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
Introduction and motivations

Outline

1. Introduction and motivations
2. A general consistency result
3. Examples
Introduction and motivations

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 \(Y \in \mathbb{R}\)
Introduction and motivations

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 \(Y \in \mathbb{R}\)
- \(X \in (\mathcal{X}, \langle ., . \rangle_\mathcal{X})\), an infinite dimensional Hilbert space.
Introduction and motivations

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 $Y \in \mathbb{R}$
- $X \in (X, \langle ., . \rangle_X)$, 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 $(X, Y)$. 
Introduction and motivations

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 $Y \in \mathbb{R}$
- $X \in (\mathcal{X}, \langle ., . \rangle_{\mathcal{X}})$, 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 $\phi_n : \mathcal{X} \to \{-1, 1\}$ or $\mathbb{R}$, that is universally consistent:

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

**Regression case**: $\lim_{n \to +\infty} E \left[ (\phi_n(X) - Y)^2 \right] = L^*$ where $L^* = \inf_{\phi : \mathcal{X} \to \mathbb{R}} E \left[ (\phi(X) - Y)^2 \right]$ will also be called the Bayes risk.
Introduction and motivations

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 \(Y \in \mathbb{R}\)
- \(X \in (\mathcal{X}, \langle ., . \rangle_\mathcal{X})\), 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: \(\lim_{n \rightarrow +\infty} \mathbb{E}\left([\phi_n(X) - Y]^2\right) = L^*\) where

\[ L^* = \inf_{\phi : \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}\left([\phi(X) - Y]^2\right) \]

will also be called the Bayes risk.
Predicting the rate of yellow berry in durum wheat from its NIR spectrum.
Introduction and motivations

Using derivatives

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

Using derivatives

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

Second derivative: infrared-spectrum of meat
Practically, $X^{(m)}$ is often more relevant than $X$ for the prediction. But $X \rightarrow X^{(m)}$ induces information loss and

\[
\inf_{\phi: D^{m}X \rightarrow \{-1, 1\}} P(\phi(X^{(m)}) \neq Y) \geq \inf_{\phi: X \rightarrow \{-1, 1\}} P(\phi(X) \neq Y) = L^* ~ \text{and} ~ \inf_{\phi: D^{m}X \rightarrow \mathbb{R}} \mathbb{E}\left([\phi(X^{(m)}) - Y]^2\right) \geq \inf_{\phi: X \rightarrow \mathbb{R}} P\left([\phi(X) - Y]^2\right) = L^*. 
\]
**Practically**, \((X_i)_i\) are not perfectly known; only a discrete sampling is given: 
\[X^\tau_d = (X_i(t))_{t \in \tau_d},\]
where 
\[\tau_d = \{t^\tau_d_1, \ldots, t^\tau_d_{|\tau_d|}\}.\]
**Introduction and motivations**

**Sampled functions**

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

The sampling can be non uniform...
Practically, \((X_i)_i\) are not perfectly known; only a discrete sampling is given: \(X^\tau_d_i = (X_i(t))_{t \in \tau_d}\) where \(\tau_d = \{t_1^\tau_d, \ldots, t_{|\tau_d|}^\tau_d\}\).

\[
\begin{align*}
\text{Non uniform sampling,} \\
\text{noisy data}
\end{align*}
\]

... and the data can be corrupted by noise.
**Introduction and motivations**

**Sampled functions**

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

Then, \(X_i^{(m)}\) is estimated from \(X_{i}^{\tau_d}\), by \(\hat{X}_{\tau_d}^{(m)}\), which also induces information loss:

\[
\inf_{\phi: D^mX \rightarrow \{-1,1\}} \mathbb{P}\left(\phi(\hat{X}_{\tau_d}^{(m)}) \neq Y\right) \geq \inf_{\phi: D^mX \rightarrow \{-1,1\}} \mathbb{P}\left(\phi(X^{(m)}) \neq Y\right) \geq L^*
\]

and

\[
\inf_{\phi: D^mX \rightarrow \mathbb{R}} \mathbb{E}\left(\left[\phi(\hat{X}_{\tau_d}^{(m)}) - Y\right]^2\right) \geq \inf_{\phi: D^mX \rightarrow \mathbb{R}} \mathbb{E}\left(\left[\phi(X^{(m)}) - Y\right]^2\right) \geq L^*.
\]
Find a classifier or a regression function $\phi_{n,\tau_d}$ built from $\hat{X}^{(m)}_{\tau_d}$ such that the risk of $\phi_{n,\tau_d}$ asymptotically reaches the Bayes risk $L^*$:

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

or

$$\lim_{|\tau_d| \to +\infty} \lim_{n \to +\infty} E\left(\left[\phi_{n,\tau_d}(\hat{X}^{(m)}_{\tau_d}) - Y\right]^2\right) = L^*$$
Find a classifier or a regression function $\phi_{n,\tau_d}$ built from $\hat{X}_{\tau_d}^{(m)}$ such that the risk of $\phi_{n,\tau_d}$ \textbf{asymptotically reaches} the Bayes risk $L^*$:

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

or

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

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

1. Introduction and motivations
2. A general consistency result
3. Examples
A general consistency result

Basics about smoothing splines I

Suppose that $X$ is the Sobolev space

$$\mathcal{H}^m = \{ h \in L^2_\omega | \forall j = 1, \ldots, m, D^j h \text{ exists (weak sense) and } D^m h \in L^2 \}$$

Suppose that $X$ is the Sobolev space

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

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 $\text{Ker} B \cap \mathbb{P}^{m-1} = \{0\}$. 

A general consistency result

Basics about smoothing splines I

Suppose that $X$ is the Sobolev space

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

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 $\text{Ker}B \cap \mathbb{P}^{m-1} = \{0\}$. $(\mathcal{H}^m, \langle ., . \rangle_{\mathcal{H}^m})$ 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

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

and

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

A general consistency result

Basics about smoothing splines II

A simple example of boundary conditions:

\[ h(0) = h^{(1)}(0) = \ldots = h^{(m-1)}(0) = 0. \]

Then,

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

and

\[ k_1(s, t) = \int_0^1 \frac{(t - w)^{m-1}(s - w)^{m-1}}{(m - 1)!} dw. \]
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 $t \in \tau_d$
A general consistency result

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 $h \to h(t)$ for all $t \in \tau_d$

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

solution of

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

and $\hat{x}_{\lambda, \tau_d} = S_{\lambda, \tau_d} x^{\tau_d}$ where $S_{\lambda, \tau_d} : \mathbb{R}^{|\tau_d|} \to \mathcal{H}^m$. 
A general consistency result

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 $x^{\tau_d}$ in $\mathbb{R}^{\tau_d}$, $\exists ! \hat{x}_{\lambda, \tau_d} \in H^m$

solution of

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

and $\hat{x}_{\lambda, \tau_d} = S_{\lambda, \tau_d} x^{\tau_d}$ where $S_{\lambda, \tau_d} : \mathbb{R}^{\tau_d} \rightarrow H^m$.

These assumptions are fulfilled by the previous simple example as long as $0 \not\in \tau_d$. 

Nathalie Villa-Vialaneix
A general consistency result

Estimating the predictors with smoothing splines II

$S_{\lambda, \tau_d}$ is given by:

\[
S_{\lambda, \tau_d} = \omega^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|}) U^T)^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} \\
+ \eta^T (K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} (\mathbb{I}_{|\tau_d|} - U^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1}) U \omega \\
= \omega^T M_0 + \eta^T M_1
\]

with

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

Estimating the predictors with smoothing splines II

$S_{\lambda,\tau_d}$ is given by:

$$S_{\lambda,\tau_d} = \omega^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|}) U^T)^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1}$$

$$+ \eta^T (K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} (\mathbb{I}_{|\tau_d|} - U^T (U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1} U(K_1 + \lambda \mathbb{I}_{|\tau_d|})^{-1})$$

$$= \omega^T M_0 + \eta^T M_1$$

with

- $\{\omega_1, \ldots, \omega_m\}$ is a basis of $\mathbb{P}^{m-1}$, $\omega = (\omega_1, \ldots, \omega_m)^T$ and $U = (\omega_i(t))_{i=1,\ldots,m \ t \in \tau_d}$;

- $\eta = (k_1(t,.))^T_{t \in \tau_d}$ 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 $X^{\tau_d}$ by $\hat{X}_{\lambda,\tau_d}$.
A general consistency result

Two important consequences

1. No information loss

\[
\inf_{\phi: H^m \to \{-1, 1\}} P(\phi(\tilde{X}_{\lambda, \tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \to \{-1, 1\}} P(\phi(X^{\tau_d}) \neq Y)
\]

and

\[
\inf_{\phi: H^m \to \{-1, 1\}} \mathbb{E}\left( [\phi(\tilde{X}_{\lambda, \tau_d}) - Y]^2 \right) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \to \{-1, 1\}} P\left( [\phi(X^{\tau_d}) - Y]^2 \right)
\]
A general consistency result

Two important consequences

1. **No information loss**

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{P}(\phi(\hat{X}_{\lambda,\tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{\lvert \tau_d \rvert} \to \{-1,1\}} \mathbb{P}(\phi(X_{\tau_d}) \neq Y)
\]

and

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{E}\left( [\phi(\hat{X}_{\lambda,\tau_d}) - Y]^2 \right) = \inf_{\phi: \mathbb{R}^{\lvert \tau_d \rvert} \to \{-1,1\}} \mathbb{P}\left( [\phi(X_{\tau_d}) - Y]^2 \right)
\]

2. **Easy way to use derivatives:**

\[
\langle S_{\lambda,\tau_d} u_{\tau_d}, S_{\lambda,\tau_d} v_{\tau_d} \rangle_{\mathcal{H}^m} = \langle \hat{u}_{\lambda,\tau_d}, \hat{v}_{\lambda,\tau_d} \rangle_{\mathcal{H}^m}
\]

Remark: \( Q_{\lambda,\tau_d} \) is calculated only from the RKHS, \( \lambda \) and \( \tau_d \); it does not depend on the data set.
A general consistency result

Two important consequences

1. No information loss

\[ \inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{P}(\phi(\hat{X}_{\lambda,\tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \to \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^\tau_d) \neq Y) \]

and

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

2. Easy way to use derivatives:

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

where \( K_1, M_0 \) and \( M_1 \) have been previously defined and

\[ W = (\langle \omega_i, \omega_j \rangle_{\mathcal{H}^m})_{i,j=1,...,m}. \]
A general consistency result

Two important consequences

1. **No information loss**

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{P}(\phi(\hat{X}_{\lambda,\tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{\mid\tau_d\mid} \to \{-1,1\}} \mathbb{P}(\phi(X^{\tau_d}) \neq Y)
\]

and

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{E}
\left(\left[\phi(\hat{X}_{\lambda,\tau_d}) - Y\right]^2\right) = \inf_{\phi: \mathbb{R}^{\mid\tau_d\mid} \to \{-1,1\}} \mathbb{P}\left(\left[\phi(X^{\tau_d}) - Y\right]^2\right)
\]

2. **Easy way to use derivatives:**

\[
(u^{\tau_d})^T M_{\lambda,\tau_d} v^{\tau_d} = \langle \hat{u}_{\lambda,\tau_d}, \hat{v}_{\lambda,\tau_d} \rangle_{\mathcal{H}^m}
\]

where \(M_{\lambda,\tau_d}\) is symmetric, definite positive.
A general consistency result

Two important consequences

1. **No information loss**

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{P} \left( \phi(\hat{X}_{\lambda,\tau_d}) \neq Y \right) = \inf_{\phi: \mathbb{R}^{1\tau_d} \to \{-1,1\}} \mathbb{P} \left( \phi(X_{\tau_d}) \neq Y \right)
\]

and

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{E} \left( \left[ \phi(\hat{X}_{\lambda,\tau_d}) - Y \right]^2 \right) = \inf_{\phi: \mathbb{R}^{1\tau_d} \to \{-1,1\}} \mathbb{P} \left( \left[ \phi(X_{\tau_d}) - Y \right]^2 \right)
\]

2. **Easy way to use derivatives:**

\[
(Q_{\lambda,\tau_d} u^{\tau_d})^T (Q_{\lambda,\tau_d} v^{\tau_d}) = \langle \hat{u}_{\lambda,\tau_d}, \hat{v}_{\lambda,\tau_d} \rangle_{\mathcal{H}^m}
\]

where \(Q_{\lambda,\tau_d}\) is the Choleski triangle of \(M_{\lambda,\tau_d} : Q_{\lambda,\tau_d}^T Q_{\lambda,\tau_d} = M_{\lambda,\tau_d}\).

**Remark:** \(Q_{\lambda,\tau_d}\) is calculated only from the RKHS, \(\lambda\) and \(\tau_d\): it does not depend on the data set.
A general consistency result

Two important consequences

1. **No information loss**

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{P}(\phi(\hat{X}_{\lambda,\tau_d}) \neq Y) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \to \{-1,1\}} \mathbb{P}(\phi(X^{\tau_d}) \neq Y)
\]

and

\[
\inf_{\phi: \mathcal{H}^m \to \{-1,1\}} \mathbb{E}\left(\left[\phi(\hat{X}_{\lambda,\tau_d}) - Y\right]^2\right) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \to \{-1,1\}} \mathbb{P}\left([\phi(X^{\tau_d}) - Y]^2\right)
\]

2. **Easy way to use derivatives:**

\[
(Q_{\lambda,\tau_d} u^{\tau_d})^T (Q_{\lambda,\tau_d} v^{\tau_d}) = \langle \hat{u}_{\lambda,\tau_d}, \hat{v}_{\lambda,\tau_d} \rangle_{\mathcal{H}^m}
\]

\[
\simeq \langle u^{(m)}_{\lambda,\tau_d}, v^{(m)}_{\lambda,\tau_d} \rangle_{L^2}
\]

where \(Q_{\lambda,\tau_d}\) is the Choleski triangle of \(M_{\lambda,\tau_d}: Q_{\lambda,\tau_d}^T Q_{\lambda,\tau_d} = M_{\lambda,\tau_d}\).

**Remark:** \(Q_{\lambda,\tau_d}\) is calculated only from the RKHS, \(\lambda\) and \(\tau_d\): it does not depend on the data set.
A general consistency result

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

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

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

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 $Q_{\lambda,\tau_d}$:

**Example:** Nonparametric kernel regression

$$\phi_{n,d} = \psi \circ Q_{\lambda,\tau_d} : x \in \mathcal{H}^m \rightarrow \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\|Q_{\lambda,\tau_d} x_{\tau_d} - Q_{\lambda,\tau_d} X_{\tau_d}^i\|_{\mathbb{R}^{\tau_d}}}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{\|Q_{\lambda,\tau_d} x_{\tau_d} - Q_{\lambda,\tau_d} X_{\tau_d}^i\|_{\mathbb{R}^{\tau_d}}}{h_n}\right)}$$
A general consistency result

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 $Q_{\lambda,\tau_d}$:

**Example:** Nonparametric kernel regression

$$
\phi_{n,d} = \psi \circ Q_{\lambda,\tau_d} : x \in \mathcal{H}^m \rightarrow \frac{\sum_{i=1}^{n} Y_i K \left( \frac{\|Q_{\lambda,\tau_d} x^{\tau_d} - Q_{\lambda,\tau_d} X^{\tau_d}_i \|_{\mathbb{R}^{|\tau_d|}}} {h_n} \right)} {\sum_{i=1}^{n} K \left( \frac{\|Q_{\lambda,\tau_d} x^{\tau_d} - Q_{\lambda,\tau_d} X^{\tau_d}_i \|_{\mathbb{R}^{|\tau_d|}}} {h_n} \right)}
$$
A general consistency result

Remark for consistency

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

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

where \( L_d^* = \inf_{\phi: \mathbb{R}^{\tau_d} \rightarrow \{-1, 1\}} P(\phi(X^{\tau_d}) \neq Y). \)
A general consistency result

Remark for consistency

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

\[ P\left( \phi_{n,\tau_d}(\hat{X}_{\lambda,\tau_d}) \neq Y \right) - L^* = P\left( \phi_{n,\tau_d}(\hat{X}_{\lambda,\tau_d}) \neq Y \right) - L^*_d + L^*_d - L^* \]

where \( L^*_d = \inf_{\phi: \mathbb{R}^{|\tau_d|} \to \{-1,1\}} P(\phi(X^{\tau_d}) \neq Y). \)

1. For all fixed \( d, \)

\[ \lim_{n \to +\infty} P\left( \phi_{n,\tau_d}(\hat{X}_{\lambda,\tau_d}) \neq Y \right) = L^*_d \]

as long as the \( \mathbb{R}^{|\tau_d|}\)-classifier is consistent because there is a one-to-one mapping between \( X^{\tau_d} \) and \( \hat{X}_{\lambda,\tau_d}. \)
A general consistency result

Remark for consistency

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

\[ P\left( \phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y \right) - L^* = P\left( \phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y \right) - L^*_d + L^*_d - L^* \]

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

1. For all fixed \( d \),

\[ \lim_{n \rightarrow +\infty} P\left( \phi_{n,\tau_d}(\widehat{X}_{\lambda,\tau_d}) \neq Y \right) = L^*_d \]

as long as the \( \mathbb{R}^{|\tau_d|} \)-classifier is consistent because there is a one-to-one mapping between \( X^{\tau_d} \) 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 \( \mathbb{E}(Y|X = .) \), consistency would be proved.
A general consistency result

Remark for consistency

**Classification case** (approximatively the same is true for regression):

\[
P(\phi_{n,\tau_d}(\hat{X}_{\lambda,\tau_d}) \neq Y) - L^* = P(\phi_{n,\tau_d}(\hat{X}_{\lambda,\tau_d}) \neq Y) - L_d^* + L_d^* - L^*
\]

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

1. For all fixed \( d \),

\[
\lim_{n \to +\infty} P(\phi_{n,\tau_d}(\hat{X}_{\lambda,\tau_d}) \neq Y) = L_d^*
\]

as long as the \( R^{\mid\tau_d\mid} \)-classifier is consistent because there is a one-to-one mapping between \( X^{\tau_d} \) and \( \hat{X}_{\lambda,\tau_d} \).

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

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

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

Spline consistency

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

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

**Assumption (A2)**

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

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

$$\|\hat{x}_{\lambda_d, \tau_d} - x\|_2 \leq (A R, m \lambda_d + B R, m 1^{m \underline{\Delta}_{\tau_d}}) \|D^m x\|_2 L^d \to +\infty \quad \Rightarrow \quad 0$$
A general consistency result

Spline consistency

Let \( \lambda \) depends on \( d \) and denote \( (\lambda_d)_d \) the series of regularization parameters. Also introduce
\[
\overline{\Delta}_{\tau_d} := \max\{t_1, t_2 - t_1, \ldots, 1 - t_{|\tau_d|}\}, \quad \Delta_{\tau_d} := \min_{1 \leq i < |\tau_d|}\{t_{i+1} - t_i\}
\]

**Assumption (A2)**

- \( \exists R \) such that \( \overline{\Delta}_{\tau_d}/\Delta_{\tau_d} \leq R \) for all \( d \);
- \( \lim_{d \to +\infty} |\tau_d| = +\infty \);
- \( \lim_{d \to +\infty} \lambda_d = 0 \).

**[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 \( \lambda_d > 0 \),
\[
\left\| \hat{x}_{\lambda_d, \tau_d} - x \right\|^2_{L^2} \leq \left( A_{R,m} \lambda_d + B_{R,m} \frac{1}{|\tau_d|^{2m}} \right) \left\| D^m x \right\|^2_{L^2} \xrightarrow{d \to +\infty} 0
\]
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)
\[ \mathbb{E}\left(\|D^m X\|_{L^2}^2\right) \text{ is finite and } Y \in \{-1, 1\}. \]

or

Assumption (A3b)
\[ \tau_d \subset \tau_{d+1} \text{ for all } d \text{ and } \mathbb{E}(Y^2) \text{ is finite.} \]
A general consistency result

Bayes risk consistency

Assumption (A3a)
\[ \mathbb{E}\left(\|D^m X\|_{L^2}^2\right) \text{ is finite and } Y \in \{-1, 1\}. \]

or

Assumption (A3b)
\[ \tau_d \subset \tau_{d+1} \text{ for all } d \text{ and } \mathbb{E}(Y^2) \text{ is finite.} \]

Under (A1)-(A3), \( \lim_{d \to +\infty} L_d^* = L^* \).
A general consistency result

Proof under assumption (A3a)

Assumption (A3a)

\[ \mathbb{E} \left( \| D^m X \|^2_L \right) \text{ is finite and } Y \in \{-1, 1\}. \]
Assumption (A3a)

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

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

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

\[ \mathbb{E}(\delta(T_d(X), X)) \xrightarrow{d \to +\infty} 0 \]

then \(\lim_{d \to +\infty} \inf_{\phi : X \to \{-1, 1\}} \mathbb{P}(\phi(T_d(X)) \neq Y) = L^*\).
A general consistency result

Proof under assumption (A3a)

Assumption (A3a)

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

The proof is based on a result of [Faragó and Győrfi, 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 Gyorfi’s Theorem.

Then the result follows.
A general consistency result

Proof under assumption (A3b)

Assumption (A3b)

\[ \tau_d \subset \tau_{d+1} \text{ for all } d \text{ and } \mathbb{E}(Y^2) \text{ is finite.} \]
A general consistency result

Proof under assumption (A3b)

Assumption (A3b)

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

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

Concluding result (consistency)

Theorem

Under assumptions (A1)-(A3),

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

and

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

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 general consistency result

A practical application to SVM I

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

\[
u \in \mathbb{R}^p \rightarrow \text{Sign}\left(\sum_{i=1}^{n} \alpha_i T_i e^{-\gamma \|u-U_i\|_2^2_{\mathbb{R}^p}}\right)\]

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

\[
\arg\min_w \sum_{i=1}^{n} \left|1 - T_i w(U_i)\right|_+ + C \|w\|^2_S
\]

where \(w(u) = \sum_{i=1}^{n} \alpha_i e^{-\gamma \|u-U_i\|_2^2_{\mathbb{R}^p}}\) and \(S\) is the RKHS associated with the gaussian kernel and \(C\) is a regularization parameter.
A general consistency result

A practical application to SVM I

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

\[
u \in \mathbb{R}^p \rightarrow \text{Sign}\left( \sum_{i=1}^{n} \alpha_i T_i e^{-\gamma \|u-U_i\|_2^2} \right)
\]

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

\[
\arg\min_w \sum_{i=1}^{n} \left| 1 - T_i w(U_i) \right|_+ + C \|w\|_S^2
\]

where \( w(u) = \sum_{i=1}^{n} \alpha_i e^{-\gamma \|u-U_i\|_2^2} \) and \( S \) is the RKHS associated with the Gaussian kernel and \( C \) is a regularization parameter.

A general consistency result

A practical application to SVM II

**Additional assumptions related to SVM: Assumptions (A4)**

- For all $d$, the regularization parameter depends on $n$ such that
  \[
  \lim_{n \to +\infty} nC_n^d = +\infty \quad \text{and} \quad C_n^d = O_n\left(n^{\beta_d - 1}\right) \quad \text{for a } 0 < \beta_d < 1/d.
  \]

- For all $d$, there is a bounded subset of $\mathbb{R}^{|\tau_d|}$, $\mathcal{B}_d$, such that $X^{\tau_d}$ belongs to $\mathcal{B}_d$. 

A general consistency result

A practical application to SVM II

Additional assumptions related to SVM: Assumptions (A4)

- For all $d$, the regularization parameter depends on $n$ such that
  \[
  \lim_{n \to +\infty} nC_n^d = +\infty \quad \text{and} \quad C_n^d = O_n\left(n^{\beta_d - 1}\right)
  \]
  for a $0 < \beta_d < 1/d$.

- For all $d$, there is a bounded subset of $\mathbb{R}^{|	au_d|}$, $\mathcal{B}_d$, such that $X_{\tau_d}$ belongs to $\mathcal{B}_d$.

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

\[
\text{Sign}\left(\sum_{i=1}^{n} \alpha_i Y_i e^{-\gamma \|Q_{\lambda_d,\tau_d} x_{\tau_d} - Q_{\lambda_d,\tau_d} X_{\tau_d}^i\|_2^2}\right) \approx \text{Sign}\left(\sum_{i=1}^{n} \alpha_i Y_i e^{-\gamma \|x^{(m)}_i - X^{(m)}_i\|_{L2}^2}\right)
\]

is consistent: $\lim_{|	au_d| \to +\infty} \lim_{n \to +\infty} \mathbb{P}\left(\phi_{n,\tau_d}(\widehat{X}_{\lambda_d,\tau_d}) \neq Y\right) = L^*$.
A general consistency result

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

Under suitable (and usual) regularity assumptions on $\mathbb{E}(Y|X=.)$ 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 $\nu$ is either equal to $m$ or to a Lipchitz constant related to $\mathbb{E}(Y|X=.)$. 
Outline

1 Introduction and motivations

2 A general consistency result

3 Examples
Examples

Chosen regression method: Regression with kernel ridge regression

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

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

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 \( \mathbb{R}^p \times \mathbb{R} \).
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 $\mathbb{R}^p \times \mathbb{R}$. In the following examples, $U_i$ is either:

- the original (sampled) functions $X_i$ (viewed as $\mathbb{R}^{|	au_d|}$ vectors);
- $Q_{\lambda, \tau_d} X_i^{\tau_d}$ 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.

The differences are significant between \( G(2) / G(1) \) and between \( G(1) / G(2) \).
Examples

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

<table>
<thead>
<tr>
<th>Kernel (SVM)</th>
<th>MSE on test (and sd ×10^{-3})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear (L)</td>
<td>0.122 (8.77)</td>
</tr>
<tr>
<td>Linear on derivatives (L^{(1)})</td>
<td>0.138 (9.53)</td>
</tr>
<tr>
<td>Linear on second derivatives (L^{(2)})</td>
<td>0.122 (1.71)</td>
</tr>
<tr>
<td>Gaussian (G)</td>
<td>0.110 (20.2)</td>
</tr>
<tr>
<td>Gaussian on derivatives (G^{(1)})</td>
<td>0.098 (7.92)</td>
</tr>
<tr>
<td><strong>Gaussian on second derivatives (G^{(2)})</strong></td>
<td><strong>0.094 (8.35)</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>MSE (mean)</th>
<th>MSE (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>0.154</td>
<td>0.012</td>
</tr>
<tr>
<td>Kernel PLS</td>
<td>0.154</td>
<td>0.013</td>
</tr>
<tr>
<td>KRR splines (reg. $D^2$)</td>
<td>0.094</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Error decrease: almost 40 %
Example 2: Simulated noisy spectra

Original data:

Variable to predict: Fat content of pieces of meat.
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.:
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 $X_i$ (viewed as $\mathbb{R}^{\tau_d}$ vectors);
  - $Q_{\lambda,\tau_d} X^\tau_{i}$ for derivatives of order 1 or 2: **smoothing splines derivatives**;
  - $Q_{0,\tau_d} X^\tau_{i}$ 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
Examples

Performances

Noise with sd = 0.2

[Box plots showing mean squared error for different conditions (O, S1, FD1, S2)].
References

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

On the continuity of the error distortion function for multiple-hypothesis decisions.

Some results on Tchebycheffian spline functions.

Error bounds for derivative estimation based on spline smoothing of exact or noisy data.

Support vector machines are universally consistent.

Any question?