$(\mathbf{u}^{\tau_d})^T M_0^T W M_0 \mathbf{v}^{\tau_d} + (\mathbf{u}^{\tau_d})^T M_1^T K_1 M_1 \mathbf{v}^{\tau_d} = \langle \widehat{\mathbf{u}}_{\lambda, \tau_d}, \widehat{\mathbf{v}}_{\lambda, \tau_d} \rangle_{\mathcal{H}^m}$$

where K_1 , M_0 and M_1 have been previously defined and $W = (\langle \omega_i, \omega_j \rangle_{\mathcal{H}^m})_{i,j=1,\dots,m}$.



Two important consequences

1 No information loss

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$$

and

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{E}([\phi(\widehat{X}_{\lambda, \tau_d}) - Y]^2) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}([\phi(\mathbf{X}^{\tau_d}) - Y]^2)$$

2 Easy way to use derivatives:

$$(\mathbf{u}^{\tau_d})^T \mathbf{M}_{\lambda, \tau_d} \mathbf{v}^{\tau_d} = \langle \widehat{\mathbf{u}}_{\lambda, \tau_d}, \widehat{\mathbf{v}}_{\lambda, \tau_d} \rangle_{\mathcal{H}^m}$$

where $\mathbf{M}_{\lambda, \tau_d}$ is symmetric, definite positive.



A general consistency result

Two important consequences

1 No information loss

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{\tau_d} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$$

and

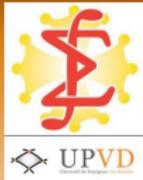
$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{E}([\phi(\widehat{X}_{\lambda, \tau_d}) - Y]^2) = \inf_{\phi: \mathbb{R}^{\tau_d} \rightarrow \{-1,1\}} \mathbb{P}([\phi(\mathbf{X}^{\tau_d}) - Y]^2)$$

2 Easy way to use derivatives:

$$(\mathbf{Q}_{\lambda, \tau_d} \mathbf{u}^{\tau_d})^T (\mathbf{Q}_{\lambda, \tau_d} \mathbf{v}^{\tau_d}) = \langle \widehat{\mathbf{u}}_{\lambda, \tau_d}, \widehat{\mathbf{v}}_{\lambda, \tau_d} \rangle_{\mathcal{H}^m}$$

where $\mathbf{Q}_{\lambda, \tau_d}$ is the Choleski triangle of $\mathbf{M}_{\lambda, \tau_d}$: $\mathbf{Q}_{\lambda, \tau_d}^T \mathbf{Q}_{\lambda, \tau_d} = \mathbf{M}_{\lambda, \tau_d}$.

Remark: $\mathbf{Q}_{\lambda, \tau_d}$ is calculated only from the RKHS, λ and τ_d : it does not depend on the data set.



A general consistency result

Two important consequences

1 No information loss

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$$

and

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1,1\}} \mathbb{E}([\phi(\widehat{X}_{\lambda, \tau_d}) - Y]^2) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}([\phi(\mathbf{X}^{\tau_d}) - Y]^2)$$

2 Easy way to use derivatives:

$$\begin{aligned} (\mathbf{Q}_{\lambda, \tau_d} \mathbf{u}^{\tau_d})^T (\mathbf{Q}_{\lambda, \tau_d} \mathbf{v}^{\tau_d}) &= \langle \widehat{\mathbf{u}}_{\lambda, \tau_d}, \widehat{\mathbf{v}}_{\lambda, \tau_d} \rangle_{\mathcal{H}^m} \\ &\simeq \langle \widehat{\mathbf{u}}_{\lambda, \tau_d}^{(m)}, \widehat{\mathbf{v}}_{\lambda, \tau_d}^{(m)} \rangle_{L^2} \end{aligned}$$

where $\mathbf{Q}_{\lambda, \tau_d}$ is the Choleski triangle of $\mathbf{M}_{\lambda, \tau_d}$: $\mathbf{Q}_{\lambda, \tau_d}^T \mathbf{Q}_{\lambda, \tau_d} = \mathbf{M}_{\lambda, \tau_d}$.

Remark: $\mathbf{Q}_{\lambda, \tau_d}$ is calculated only from the RKHS, λ and τ_d : it does not depend on the data set.



Classification and regression based on derivatives

Suppose that we know a **consistent classifier or regression function in $\mathbb{R}^{|\tau_d|}$** that is based on $\mathbb{R}^{|\tau_d|}$ scalar product or norm.

Example: Nonparametric kernel regression

$$\Psi : u \in \mathbb{R}^{|\tau_d|} \rightarrow \frac{\sum_{i=1}^n T_i K\left(\frac{\|u - U_i\|_{\mathbb{R}^{|\tau_d|}}}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\|u - U_i\|_{\mathbb{R}^{|\tau_d|}}}{h_n}\right)}$$

where $(U_i, T_i)_{i=1, \dots, n}$ is a learning set in $\mathbb{R}^{|\tau_d|} \times \mathbb{R}$.



Classification and regression based on derivatives

Suppose that we know a **consistent classifier or regression function in $\mathbb{R}^{|\tau_d|}$** that is based on $\mathbb{R}^{|\tau_d|}$ scalar product or norm. The **corresponding derivative based classifier or regression function** is given by using the norm induced by $\mathbf{Q}_{\lambda, \tau_d}$:

Example: Nonparametric kernel regression

$$\phi_{n,d} = \Psi \circ \mathbf{Q}_{\lambda, \tau_d} : x \in \mathcal{H}^m \rightarrow \frac{\sum_{i=1}^n Y_i K \left(\frac{\|\mathbf{Q}_{\lambda, \tau_d} \mathbf{x}^{\tau_d} - \mathbf{Q}_{\lambda, \tau_d} \mathbf{x}_i^{\tau_d}\|_{\mathbb{R}^{|\tau_d|}}}{h_n} \right)}{\sum_{i=1}^n K \left(\frac{\|\mathbf{Q}_{\lambda, \tau_d} \mathbf{x}^{\tau_d} - \mathbf{Q}_{\lambda, \tau_d} \mathbf{x}_i^{\tau_d}\|_{\mathbb{R}^{|\tau_d|}}}{h_n} \right)}$$



Classification and regression based on derivatives

Suppose that we know a **consistent classifier or regression function in $\mathbb{R}^{|\tau_d|}$** that is based on $\mathbb{R}^{|\tau_d|}$ scalar product or norm. The **corresponding derivative based classifier or regression function** is given by using the norm induced by $\mathbf{Q}_{\lambda, \tau_d}$:

Example: Nonparametric kernel regression

$$\begin{aligned}
 \phi_{n,d} = \Psi \circ \mathbf{Q}_{\lambda, \tau_d} : x \in \mathcal{H}^m &\rightarrow \frac{\sum_{i=1}^n Y_i K \left(\frac{\|\mathbf{Q}_{\lambda, \tau_d} \mathbf{x}^{\tau_d} - \mathbf{Q}_{\lambda, \tau_d} \mathbf{X}_i^{\tau_d}\|_{\mathbb{R}^{|\tau_d|}}}{h_n} \right)}{\sum_{i=1}^n K \left(\frac{\|\mathbf{Q}_{\lambda, \tau_d} \mathbf{x}^{\tau_d} - \mathbf{Q}_{\lambda, \tau_d} \mathbf{X}_i^{\tau_d}\|_{\mathbb{R}^{|\tau_d|}}}{h_n} \right)} \\
 &\downarrow \approx \\
 &\frac{\sum_{i=1}^n Y_i K \left(\frac{\|x^{(m)} - X_i^{(m)}\|_{L^2}}{h_n} \right)}{\sum_{i=1}^n K \left(\frac{\|x^{(m)} - X_i^{(m)}\|_{L^2}}{h_n} \right)}
 \end{aligned}$$



A general consistency result

Remark for consistency

Classification case (approximately the same is true for regression):

$$\mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L^* = \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L_d^* + L_d^* - L^*$$

where $L_d^* = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$.



Remark for consistency

Classification case (approximately the same is true for regression):

$$\mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L^* = \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L_d^* + L_d^* - L^*$$

where $L_d^* = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$.

① For all fixed d ,

$$\lim_{n \rightarrow +\infty} \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) = L_d^*$$

as long as the $\mathbb{R}^{|\tau_d|}$ -classifier is consistent because there is a one-to-one mapping between \mathbf{X}^{τ_d} and $\widehat{X}_{\lambda,\tau_d}$.



Remark for consistency

Classification case (approximately the same is true for regression):

$$\mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L^* = \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L_d^* + L_d^* - L^*$$

where $L_d^* = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$.

- 1 For all fixed d ,

$$\lim_{n \rightarrow +\infty} \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) = L_d^*$$

as long as the $\mathbb{R}^{|\tau_d|}$ -classifier is consistent because there is a one-to-one mapping between \mathbf{X}^{τ_d} and $\widehat{X}_{\lambda,\tau_d}$.

- 2 $L_d^* - L^* \leq \mathbb{E} \left(\left| \mathbb{E}(Y|\widehat{X}_{\lambda,\tau_d}) - \mathbb{E}(Y|X) \right| \right)$

with consistency of spline estimate $\widehat{X}_{\lambda,\tau_d}$ and assumption on the regularity of $\mathbb{E}(Y|X = \cdot)$, consistency would be proved.



Remark for consistency

Classification case (approximately the same is true for regression):

$$\mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L^* = \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) - L_d^* + L_d^* - L^*$$

where $L_d^* = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y)$.

- 1 For all fixed d ,

$$\lim_{n \rightarrow +\infty} \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y) = L_d^*$$

as long as the $\mathbb{R}^{|\tau_d|}$ -classifier is consistent because there is a one-to-one mapping between \mathbf{X}^{τ_d} and $\widehat{X}_{\lambda,\tau_d}$.

- 2 $L_d^* - L^* \leq \mathbb{E} \left(\left| \mathbb{E}(Y|\widehat{X}_{\lambda,\tau_d}) - \mathbb{E}(Y|X) \right| \right)$

with consistency of spline estimate $\widehat{X}_{\lambda,\tau_d}$ and assumption on the regularity of $\mathbb{E}(Y|X = \cdot)$, consistency would be proved.

But continuity of $\mathbb{E}(Y|X = \cdot)$ is a strong assumption in infinite dimensional case, and is not easy to check.



A general consistency result

Spline consistency

Let λ depends on d and denote $(\lambda_d)_d$ the series of regularization parameters. Also introduce

$$\overline{\Delta}_{\tau_d} := \max\{t_1, t_2 - t_1, \dots, 1 - t_{|\tau_d|}\}, \quad \underline{\Delta}_{\tau_d} := \min_{1 \leq i < |\tau_d|} \{t_{i+1} - t_i\}$$

Assumption (A2)

- $\exists R$ such that $\overline{\Delta}_{\tau_d} / \underline{\Delta}_{\tau_d} \leq R$ for all d ;
- $\lim_{d \rightarrow +\infty} |\tau_d| = +\infty$;
- $\lim_{d \rightarrow +\infty} \lambda_d = 0$.



Spline consistency

Let λ depends on d and denote $(\lambda_d)_d$ the series of regularization parameters. Also introduce

$$\overline{\Delta}_{\tau_d} := \max\{t_1, t_2 - t_1, \dots, 1 - t_{|\tau_d|}\}, \quad \underline{\Delta}_{\tau_d} := \min_{1 \leq i < |\tau_d|} \{t_{i+1} - t_i\}$$

Assumption (A2)

- $\exists R$ such that $\overline{\Delta}_{\tau_d} / \underline{\Delta}_{\tau_d} \leq R$ for all d ;
- $\lim_{d \rightarrow +\infty} |\tau_d| = +\infty$;
- $\lim_{d \rightarrow +\infty} \lambda_d = 0$.

[Ragozin, 1983]: Under (A1) and (A2), $\exists A_{R,m}$ and $B_{R,m}$ such that for any $x \in \mathcal{H}^m$ and any $\lambda_d > 0$,

$$\|\hat{x}_{\lambda_d, \tau_d} - x\|_{L^2}^2 \leq \left(A_{R,m} \lambda_d + B_{R,m} \frac{1}{|\tau_d|^{2m}} \right) \|D^m x\|_{L^2}^2 \xrightarrow{d \rightarrow +\infty} 0$$



A general consistency result

Bayes risk consistency

Assumption (A3a)

$\mathbb{E} \left(\|D^m X\|_{L^2}^2 \right)$ is finite and $Y \in \{-1, 1\}$.



A general consistency result

Bayes risk consistency

Assumption (A3a)

$\mathbb{E}(\|D^m X\|_{L^2}^2)$ is finite and $Y \in \{-1, 1\}$.

or

Assumption (A3b)

$\tau_d \subset \tau_{d+1}$ for all d and $\mathbb{E}(Y^2)$ is finite.



A general consistency result

Bayes risk consistency

Assumption (A3a)

$\mathbb{E}(\|D^m X\|_{L^2}^2)$ is finite and $Y \in \{-1, 1\}$.

or

Assumption (A3b)

$\tau_d \subset \tau_{d+1}$ for all d and $\mathbb{E}(Y^2)$ is finite.

Under (A1)-(A3), $\lim_{d \rightarrow +\infty} L_d^* = L^*$.



A general consistency result

Proof under assumption (A3a)

Assumption (A3a)

$\mathbb{E}(\|D^m X\|_{L^2}^2)$ is finite and $Y \in \{-1, 1\}$.



Proof under assumption (A3a)

Assumption (A3a)

$\mathbb{E}(\|D^m X\|_{L^2}^2)$ is finite and $Y \in \{-1, 1\}$.

The proof is based on a result of **[Faragó and Györfi, 1975]**:

For a pair of random variables (X, Y) taking their values in $\mathcal{X} \times \{-1, 1\}$ where \mathcal{X} is an arbitrary metric space and for a series of functions $T_d : \mathcal{X} \rightarrow \mathcal{X}$ such that

$$\mathbb{E}(\delta(T_d(X), X)) \xrightarrow{d \rightarrow +\infty} 0$$

then $\lim_{d \rightarrow +\infty} \inf_{\phi: \mathcal{X} \rightarrow \{-1, 1\}} \mathbb{P}(\phi(T_d(X)) \neq Y) = L^$.*



Proof under assumption (A3a)

Assumption (A3a)

$\mathbb{E}(\|D^m X\|_{L^2}^2)$ is finite and $Y \in \{-1, 1\}$.

The proof is based on a result of **[Faragó and Györfi, 1975]**:

- T_d is the spline estimate based on the sampling;
- the inequality of **[Ragozin, 1983]** about this estimate is exactly the assumption of Farago and Györfi's Theorem.

Then the result follows.



A general consistency result

Proof under assumption (A3b)

Assumption (A3b)

$\tau_d \subset \tau_{d+1}$ for all d and $\mathbb{E}(Y^2)$ is finite.



A general consistency result

Proof under assumption (A3b)

Assumption (A3b)

$\tau_d \subset \tau_{d+1}$ for all d and $\mathbb{E}(Y^2)$ is finite.

Under (A3b), $(\mathbb{E}(Y|\widehat{X}_{\lambda_d, \tau_d}))_d$ is a uniformly bounded martingale and thus converges for the L^1 -norm. Using the consistency of $(\widehat{X}_{\lambda_d, \tau_d})_d$ to X ends the proof.



Concluding result (consistency)

Theorem

Under assumptions (A1)-(A3),

$$\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{P}(\phi_{n,\tau_d}(\widehat{X}_{\lambda_d,\tau_d}) \neq Y) = L^*$$

and

$$\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{E}([\phi_{n,\tau_d}(\widehat{X}_{\lambda_d,\tau_d}) - Y]^2) = L^*$$

Proof: For a $\epsilon > 0$, fix d_0 such that, for all $d \geq d_0$, $L_d^* - L^* \leq \epsilon/2$. Then, by consistency of the $\mathbb{R}^{|\tau_d|}$ -classifier or regression function, conclude.



A practical application to SVM I

Recall that, for a learning set $(U_i, T_i)_{i=1, \dots, n}$ in $\mathbb{R}^p \times \{-1, 1\}$, gaussian SVM is the classifier

$$u \in \mathbb{R}^p \rightarrow \text{Sign} \left(\sum_{i=1}^n \alpha_i T_i e^{-\gamma \|u - U_i\|_{\mathbb{R}^p}^2} \right)$$

where $(\alpha_i)_i$ satisfy the following quadratic optimization problem:

$$\arg \min_w \sum_{i=1}^n |1 - T_i w(U_i)|_+ + C \|w\|_{\mathcal{S}}^2$$

where $w(u) = \sum_{i=1}^n \alpha_i e^{-\gamma \|u - U_i\|_{\mathbb{R}^p}^2}$ and \mathcal{S} is the RKHS associated with the gaussian kernel and C is a **regularization parameter**.



A practical application to SVM I

Recall that, for a learning set $(U_i, T_i)_{i=1, \dots, n}$ in $\mathbb{R}^p \times \{-1, 1\}$, gaussian SVM is the classifier

$$u \in \mathbb{R}^p \rightarrow \text{Sign} \left(\sum_{i=1}^n \alpha_i T_i e^{-\gamma \|u - U_i\|_{\mathbb{R}^p}^2} \right)$$

where $(\alpha_i)_i$ satisfy the following quadratic optimization problem:

$$\arg \min_w \sum_{i=1}^n |1 - T_i w(U_i)|_+ + C \|w\|_{\mathcal{S}}^2$$

where $w(u) = \sum_{i=1}^n \alpha_i e^{-\gamma \|u - U_i\|_{\mathbb{R}^p}^2}$ and \mathcal{S} is the RKHS associated with the gaussian kernel and C is a **regularization parameter**. Under suitable assumptions, **[Steinwart, 2002]** proves the consistency of SVM classifiers.



A practical application to SVM II

Additional assumptions related to SVM: Assumptions (A4)

- For all d , the regularization parameter depends on n such that $\lim_{n \rightarrow +\infty} nC_n^d = +\infty$ and $C_n^d = O_n(n^{\beta_d - 1})$ for a $0 < \beta_d < 1/d$.
- For all d , there is a bounded subset of $\mathbb{R}^{|\tau_d|}$, \mathcal{B}_d , such that \mathbf{X}^{τ_d} belongs to \mathcal{B}_d .



A practical application to SVM II

Additional assumptions related to SVM: Assumptions (A4)

- For all d , the regularization parameter depends on n such that $\lim_{n \rightarrow +\infty} nC_n^d = +\infty$ and $C_n^d = O_n(n^{\beta_d - 1})$ for a $0 < \beta_d < 1/d$.
- For all d , there is a bounded subset of $\mathbb{R}^{|\tau_d|}$, \mathcal{B}_d , such that \mathbf{X}^{τ_d} belongs to \mathcal{B}_d .

Result: Under assumptions (A1)-(A4), the SVM $\phi_{n,d} : x \in \mathcal{H}^m \rightarrow$

$$\text{Sign} \left(\sum_{i=1}^n \alpha_i Y_i e^{-\gamma \|\mathbf{Q}_{\lambda_d, \tau_d} \mathbf{x}^{\tau_d} - \mathbf{Q}_{\lambda_d, \tau_d} \mathbf{x}_i^{\tau_d}\|_{\mathbb{R}^d}^2} \right) \simeq \text{Sign} \left(\sum_{i=1}^n \alpha_i Y_i e^{-\gamma \|x^{(m)} - X_i^{(m)}\|_{L^2}^2} \right)$$

is consistent: $\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{P}(\phi_{n, \tau_d}(\widehat{X}_{\lambda_d, \tau_d}) \neq Y) = L^*$.



Additional remark about the link between n and $|\tau_d|$

Under suitable (and usual) regularity assumptions on $\mathbb{E}(Y|X = \cdot)$ and if $n \sim \nu^{|\tau_d| \log |\tau_d|}$, the **rate of convergence** of this method is of order $d^{-\frac{2\nu}{2\nu+1}}$ where ν is either equal to m or to a Lipschitz constant related to $\mathbb{E}(Y|X = \cdot)$.



Examples

Outline

- 1 Introduction and motivations
- 2 A general consistency result
- 3 Examples**



Chosen regression method: Regression with kernel ridge regression

Recall that **kernel ridge regression** in \mathbb{R}^p is given by solving

$$\arg \min_w \sum_{i=1}^n (T_i - w(U_i))^2 + C \|w\|_S^2$$

where S is a RKHS induced by a given kernel (such as the Gaussian kernel) and $(U_i, T_i)_i$ is a training sample in $\mathbb{R}^p \times \mathbb{R}$.



Chosen regression method: Regression with kernel ridge regression

Recall that **kernel ridge regression** in \mathbb{R}^p is given by solving

$$\arg \min_w \sum_{i=1}^n (T_i - w(U_i))^2 + C \|w\|_S^2$$

where S is a RKHS induced by a given kernel (such as the Gaussian kernel) and $(U_i, T_i)_i$ is a training sample in $\mathbb{R}^p \times \mathbb{R}$. In the following examples, U_i is either:

- the original (sampled) functions \mathbf{X}_i (viewed as $\mathbb{R}^{|\tau_d|}$ vectors);
- $\mathbf{Q}_{\lambda, \tau_d} \mathbf{X}_i^{\tau_d}$ for derivatives of order 1 or 2.



Example 1: Predicting yellow berry in durum wheat from NIR spectra

953 wheat samples were analyzed:

- **NIR spectrometry**: 1049 wavelengths regularly ranged from 400 to 2498 nm;
- **Yellow berry**: manual count (%) of affected grains.



Example 1: Predicting yellow berry in durum wheat from NIR spectra

953 wheat samples were analyzed:

- **NIR spectrometry**: 1049 wavelengths regularly ranged from 400 to 2498 nm;
- **Yellow berry**: manual count (%) of affected grains.

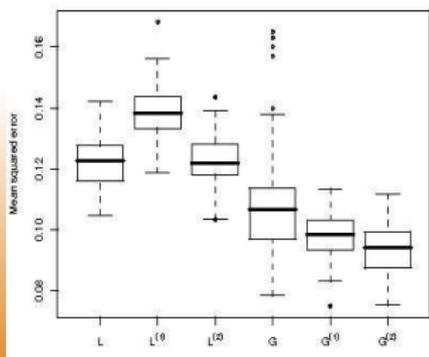
Methodology for comparison:

- **Split the data** into train/test sets (50 times);
- **Train** 50 regression functions for the 50 train sets (hyper-parameters were tuned by CV);
- **Evaluate** these regression functions by calculating the **MSE** for the 50 corresponding test sets.



Example 1: Predicting yellow berry in durum wheat from NIR spectra

Kernel (SVM)	MSE on test (and sd $\times 10^{-3}$)
Linear (L)	0.122 (8.77)
Linear on derivatives ($L^{(1)}$)	0.138 (9.53)
Linear on second derivatives ($L^{(2)}$)	0.122 (1.71)
Gaussian (G)	0.110 (20.2)
Gaussian on derivatives ($G^{(1)}$)	0.098 (7.92)
Gaussian on second derivatives ($G^{(2)}$)	0.094 (8.35)



The differences are significant between $G^{(2)}$ / $G^{(1)}$ and between $G^{(1)}$ / G .

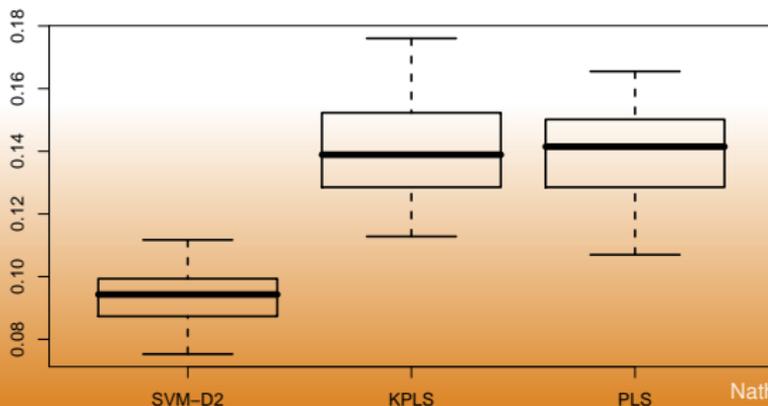


Examples

Comparison with PLS...

	MSE (mean)	MSE (sd)
PLS	0.154	0.012
Kernel PLS	0.154	0.013
KRR splines (reg. D^2)	0.094	0.008

Error decrease: almost 40 %

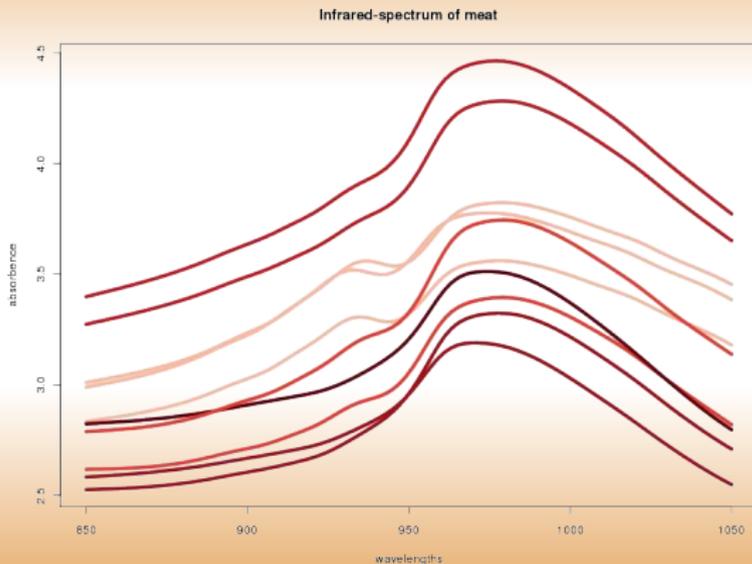




Examples

Example 2: Simulated noisy spectra

Original data:



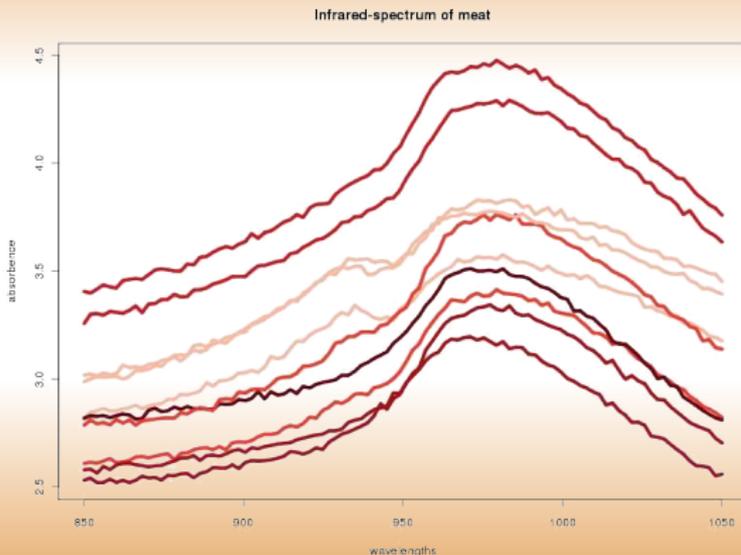
Variable to predict: Fat content of pieces of meat.



Examples

Example 2: Simulated noisy spectra

Noisy data: $X_i^b(t) = X_i(t) + \epsilon_{it}$, $\epsilon_{it} \sim \mathcal{N}(0, 0.01)$, i.i.d.:

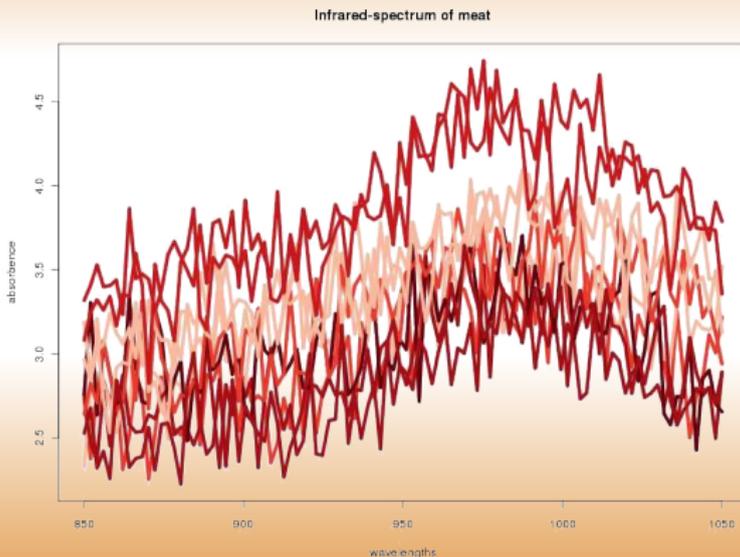


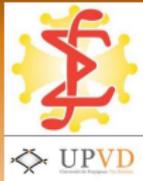


Examples

Example 2: Simulated noisy spectra

Worse noisy data: $X_i^b(t) = X_i(t) + \epsilon_{it}$, $\epsilon_{it} \sim \mathcal{N}(0, 0.2)$, i.i.d.:





Methodology for comparison

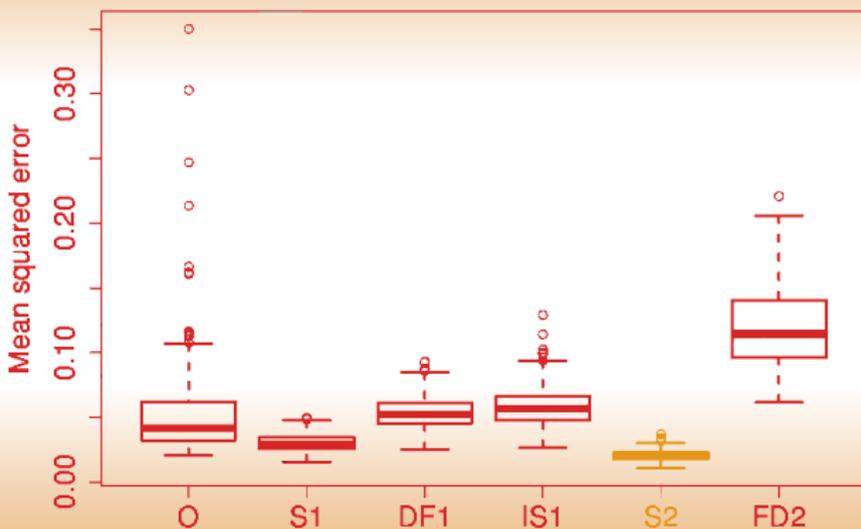
- **Split the data** into train/test sets (250 times);
- **Train** 250 regression functions for the 250 train sets (hyper-parameters were tuned by CV) with the predictors being
 - the original (sampled) functions \mathbf{X}_i (viewed as $\mathbb{R}^{|\tau_d|}$ vectors);
 - $\mathbf{Q}_{\lambda, \tau_d} \mathbf{X}_i^{\tau_d}$ for derivatives of order 1 or 2: **smoothing splines derivatives**;
 - $\mathbf{Q}_{0, \tau_d} \mathbf{X}_i^{\tau_d}$ for derivatives of order 1 or 2: **interpolating splines derivatives**;
 - derivatives of order 1 or 2 evaluated by $\frac{X_i(t_{j+1}) - X_i(t_j)}{t_{j+1} - t_j}$: **finite differences derivatives**;
- **Evaluate** these regression functions by calculating the **MSE** for the 50 corresponding test sets.



Examples

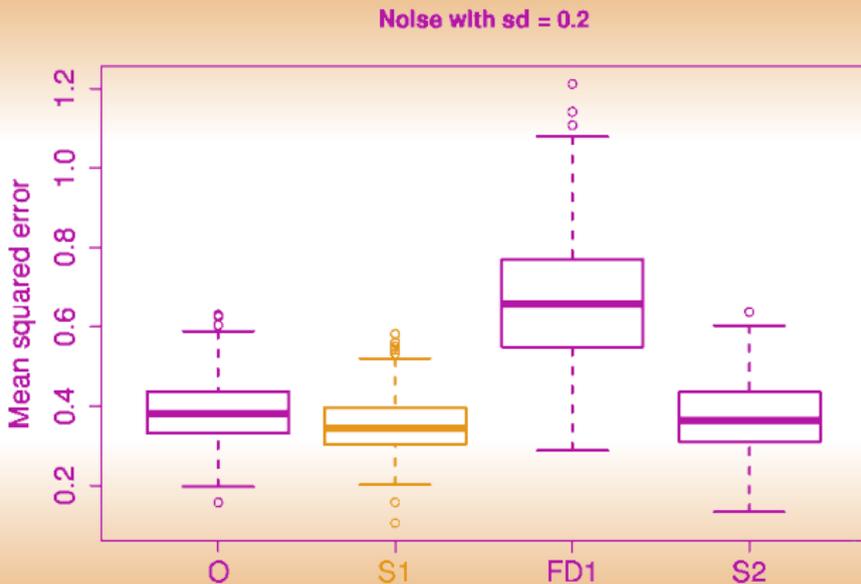
Performances

Noise with $sd = 0.01$





Performances



References



Berlinet, A. and Thomas-Agnan, C. (2004).

Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer Academic Publisher.



Faragó, T. and Györfi, L. (1975).

On the continuity of the error distortion function for multiple-hypothesis decisions.
IEEE Transactions on Information Theory, 21(4):458–460.



Kimeldorf, G. and Wahba, G. (1971).

Some results on Tchebycheffian spline functions.
Journal of Mathematical Analysis and Applications, 33(1):82–95.



Ragozin, D. (1983).

Error bounds for derivative estimation based on spline smoothing of exact or noisy data.
Journal of Approximation Theory, 37:335–355.



Steinwart, I. (2002).

Support vector machines are universally consistent.
Journal of Complexity, 18:768–791.

Any question?