

Consistency of Functional Learning Methods Based on Derivatives

Fabrice Rossi^a, Nathalie Villa-Vialaneix^{b,c,*}

^a*Télécom ParisTech, LTCI - UMR CNRS 5141, France*

^b*IUT de Perpignan (Dpt STID, Carcassonne), Université de Perpignan Via Domitia, France*

^c*Institut de Mathématiques de Toulouse, Université de Toulouse, France*

Abstract

In some real world applications, such as spectrometry, functional models achieve better predictive performances if they work on the derivatives of order m of their inputs rather than on the original functions. As a consequence, the use of derivatives is a common practice in functional data analysis, despite a lack of theoretical guarantees on the asymptotically achievable performances of a derivative based model. In this paper, we show that a smoothing spline approach can be used to preprocess multivariate observations obtained by sampling functions on a discrete and finite sampling grid in a way that leads to a consistent scheme on the original infinite dimensional functional problem. This work extends Mas and Pumo (2009) to nonparametric approaches and incomplete knowledge. To be more precise, the paper tackles two difficulties in a nonparametric framework: the information loss due to the use of the derivatives instead of the original functions and the information loss due to the fact that the functions are observed through a discrete sampling and are thus also imperfectly known: the use of a smoothing spline based approach solves these two problems. Finally, the proposed approach is tested on two real world datasets and the approach is experimentally proven to be a good solution in the case of noisy functional predictors.

Keywords: Functional Data Analysis, Consistency, Statistical learning, Derivatives, SVM, Smoothing splines, RKHS, Kernel

*Corresponding author.

Email addresses: `Fabrice.Rossi@telecom-paristech.fr` (Fabrice Rossi),
`nathalie.villa@math.univ-toulouse.fr` (Nathalie Villa-Vialaneix)

1. Introduction

As the measurement techniques are developing, more and more data are high dimensional vectors generated by measuring a continuous process on a discrete sampling grid. Many examples of this type of data can be found in real world applications, in various fields such as spectrometry, voice recognition, time series analysis, etc.

Data of this type should not be handled in the same way as standard multivariate observations but rather analysed as *functional* data: each observation is a function coming from an input space with infinite dimension, sampled on a high resolution sampling grid. This leads to a large number of variables, generally more than the number of observations. Moreover, functional data are frequently smooth and generate highly correlated variables as a consequence. Applied to the obtained high dimensional vectors, classical statistical methods (e.g., linear regression, factor analysis) often lead to ill-posed problems, especially when a covariance matrix has to be inverted (this is the case, e.g., in linear regression, in discriminant analysis and also in sliced inverse regression). Indeed, the number of observed values for each function is generally larger than the number of functions itself and these values are often strongly correlated. As a consequence, when these data are considered as multidimensional vectors, the covariance matrix is ill-conditioned and leads to unstable and unaccurate solutions in models where its inverse is required. Thus, these methods cannot be directly used. During past years, several methods have been adapted to that particular context and grouped under the generic name of Functional Data Analysis (FDA) methods. Seminal works focused on linear methods such as factorial analysis (Deville (1974); Dauxois and Pousse (1976); Besse and Ramsay (1986); James et al. (2000), among others) and linear models Ramsay and Dalzell (1991); Cardot et al. (1999); James and Hastie (2001); a comprehensive presentation of linear FDA methods is given in Ramsay and Silverman (1997, 2002). More recently, nonlinear functional models have been extensively developed and include generalized linear models James (2002); James and Silverman (2005), kernel nonparametric regression Ferraty and Vieu (2006), Functional Inverse Regression Ferré and Yao (2003), neural networks Rossi and Conan-Guez (2005); Rossi et al. (2005), k -nearest neighbors Biau et al. (2005); Laloë (2008), Support Vector Machines (SVM), Rossi and Villa (2006), among a very large variety

36 of methods.

37 In previous works, numerous authors have shown that the derivatives
38 of the functions lead sometimes to better predictive performances than the
39 functions themselves in inference tasks, as they provide information about
40 the shape or the regularity of the function. In particular applications such
41 as spectrometry Ferraty and Vieu (2006); Rossi et al. (2005); Rossi and Villa
42 (2006), micro-array data Dejean et al. (2007) and handwriting recognition
43 Williams et al. (2006); Bahlmann and Burkhardt (2004), these characteristics
44 lead to accurate predictive models. But, on a theoretical point of the view,
45 limited results about the effect of the use of the derivatives instead of the
46 original functions are available: Mas and Pumo (2009) studies this problem
47 for a linear model built on the first derivatives of the functions. In the present
48 paper, we also focus on the theoretical relevance of this common practice and
49 extend Mas and Pumo (2009) to nonparametric approaches and incomplete
50 knowledge.

51 More precisely, we address the problem of the estimation of the condi-
52 tional expectation $\mathbb{E}(Y|X)$ of a random variable Y given a functional random
53 variable X . Y is assumed to be either real valued (leading to a regression
54 problem) or to take values in $\{-1, 1\}$ (leading to a binary classification prob-
55 lem). We target two theoretical difficulties. The first difficulty is the po-
56 tential information loss induced by using a derivative instead of the original
57 function: when one replaces X by its order m derivative $X^{(m)}$, consistent
58 estimators (such as kernel models Ferraty and Vieu (2006)) guarantee an
59 asymptotic estimation of $\mathbb{E}(Y|X^{(m)})$ but cannot be used directly to address
60 the original problem, namely estimating $\mathbb{E}(Y|X)$. This is a simple conse-
61 quence of the fact that $X \mapsto X^{(m)}$ is not a one to one mapping. The second
62 difficulty is induced by sampling: in practice, functions are never observed
63 exactly but rather, as explained above, sampled on a discrete sampling grid.
64 As a consequence, one relies on approximate derivatives, $\widehat{X}_\tau^{(m)}$ (where τ de-
65 notes the sampling grid). This approach induces even more information loss
66 with respect to the underlying functional variable X : in general, a consistent
67 estimator of $\mathbb{E}(Y|\widehat{X}_\tau^{(m)})$ will not provide a consistent estimation of $\mathbb{E}(Y|X)$
68 and the optimal predictive performances for Y given $\widehat{X}_\tau^{(m)}$ will be lower than
69 the optimal predictive performances for Y given X .

70 We show in this paper that the use of a smoothing spline based approach
71 solves both problems. Smoothing splines are used to estimate the functions
72 from their sampled version in a convergent way. In addition, properties of

73 splines are used to obtain estimates of the derivatives of the functions with no
74 induced information loss. Both aspects are implemented as a preprocessing
75 step applied to the multivariate observations generated via the sampling grid.
76 The preprocessed observations can then be fed into any finite dimensional
77 consistent regression estimator or classifier, leading to a consistent estima-
78 tor for the original infinite dimensional problem (in real world applications,
79 we instantiate the general scheme in the particular case of kernel machines
80 Shawe-Taylor and Cristianini (2004)).

81 The remainder of the paper is organized as follows: Section 2 introduces
82 the model, the main smoothness assumption and the notations. Section 3
83 recalls important properties of spline smoothing. Section 4 presents approx-
84 imation results used to build a general consistent classifier or a general con-
85 sistent regression estimator in Section 5. Finally, Section 6 illustrates the
86 behavior of the proposed method for two real world spectrometric problems.
87 The proofs are given at the end of the article.

88 2. Setup and notations

89 2.1. Consistent classifiers and regression functions

90 We consider a pair of random variables (X, Y) where X takes values in
91 a functional space \mathcal{X} and Y is either a real valued random variable (regres-
92 sion case) or a random variable taking values in $\{-1, 1\}$ (binary classifica-
93 tion case). From this, we are given a learning set $S_n = \{(X_i, Y_i)\}_{i=1}^n$ of n
94 independent copies of (X, Y) . Moreover, the functions X_i are not entirely
95 known but sampled according to a non random sampling grid of finite length,
96 $\tau_d = (t_l)_{l=1}^{|\tau_d|}$: we only observe $\mathbf{X}_i^{\tau_d} = (X_i(t_1), \dots, X_i(t_{|\tau_d|}))^T$, a vector of $\mathbb{R}^{|\tau_d|}$
97 and denote S_{n, τ_d} the corresponding learning set. Our goal is to construct:

- 98 1. *in the binary classification case*: a classifier, ϕ_{n, τ_d} , whose misclassifica-
99 tion probability

$$L(\phi_{n, \tau_d}) = \mathbb{P}(\phi_{n, \tau_d}(\mathbf{X}^{\tau_d}) \neq Y)$$

100 asymptotically reaches the Bayes risk

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

101 i.e., $\lim_{|\tau_d| \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{E}(L(\phi_{n, \tau_d})) = L^*$;

102 2. *in the regression case*: a regression function, ϕ_{n,τ_d} , whose L^2 error

$$L(\phi_{n,\tau_d}) = \mathbb{E}([\phi_{n,\tau_d}(\mathbf{X}^{\tau_d}) - Y]^2)$$

103 asymptotically reaches the minimal L^2 error

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

104 i.e., $\lim_{|\tau_d|\rightarrow+\infty} \lim_{n\rightarrow+\infty} L(\phi_{n,\tau_d}) = L^*$.

105 This definition implicitly requires $\mathbb{E}(Y^2) < \infty$ and as a consequence,
 106 corresponds to a L^2 convergence of ϕ_{n,τ_d} to the conditional expectation
 107 $\phi^* = \mathbb{E}(Y|X)$, i.e., to $\lim_{|\tau_d|\rightarrow+\infty} \lim_{n\rightarrow+\infty} \mathbb{E}([\phi_{n,\tau_d}(\mathbf{X}^{\tau_d}) - \phi^*(X)]^2) =$
 108 0.

109 Such ϕ_{n,τ_d} are said to be (*weakly consistent*) Devroye et al. (1996); Györfi
 110 et al. (2002). We have deliberately used the same notations for the (optimal)
 111 predictive performances in both the binary classification and the regression
 112 case. We will call L^* the Bayes risk even in the case of regression. Most of
 113 the theoretical background of this paper is common to both the regression
 114 case and the classification case: the distinction between both cases will be
 115 made only when necessary.

116 As pointed out in the introduction, the main difficulty is to show that
 117 the performances of a model built on the $\mathbf{X}_i^{\tau_d}$ asymptotically reach the best
 118 performance achievable on the original functions X_i . In addition, we will
 119 build the model on derivatives estimated from the $\mathbf{X}_i^{\tau_d}$.

120 2.2. Smoothness assumption

121 Our goal is to leverage the functional nature of the data by allow-
 122 ing differentiation operators to be applied to functions prior their submis-
 123 sion to a more common classifier or regression function. Therefore we as-
 124 sume that the functional space \mathcal{X} contains only differentiable functions.
 125 More precisely, \mathcal{X} is the Sobolev space $\mathcal{H}^m = \left\{ h \in L^2([0, 1]) \mid \forall j = \right.$
 126 $1, \dots, m, D^j h$ exists in the weak sense, and $D^m h \in L^2([0, 1]) \left. \right\}$, where $D^j h$
 127 is the j -th derivative of h (also denoted by $h^{(j)}$) and for an integer $m \geq 1$.
 128 Of course, by a straightforward generalization, any bounded interval can be
 129 considered instead of $[0, 1]$.

130 To estimate the underlying functions X_i and their derivatives from sam-
 131 pled data, we rely on smoothing splines. More precisely, let us consider
 132 a deterministic function $x \in \mathcal{H}^m$ sampled on the aforementioned grid. A
 133 smoothing spline estimate of x is the solution, $\hat{x}_{\lambda, \tau_d}$, of

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

134 where λ is a regularization parameter that balances interpolation error and
 135 smoothness (measured by the L^2 norm of the m -th derivative of the esti-
 136 mate). The goal is to show that a classifier or a regression function built
 137 on $\hat{X}_{\lambda, \tau_d}^{(m)}$ is consistent for the original problem (i.e., the problem defined by
 138 the pair (X, Y)): this means that using $\hat{X}_{\lambda, \tau_d}^{(m)}$ instead of X has no dramatic
 139 consequences on the accuracy of the classifier or of the regression function.
 140 In other words, asymptotically, no information loss occurs when one replaces
 141 X by $\hat{X}_{\lambda, \tau_d}^{(m)}$.

142 The proof is based on the following steps:

- 143 1. First, we show that building a classifier or a regression function on
 144 $\hat{X}_{\lambda, \tau_d}^{(m)}$ is approximately equivalent to building a classifier or a regression
 145 function on $\mathbf{X}^{\tau_d} = (X(t_l))_{l=1}^{|\tau_d|}$ using a specific metric. This is done by
 146 leveraging the Reproducing Kernel Hilbert Space (RKHS) structure of
 147 \mathcal{H}^m . This part serves one main purpose: it provides a solution to work
 148 with estimation of the derivatives of the original function in a way
 149 that preserves all the information available in \mathbf{X}^{τ_d} . In other words, the
 150 best predictive performances for Y theoretically available by building a
 151 multivariate model on \mathbf{X}^{τ_d} are equal to the best predictive performances
 152 obtained by building a functional model on $\hat{X}_{\lambda, \tau_d}^{(m)}$.
- 153 2. Then, we link $\mathbb{E}(Y|\hat{X}_{\lambda, \tau_d})$ with $\mathbb{E}(Y|X)$ by approximation results
 154 available for smoothing splines. This part of the proof handles the
 155 effects of sampling.
- 156 3. Finally, we glue both results via standard $\mathbb{R}^{|\tau_d|}$ consistency results.

157 **3. Smoothing splines and differentiation operators**

158 *3.1. RKHS and smoothing splines*

159 As we want to work on derivatives of functions from \mathcal{H}^m , a natural inner
 160 product for two functions of \mathcal{H}^m would be $(u, v) \rightarrow \int_0^1 u^{(m)}(t)v^{(m)}(t)dt$.
 161 However, we prefer to use an inner product of \mathcal{H}^m ($\int_0^1 u^{(m)}(t)v^{(m)}(t)dt$ only
 162 induces a semi-norm on \mathcal{H}^m) because, as will be shown later, such an inner
 163 product is related to an inner product between the sampled functions
 164 considered as vectors of $\mathbb{R}^{|\tau_d|}$.

165 This can be done by decomposing \mathcal{H}^m into $\mathcal{H}^m = \mathcal{H}_0^m \oplus \mathcal{H}_1^m$ Kimeldorf
 166 and Wahba (1971), where $\mathcal{H}_0^m = \text{Ker}D^m = \mathbb{P}^{m-1}$ (the space of polynomial
 167 functions of degree less or equal to $m - 1$) and \mathcal{H}_1^m is an infinite dimensional
 168 subspace of \mathcal{H}^m defined via m boundary conditions. The boundary condi-
 169 tions are given by a full rank linear operator from \mathcal{H}^m to \mathbb{R}^m , denoted B , such
 170 that $\text{Ker}B \cap \mathbb{P}^{m-1} = \{0\}$. Classical examples of boundary conditions include
 171 the case of “natural splines” (for $m = 2$, $h(0) = h(1) = 0$) and constraints
 172 that target only the first values of h and its derivatives at a fixed position, for
 173 instance the conditions: $h(0) = \dots = h^{(m-1)}(0) = 0$. Other boundary condi-
 174 tions can be used Berlinet and Thomas-Agnan (2004); Besse and Ramsay
 175 (1986); Craven and Wahba (1978), depending on the application.

176 Once the boundary conditions are fixed, an inner product on both \mathcal{H}_0^m
 177 and \mathcal{H}_1^m can be defined:

$$\langle u, v \rangle_1 = \langle D^m u, D^m v \rangle_{L^2} = \int_0^1 u^{(m)}(t)v^{(m)}(t)dt$$

178 is an inner product on \mathcal{H}_1^m (as $h \in \mathcal{H}_1^m$ and $D^m h \equiv 0$ give $h \equiv 0$). Moreover,
 179 if we denote $B = (B^j)_{j=1}^m$, then $\langle u, v \rangle_0 = \sum_{j=1}^m B^j u B^j v$ is an inner product
 180 on \mathcal{H}_0^m . We obtain this way an inner product on \mathcal{H}^m given by

$$\begin{aligned} \langle u, v \rangle_{\mathcal{H}^m} &= \int_0^1 u^{(m)}(t)v^{(m)}(t)dt + \sum_{j=1}^m B^j u B^j v \\ &= \langle \mathcal{P}_1^m(u), \mathcal{P}_1^m(v) \rangle_1 + \langle \mathcal{P}_0^m(u), \mathcal{P}_0^m(v) \rangle_0 \end{aligned}$$

181 where \mathcal{P}_i^m is the projector on \mathcal{H}_i^m .

182 Equipped with $\langle \cdot, \cdot \rangle_{\mathcal{H}^m}$, \mathcal{H}^m is a Reproducing Kernel Hilbert Space
 183 (RKHS, see e.g. Berlinet and Thomas-Agnan (2004); Heckman and Ramsay
 184 (2000); Wahba (1990)). More precisely, it exists a kernel $k : [0, 1]^2 \rightarrow \mathbb{R}$ such

185 that, for all $u \in \mathcal{H}^m$ and all $t \in [0, 1]$, $\langle u, k(t, \cdot) \rangle_{\mathcal{H}^m} = u(t)$. The same occurs
 186 for \mathcal{H}_0^m and \mathcal{H}_1^m which respectively have reproducing kernels denoted by k_0
 187 and k_1 . We have $k = k_0 + k_1$.

188 In the most common cases, k_0 and k_1 have already been explicitly cal-
 189 culated (see e.g., Berlinet and Thomas-Agnan (2004), especially chapter 6,
 190 sections 1.1 and 1.6.2). For example, for $m \geq 1$ and the boundary conditions
 191 $h(0) = h'(0) = \dots = h^{(m-1)}(0) = 0$, we have:

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

192 and

$$k_1(s, t) = \int_0^1 \frac{(t-w)_+^{m-1} (s-w)_+^{m-1}}{(m-1)!^2} dw.$$

193 3.2. Computing the splines

194 We need now to compute to $\hat{x}_{\lambda, \tau_d}$ starting with $\mathbf{x}^{\tau_d} = (x(t))_{t \in \tau_d}^T$. This
 195 can be done via a theorem from Kimeldorf and Wahba (1971). We need the
 196 following compatibility assumptions between the sampling grid τ_d and the
 197 boundary conditions operator B :

198 **Assumption 1.** *The sampling grid $\tau_d = (t_l)_{l=1}^{|\tau_d|}$ is such that*

- 199 1. *sampling points are distinct in $[0, 1]$ and $|\tau_d| \geq m - 1$*
- 200 2. *the m boundary conditions B^j are linearly independent from the $|\tau_d|$*
 201 *linear forms $h \mapsto h(t_l)$, for $l = 1, \dots, |\tau_d|$ (defined on \mathcal{H}^m)*

202 Then $\hat{x}_{\lambda, \tau_d}$ and $\mathbf{x}^{\tau_d} = (x(t))_{t \in \tau_d}^T$ are linked by the following result:

203 **Theorem 1** (Kimeldorf and Wahba (1971)). *Under Assumption (A1), the*
 204 *unique solution $\hat{x}_{\lambda, \tau_d}$ to equation (1) is given by:*

$$\hat{x}_{\lambda, \tau_d} = \mathcal{S}_{\lambda, \tau_d} \mathbf{x}^{\tau_d}, \quad (2)$$

205 where $\mathcal{S}_{\lambda, \tau_d}$ is a full rank linear operator from $\mathbb{R}^{|\tau_d|}$ to \mathcal{H}^m defined by:

$$\mathcal{S}_{\lambda, \tau_d} = \omega^T M_0 + \eta^T M_1 \quad (3)$$

206 with

- 207 • $M_0 = (U(K_1 + \lambda I_d)^{-1} U^T)^{-1} U(K_1 + \lambda I_d)^{-1}$
- 208 • $M_1 = (K_1 + \lambda I_d)^{-1} (I_d - U^T M_0)$;
- 209 • $\{\omega_1, \dots, \omega_m\}$ is a basis of \mathbb{P}^{m-1} , $\omega = (\omega_1, \dots, \omega_m)^T$ and $U =$
210 $(\omega_i(t))_{i=1, \dots, m}^T_{t \in \tau_d}$;
- 211 • $\eta = (k_1(t, \cdot))_{t \in \tau_d}^T$ and $K_1 = (k_1(t, t'))_{t, t' \in \tau_d}$.

212 3.3. No information loss

213 The first important consequence of Theorem 1 is that building a model
214 on $\widehat{X}_{\lambda, \tau_d}$ or on \mathbf{X}^{τ_d} leads to the same optimal predictive performances (to the
215 same Bayes risk). This is formalized by the following corollary:

216 **Corollary 1.** *Under Assumption (A1), we have*

- 217 • *in the binary classification case:*

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1, 1\}} \mathbb{P} \left(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y \right) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1, 1\}} \mathbb{P} (\phi(\mathbf{X}^{\tau_d}) \neq Y) \quad (4)$$

- 218 • *in the regression case:*

$$\inf_{\phi: \mathcal{H}^m \rightarrow \mathbb{R}} \mathbb{E} \left(\left[\phi \left(\widehat{X}_{\lambda, \tau_d} \right) - Y \right]^2 \right) = \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \mathbb{R}} \mathbb{E} ([\phi(\mathbf{X}^{\tau_d}) - Y]^2) \quad (5)$$

219 3.4. Differentiation operator

220 The second important consequence of Theorem 1 is that the inner product
221 $\langle \cdot, \cdot \rangle_{\mathcal{H}^m}$ is equivalent to a specific inner product on $\mathbb{R}^{|\tau_d|}$ given in the following
222 corollary:

223 **Corollary 2.** *Under Assumption (A1) and for any $\mathbf{u}^{\tau_d} = (u(t))_{t \in \tau_d}^T$ and*
224 $\mathbf{v}^{\tau_d} = (v(t))_{t \in \tau_d}^T$ *in $\mathbb{R}^{|\tau_d|}$,*

$$\langle \widehat{u}_{\lambda, \tau_d}, \widehat{v}_{\lambda, \tau_d} \rangle_{\mathcal{H}^m} = (\mathbf{u}^{\tau_d})^T \mathbf{M}_{\lambda, \tau_d} \mathbf{v}^{\tau_d} \quad (6)$$

225 where $\mathbf{M}_{\lambda, \tau_d} = M_0^T W M_0 + M_1^T K_1 M_1$ with $W = (\langle w_i, w_j \rangle_0)_{i, j=1, \dots, m}$. The
226 matrix $\mathbf{M}_{\lambda, \tau_d}$ is symmetric and positive definite and defines an inner product
227 on $\mathbb{R}^{|\tau_d|}$.

228 The corollary is a direct consequence of equations (2) and (3).

229 In practice, the corollary means that the euclidean space $(\mathbb{R}^{|\tau_d|}, \langle \cdot, \cdot \rangle_{\mathbf{M}_{\lambda, \tau_d}})$
 230 is isomorphic to $(\mathcal{I}_{\lambda, \tau_d}, \langle \cdot, \cdot \rangle_{\mathcal{H}^m})$, where $\mathcal{I}_{\lambda, \tau_d}$ is the image of $\mathbb{R}^{|\tau_d|}$ by $\mathcal{S}_{\lambda, \tau_d}$. As
 231 a consequence, one can use the Hilbert structure of \mathcal{H}^m directly in $\mathbb{R}^{|\tau_d|}$ via
 232 $\mathbf{M}_{\lambda, \tau_d}$: as the inner product of \mathcal{H}^m is defined on the order m derivatives of the
 233 functions, this corresponds to using those derivatives instead of the original
 234 functions.

235 More precisely, let $\mathbf{Q}_{\lambda, \tau_d}$ be the transpose of the Cholesky triangle of $\mathbf{M}_{\lambda, \tau_d}$
 236 (given by the Cholesky decomposition $\mathbf{Q}_{\lambda, \tau_d}^T \mathbf{Q}_{\lambda, \tau_d} = \mathbf{M}_{\lambda, \tau_d}$). Corollary 2
 237 shows that $\mathbf{Q}_{\lambda, \tau_d}$ acts as an approximate differentiation operation on sampled
 238 functions.

239 Let us indeed consider an estimation method for multivariate inputs based
 240 only on inner products or norms (that are directly derived from the inner
 241 products), such as, e.g., Kernel Ridge Regression Saunders et al. (1998);
 242 Shawe-Taylor and Cristianini (2004). In this latter case, if a Gaussian kernel
 243 is used, the regression function has the following form:

$$u \mapsto \sum_{i=1}^n T_i \alpha_i e^{-\gamma \|U_i - u\|_{\mathbb{R}^p}^2} \quad (7)$$

244 where $(U_i, T_i)_{1 \leq i \leq n}$ are learning examples in $\mathbb{R}^p \times \{-1, 1\}$ and the α_i are non
 245 negative real values obtained by solving a quadratic programming problem
 246 and γ is a parameter of the method. Then, if we use Kernel Ridge Regression
 247 on the training set $\{(\mathbf{Q}_{\lambda, \tau_d} \mathbf{X}_i^{\tau_d}, Y_i)\}_{i=1}^n$ (rather than the original training set
 248 $\{(\mathbf{X}_i^{\tau_d}, Y_i)\}_{i=1}^n$), it will work on the norm in L^2 of the derivatives of order
 249 m of the spline estimates of the X_i (up to the boundary conditions). More
 250 precisely, the regression function will have the following form:

$$\begin{aligned} \mathbf{x}^{\tau_d} &\mapsto \sum_{i=1}^n Y_i \alpha_i e^{-\gamma \|\mathbf{Q}_{\lambda, \tau_d} \mathbf{X}_i^{\tau_d} - \mathbf{Q}_{\lambda, \tau_d} \mathbf{x}^{\tau_d}\|_{\mathbb{R}^{|\tau_d|}}^2} \\ &\mapsto \sum_{i=1}^n Y_i \alpha_i e^{-\gamma \|D^m \widehat{X}_{i, \lambda, \tau_d} - D^m \widehat{x}_{\lambda, \tau_d}\|_{L^2}^2} \\ &\quad \times e^{-\gamma \sum_{j=1}^m (B^j \widehat{X}_{i, \lambda, \tau_d} - B^j \widehat{x}_{\lambda, \tau_d})^2} \end{aligned}$$

251 In other words, up to the boundary conditions, an estimation method based
 252 solely on inner products, or on norms derived from these inner products,

253 can be given modified inputs that will make it work on an estimation of the
 254 derivatives of the observed functions.

255 **Remark 1.** As shown in Corollary 1 in the previous section, building a
 256 model on \mathbf{X}^{τ_d} or on $\widehat{X}_{\lambda, \tau_d}$ leads to the same optimal predictive performances.
 257 In addition, it is obvious that given any one-to-one mapping f from $\mathbb{R}^{|\tau_d|}$ to
 258 itself, building a model on $f(\mathbf{X}^{\tau_d})$ gives also the same optimal performances
 259 than building a model on \mathbf{X}^{τ_d} . Then as $\mathbf{Q}_{\lambda, \tau_d}$ is invertible, the optimal
 260 predictive performances achievable with $\mathbf{Q}_{\lambda, \tau_d} \mathbf{X}^{\tau_d}$ are equal to the optimal
 261 performances achievable with \mathbf{X}^{τ_d} or with $\widehat{X}_{\lambda, \tau_d}$.

262 In practice however, the actual preprocessing of the data can have a strong
 263 influence on the obtained performances, as will be illustrated in Section 6.
 264 The goal of the theoretical analysis of the present section is to guarantee
 265 that no systematic loss can be observed as a consequence of the proposed
 266 functional preprocessing scheme.

267 4. Approximation results

268 The previous section showed that working on \mathbf{X}^{τ_d} , $\mathbf{Q}_{\lambda, \tau_d} \mathbf{X}^{\tau_d}$ or $\widehat{X}_{\lambda, \tau_d}$
 269 makes no difference in terms of optimal predictive performances. The present
 270 section addresses the effects of sampling: asymptotically, the optimal predic-
 271 tive performances obtained on $\widehat{X}_{\lambda, \tau_d}$ converge to the optimal performances
 272 achievable on the original and unobserved functional variable X .

273 4.1. Spline approximation

274 From the sampled random function $\mathbf{X}^{\tau_d} = (X(t_1), \dots, X(t_{|\tau_d|}))$, we can
 275 build an estimate, $\widehat{X}_{\lambda, \tau_d}$, of X . To ensure consistency, we must guarantee
 276 that $\widehat{X}_{\lambda, \tau_d}$ converges to X . In the case of a deterministic function x , this
 277 problem has been studied in numerous papers, such as Craven and Wahba
 278 (1978); Ragozin (1983); Cox (1984); Utreras (1988); Wahba (1990) (among
 279 others). Here we recall one of the results which is particularly well adapted
 280 to our context.

281 Obviously, the sampling grid must behave correctly, whereas the infor-
 282 mation contained in \mathbf{X}^{τ_d} will not be sufficient to recover X . We need also
 283 the regularization parameter λ to depend on τ_d . Following Ragozin (1983),
 284 a sampling grid τ_d is characterized by two quantities:

$$\begin{aligned} \overline{\Delta}_{\tau_d} &= \max\{t_1, t_2 - t_1, \dots, 1 - t_{|\tau_d|}\} \\ \underline{\Delta}_{\tau_d} &= \min_{1 \leq i < |\tau_d|} \{t_{i+1} - t_i\}. \end{aligned} \tag{8}$$

285 One way to control the distance between X and $\widehat{X}_{\lambda, \tau_d}$ is to bound the ratio
 286 $\overline{\Delta}_{\tau_d} / \underline{\Delta}_{\tau_d}$ so as to ensure quasi-uniformity of the sampling grid.

287 More precisely, we will use the following assumption:

288 **Assumption 2.** *There is R such that $\overline{\Delta}_{\tau_d} / \underline{\Delta}_{\tau_d} \leq R$ for all d .*

289 Then we have:

290 **Theorem 2** (Ragozin (1983)). *Under Assumptions (A1) and (A2), there are*
 291 *two constants $A_{R,m}$ and $B_{R,m}$ depending only on R and m , such that for any*
 292 *$x \in \mathcal{H}^m$ and any positive λ :*

$$\|\widehat{x}_{\lambda, \tau_d} - x\|_{L^2}^2 \leq \left(A_{R,m} \lambda + B_{R,m} \frac{1}{|\tau_d|^{2m}} \right) \|D^m x\|_{L^2}^2.$$

293 This result is a rephrasing of Corollary 4.16 from Ragozin (1983) which
 294 is itself a direct consequence of Theorem 4.10 from the same paper.

295 Convergence of $\widehat{x}_{\lambda, \tau_d}$ to x is then obtained by the following simple as-
 296 sumptions:

297 **Assumption 3.** *The series of sampling points τ_d and the series of regular-*
 298 *ization parameters, λ , depending on τ_d and denoted by $(\lambda_d)_{d \geq 1}$, are such that*
 299 *$\lim_{d \rightarrow +\infty} |\tau_d| = +\infty$ and $\lim_{d \rightarrow +\infty} \lambda_d = 0$.*

300 4.2. Conditional expectation approximation

301 The next step consists in relating the optimal predictive performances
 302 for the regression and the classification problem (X, Y) to the performances
 303 associated to $(\widehat{X}_{\lambda_d, \tau_d}, Y)$ when d goes to infinity, i.e., relating L^* to

304 1. *binary classification case:*

$$L_d^* = \inf_{\phi: \mathcal{H}^m \rightarrow \{-1, 1\}} \mathbb{P} \left(\phi(\widehat{X}_{\lambda_d, \tau_d}) \neq Y \right),$$

305 2. *regression case:*

$$L_d^* = \inf_{\phi: \mathcal{H}^m \rightarrow \mathbb{R}} \mathbb{E} \left([\phi(\widehat{X}_{\lambda_d, \tau_d}) - Y]^2 \right)$$

306 Two sets of assumptions will be investigated to provide the convergence
 307 of the Bayes risk L_d^* to L^* :

308 **Assumption 4. Either**

309 (A4a) $\mathbb{E}(\|D^m X\|_{L^2}^2)$ is finite and $Y \in \{-1, 1\}$,

310 **or**

311 (A4b) $\tau_d \subset \tau_{d+1}$ and $\mathbb{E}(Y^2)$ is finite.

312 The first assumption (A4a) requires an additional smoothing property for
313 the predictor functional variable X and is only valid for a binary classifica-
314 tion problem whereas the second assumption (A4b) requires an additional
315 property for the sampling point series: they have to be growing sets.

316 Theorem 2 then leads to the following corollary:

317 **Corollary 3.** *Under Assumptions (A1)-(A4), we have:*

$$\lim_{d \rightarrow +\infty} L_d^* = L^*.$$

318 5. General consistent functional classifiers and regression functions

319 5.1. Definition of classifiers and regression functions on derivatives

320 Let us now consider any consistent classification or regression scheme for
321 standard multivariate data based either on the inner product or on the Eu-
322 clidean distance between observations. Examples of such classifiers are Sup-
323 port Vector Machine Steinwart (2002), the kernel classification rule Devroye
324 and Krzyżak (1989) and k -nearest neighbors Devroye and Györfi (1985);
325 Zhao (1987) to name a few. In the same way, multilayer perceptrons Lu-
326 gosi and Zeger (1990), kernel estimates Devroye and Krzyżak (1989) and
327 k -nearest neighbors regression Devroye et al. (1994) are consistent regression
328 estimators. Additional examples of consistent estimators in classification and
329 regression can be found in Devroye et al. (1996); Györfi et al. (2002).

330 We denote $\psi_{\mathcal{D}}$ the estimator constructed by the chosen scheme using a
331 dataset $\mathcal{D} = \{(U_i, T_i)_{1 \leq i \leq n}\}$, where the $(U_i, T_i)_{1 \leq i \leq n}$ are n independent copies
332 of a pair of random variables (U, T) with values in $\mathbb{R}^p \times \{-1, 1\}$ (classification)
333 or $\mathbb{R}^p \times \mathbb{R}$ (regression).

334 The proposed functional scheme consists in choosing the estimator ϕ_{n, τ_d}
335 as $\psi_{\mathcal{E}_{n, \tau_d}}$ with the dataset \mathcal{E}_{n, τ_d} defined by:

$$\mathcal{E}_{n, \tau_d} = \{(\mathbf{Q}_{\lambda_d, \tau_d} \mathbf{X}_i^{\tau_d}, Y_i)_{1 \leq i \leq n}\}$$

336 As pointed out in Section 3.4, the linear transformation $\mathbf{Q}_{\lambda_d, \tau_d}$ is an approx-
 337 imate multivariate differentiation operator: up to the boundary conditions,
 338 an estimator based on $\mathbf{Q}_{\lambda_d, \tau_d} \mathbf{X}^{\tau_d}$ is working on the m -th derivative of $\widehat{X}_{\lambda_d, \tau_d}$.

339 In more algorithmic terms, the estimator is obtained as follows:

- 340 1. choose an appropriate value for λ_d
- 341 2. compute $\mathbf{M}_{\lambda_d, \tau_d}$ using Theorem 1 and Corollary 2;
- 342 3. compute the Cholesky decomposition of $\mathbf{M}_{\lambda_d, \tau_d}$ and the transpose of
 343 the Cholesky triangle, $\mathbf{Q}_{\lambda_d, \tau_d}$ (such that $\mathbf{Q}_{\lambda_d, \tau_d}^T \mathbf{Q}_{\lambda_d, \tau_d} = \mathbf{M}_{\lambda_d, \tau_d}$);
- 344 4. compute $\mathbf{Q}_{\lambda_d, \tau_d} \mathbf{X}_i^{\tau_d}$ to obtain the transformed dataset \mathcal{E}_{n, τ_d} ;
- 345 5. build a classifier/regression function $\psi_{\mathcal{E}_{n, \tau_d}}$ with a multivariate method
 346 in $\mathbb{R}^{|\tau_d|}$ applied to the dataset \mathcal{E}_{n, τ_d} ;
- 347 6. associate to a new sampled function $\mathbf{X}_{n+1}^{\tau_d}$ the prediction
 348 $\psi_{\mathcal{E}_{n, \tau_d}}(\mathbf{Q}_{\lambda, \tau_d} \mathbf{X}_{n+1}^{\tau_d})$.

349 Figure 5.1 illustrates the way the method performs: instead of relying
 350 on an approximation of the function and then on the derivation preprocess-
 351 ing of this estimates, it directly uses an equivalent metric by applying the
 352 $\mathbf{Q}_{\lambda_d, \tau_d}$ matrix to the sampled function. The consistency result proved in The-
 353 orem 3 shows that, combined with any consistent multidimensional learning
 354 algorithm, this method is (asymptotically) equivalent to using the original
 355 function drawn at the top left side of Figure 5.1.

356 On a practical point of view, Wahba (1990) demonstrates that cross val-
 357 idated estimates of λ achieve suitable convergence rates. Hence, steps 1 and
 358 2 can be computed simultaneously by minimizing the total cross validated
 359 error for all the observations, given by

$$\sum_{i=1}^n \frac{1}{|\tau_d|} \sum_{t \in \tau_d} \frac{(x_i(t) - \widehat{x}_{i, \lambda, \tau_d}(t))^2}{(1 - A_{tt}(\lambda))^2},$$

360 where A is a $|\tau_d| \times |\tau_d|$ matrix called the *influence matrix* (see Wahba (1990)),
 361 over a finite number of λ values.

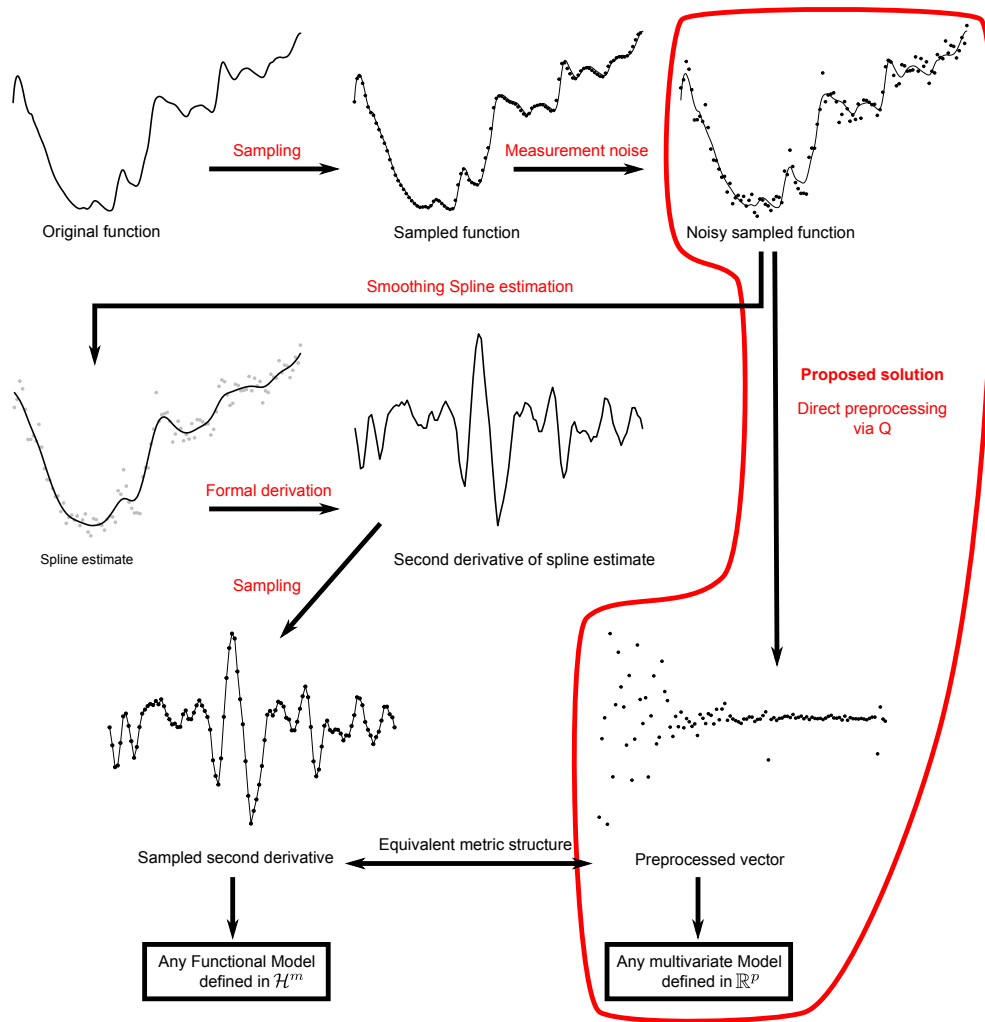


Figure 1: Method scheme and its equivalence to the usual approach for using derivatives in learning algorithms.

362 *5.2. Consistency result*

363 Corollary 1 and Corollary 3 guarantee that the estimator proposed in the
 364 previous section is consistent:

365 **Theorem 3.** *Under assumptions (A1)-(A4), the series of classi-*
 366 *fiers/regression functions $(\phi_{n,\tau_d})_{n,d}$ is consistent:*

$$\lim_{d \rightarrow +\infty} \lim_{n \rightarrow +\infty} \mathbb{E} (L\phi_{n,\tau_d}) = L^*$$

367 *5.3. Discussion*

368 While Theorem 3 is very general, it could be easily extended to cover
 369 special cases such as additional hypothesis needed by the estimation scheme
 370 or to provide data based parameter selections. We discuss briefly those issues
 371 in the present section.

372 It should first be noted that most estimation schemes, $\psi_{\mathcal{D}}$, depend on
 373 parameters that should fulfill some assumptions for the scheme to be con-
 374 sistent. For instance, in the Kernel Ridge Regression method in \mathbb{R}^p , with
 375 Gaussian kernel, $\psi_{\mathcal{D}}$ has the form given in Equation (7) where the (α_i) are
 376 the solutions of

$$\arg \min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \left(T_i - \sum_{j=1}^n T_j \alpha_j e^{-\gamma \|U_i - U_j\|_{\mathbb{R}^p}^2} \right)^2 +$$

$$\delta_n \sum_{i,j=1}^n T_i T_j \alpha_i \alpha_j e^{-\gamma \|U_i - U_j\|_{\mathbb{R}^p}^2}.$$

377 The method thus depends on the parameter of the Gaussian kernel, γ and
 378 of the regularization parameter δ_n . This method is known to be consistent if
 379 (see Theorem 9.1 of Steinwart and Christmann (2008)):

$$\delta_n \xrightarrow{n \rightarrow +\infty} 0 \quad \text{and} \quad n\delta_n^4 \xrightarrow{n \rightarrow +\infty} +\infty.$$

380 Additional conditions of this form can obviously be directly integrated in
 381 Theorem 3 to obtain consistency results specific to the corresponding algo-
 382 rithms.

383 Moreover, practitioners generally rely on data based selection of the pa-
 384 rameters of the estimation scheme $\psi_{\mathcal{D}}$ via a validation method: for instance,
 385 rather than setting δ_n to e.g., n^{-5} for n observations (a choice which is com-
 386 patible with theoretical constraints on δ_n), one chooses the value of δ_n that

387 optimizes an estimation of the performances of the regression function ob-
388 tained on an independent data set (or via a re-sampling approach).

389 In addition to the parameters of the estimation scheme, functional data
390 raise the question of the convenient order of the derivative, m , and of the
391 sampling grid optimality. In practical applications, the number of available
392 sampling points can be unnecessarily large (see Biau et al. (2005) for an ex-
393 ample with more than 8 000 sampling points). The preprocessing performed
394 by $\mathbf{Q}_{\lambda_d, \tau_d}$ do not change the dimensionality of the data which means that
395 overfitting can be observed in practice when the number of sampling points
396 is large compared to the number of functions. Moreover, processing very
397 high dimensional vectors is time consuming. It is there quite interesting in
398 practice to use a down-sampled version of the original grid.

399 To select the parameters of ψ_D , the order of the derivative and/or the
400 down-sampled grid, a validation strategy, based on splitting the dataset into
401 training and validation sets, could be used. A simple adaptation of the idea
402 of Berline et al. (2008); Biau et al. (2005); Laloë (2008); Rossi and Villa
403 (2006) shows that a penalized validation method can be used to choose any
404 combination of those parameters consistently. According to those papers,
405 the condition for the consistency of the validation strategy would simply
406 relate the shatter coefficients of the set of classifiers in \mathbb{R}^d to the penalization
407 parameter of the validation. Once again, this type of results is a rather direct
408 extension of Theorem 3.

409 6. Applications

410 In this section, we show that the proposed approach works as expected on
411 real world spectrometric examples: for some applications, the use of deriva-
412 tives leads to more accurate models than the direct processing of the spectra
413 (see e.g. Rossi et al. (2005); Rossi and Villa (2006) for other examples of such
414 a behavior based on ad hoc estimators of the spectra derivatives). It should
415 be noted that the purpose of this section is only to illustrate the behavior
416 of the proposed method on finite datasets. The theoretical results of the
417 present paper show that all consistent schemes have asymptotically identical
418 performances, and therefore that using derivatives is asymptotically useless.
419 On a finite dataset however, preprocessing can have strong influence on the
420 predictive performances, as will be illustrated in the present section. In ad-
421 dition, schemes that are not universally consistent, e.g., linear models, can
422 lead to excellent predictive performances on finite datasets; such models are

423 therefore included in the present section despite the fact the theory does not
424 apply to them.

425 6.1. Methodology

426 The methodology followed for the two illustrative datasets is roughly the
427 same:

- 428 1. the dataset is randomly split into a training set on which the model is
429 estimated and a test set on which performances are computed. The split
430 is repeated several times. The Tecator dataset (Section 6.2) is rather
431 small (240 spectra) and exhibits a rather large variability in predic-
432 tive performances between different random splits. We have therefore
433 used 250 random splits. For the Yellow-berry dataset (Section 6.3), we
434 used only 50 splits as the relative variability in performances is far less
435 important.
- 436 2. λ is chosen by a global leave-one-out strategy on the spectra contained
437 in training set (as suggested in Section 5.1). More precisely, a leave-one-
438 out estimate of the reconstruction error of the spline approximation of
439 each training spectrum is computed for a finite set of candidate values
440 for λ . Then a common λ is chosen by minimizing the average over
441 the training spectra of the leave-one-out reconstruction errors. This
442 choice is relevant as cross validation estimates of λ are known to have
443 favorable theoretical properties (see Craven and Wahba (1978); Utreras
444 (1981) among others).
- 445 3. for regression problems, a Kernel Ridge Regression (KRR) Saunders
446 et al. (1998); Shawe-Taylor and Cristianini (2004) is then performed to
447 estimate the regression function; this method is consistent when used
448 with a Gaussian kernel under additional conditions on the parameters
449 (see Theorem 9.1 of Steinwart and Christmann (2008)); as already ex-
450 plained, in the applications, Kernel Ridge Regression is performed both
451 with a Gaussian kernel and with a linear kernel (in that last case, the
452 model is essentially a ridge regression model). Parameters of the models
453 (a regularization parameter, δ_n , in all cases and a kernel parameter, γ
454 for Gaussian kernels) are chosen by a grid search that minimizes a vali-
455 dation based estimate of the performances of the model (on the training
456 set). A leave-one-out solution has been chosen: in Kernel Ridge Re-
457 gression, the leave-one-out estimate of the performances of the model is

458 obtained as a by-product of the estimation process, without additional
459 computation cost, see e.g. Cawley and Talbot (2004).
460 Additionally, for a sake of comparison with a more traditional approach
461 in FDA, Kernel Ridge Regression is compared with a nonparametric
462 kernel estimate for the Tecator dataset (Section 6.2.1). Nonparametric
463 kernel estimate is the first nonparametric approach introduced in Func-
464 tional Data Analysis Ferraty and Vieu (2006) and can thus be seen as
465 a basis for comparison in the context of regression with functional pre-
466 dictors. For this method, the same methodology as with Kernel Ridge
467 Regression was used: the parameter of the model (i.e., the bandwidth)
468 was selected on a grid search minimizing a cross-validation estimate of
469 the performances of the model. In this case, a 4-fold cross validation
470 estimate was used instead of a leave-one-out estimate to avoid a large
471 computational cost.

- 472 4. for the classification problem, a Support Vector Machine (SVM) is used
473 Shawe-Taylor and Cristianini (2004). As KRR, SVM are consistent
474 when used with a Gaussian kernel Steinwart (2002). We also use a
475 SVM with a linear kernel as this is quite adapted for classification in
476 high dimensional spaces associated to sampled function data. We also
477 use a K-nearest neighbor model (KNN) for reference. Parameters of the
478 models (a regularization parameter for both SVM, a kernel parameter,
479 γ for Gaussian kernels and number of neighbors K for KNN) are chosen
480 by a grid search that minimizes a validation based estimate of the
481 classification error: we use a 4-fold cross-validation to get this estimate.
- 482 5. We evaluate the models obtained for each random split on the test set.
483 We report the mean and the standard deviation of the performance
484 index (classification error and mean squared error, respectively) and
485 assess the significance of differences between the reported figures via
486 paired Student tests (with level 1%).
- 487 6. Finally, we compare models estimated on the raw spectra and on spec-
488 tra transformed via the $\mathbf{Q}_{\lambda_d, \tau_d}$ matrix for $m = 1$ (first derivative) and
489 $m = 2$ (second derivative). For both values of m , we used the most
490 classical boundary conditions ($x(0) = 0$ and $Dx(0) = 0$). Depending of
491 the problem, other boundary conditions could be investigated but this
492 is outside the scope of the present paper (see Besse and Ramsay (1986);
493 Heckman and Ramsay (2000) for discussion on this subject). For the

494 Tecator problem, we also compare these approaches with models es-
495 timated on first and second derivatives based on interpolating splines
496 (i.e. with $\lambda = 0$) and on first and second derivatives estimated by finite
497 differences.

498 Note that the kind of preprocessing used has almost no impact on
499 the computation time. In general, selecting the parameters of the
500 model with leave-one-out or cross-validation will use significantly more
501 computing power than constructing the splines and calculating their
502 derivatives. For instance, computing the optimal λ with the approach
503 described above takes less than 0.1 second for the Tecator dataset on a
504 standard PC using our R implementation which is negligible compared
505 to the several minutes used to select the optimal parameters of the
506 models used on the preprocessed data.

507 6.2. Tecator dataset

508 The first studied dataset is the standard Tecator dataset Thodberg (1996)
509 ¹. It consists in spectrometric data from the food industry. Each of the
510 240 observations is the near infrared absorbance spectrum of a meat sample
511 recorded on a Tecator Infratec Food and Feed Analyzer. Each spectrum is
512 sampled at 100 wavelengths uniformly spaced in the range 850–1050 nm.
513 The composition of each meat sample is determined by analytic chemistry
514 and percentages of moisture, fat and protein are associated this way to each
515 spectrum.

516 The Tecator dataset is a widely used benchmark in Functional Data Anal-
517 ysis, hence the motivation for its use for illustrative purposes. More precisely,
518 in Section 6.2.1, we address the original regression problem by predicting the
519 percentage of fat content from the spectra with various regression method
520 and various estimates of the derivative preprocessing: this analysis shows
521 that both the method and the use of derivative have a strong effect on the
522 performances whereas the way the derivatives are estimated has almost no
523 effect. Additionally, in Section 6.2.2, we apply a noise (with various vari-
524 ances) to the original spectra in order to study the influence of smoothing
525 in the case of noisy predictors: this section shows the relevance of the use of
526 a smoothing spline approach when the data are noisy. Finally, Section 6.2.3
527 deals with a classification problem derived from the original Tecator problem

¹Data are available on statlib at <http://lib.stat.cmu.edu/datasets/tecator>

528 (in the same way as what was done in Ferraty and Vieu (2003)): conclusions
529 of this section are similar to the ones of the regression study.

530 6.2.1. Fat content prediction

531 As explained above, we first address the regression problem that consists
532 in predicting the fat content of peaces of meat from the Tecator dataset. The
533 parameters of the model are optimized with a grid search using the leave-one-
534 out estimate of the predictive performances (both models use a regularization
535 parameter, with an additional width parameter in the Gaussian kernel case).
536 The original data set is split randomly into 160 spectra for learning and 80
537 spectra for testing. As shown in the result Table 1, the data exhibit a rather
538 large variability; we use therefore 250 random split to assess the differences
539 between the different approaches.

540 The performance indexes are the mean squared error (M.S.E.) and the
541 R^2 .² As a reference, the target variable (fat) has a variance equal to 14.36.
542 Results are summarized in Table 1.

543 The first conclusion is that the method itself has a strong effect on the
544 performances of the prediction: for this application, a linear method is not
545 appropriate (mean squared errors are much greater for linear methods than
546 for the kernel ridge regression used with a Gaussian kernel) and the non-
547 parametric kernel estimate gives worse performances than the kernel ridge
548 regression (indeed, they are about 10 times worse). Nevertheless, for non-
549 parametric approaches (Gaussian KKR and NKE), the use of derivatives
550 has also a strong impact on the performances: for kernel ridge regression,
551 e.g., preprocessing by estimating the first order derivative leads to a strong
552 decrease of the mean squared error.

553 Differences between the average MSEs are not always significant, but
554 we can nevertheless rank the methods in increasing order of modeling error
555 (using notations explained in Table 1) for Gaussian kernel ridge regression:

$$\text{FD1} \leq \text{IS1} \leq \text{S1} < \text{DF2} \leq \text{SS2} < \text{IS2} < \text{O}$$

556 where $<$ corresponds to a significant difference (for a paired Student test
557 with level 1%) and \leq to a non significant one. In this case, the data are very
558 smooth and thus the use of smoothing splines instead of a finite differences

² $R^2 = 1 - \frac{\text{M.S.E}}{\text{Var}(y)}$ where $\text{Var}(y)$ is the (empirical) variance of the target variable on the test set.

Method	Data	Average M.S.E. and SD	Average R^2
KRR Linear	O	8.69 (4.47)	95.7%
	S1	8.09 (3.85)	96.1%
	IS1	8.09 (3.85)	96.1%
	FD1	8.27 (4.17)	96.0%
	S2	9.64 (4.98)	95.3%
	IS2	9.87 (5.84)	95.2%
	FD2	8.45 (4.18)	95.9%
KRR Gaussian	O	5.02 (11.47)	97.6%
	S1	0.485 (0.385)	99.8%
	IS1	0.485 (0.385)	99.8%
	FD1	0.484 (0.387)	99.8%
	S2	0.584 (0.303)	99.7%
	IS2	0.586 (0.303)	99.7%
	FD2	0.569 (0.281)	99.7%
NKE	O	73.1 (16.5)	64.2%
	S1	4.59 (1.09)	97.7%
	IS1	4.59 (1.09)	97.7%
	FD1	4.59 (1.09)	97.7%
	S2	3.75 (1.22)	98.2%
	IS2	3.75 (1.22)	98.2%
	FD2	3.67 (1.18)	98.2%

Table 1: Summary of the performances of the chosen models on the test set (fat Tecator regression problem) when using either a kernel ridge regression (KRR) with linear kernel or with Gaussian kernel or when using a nonparametric kernel estimate (NKE) with various inputs: O (original data), S1 (smoothing splines with order 1 derivatives), IS1 (interpolating splines with order 1 derivatives), FD1 (order 1 derivatives estimated by finite differences) and S2, IS2 and FD2 (the same as previously with order 2 derivatives).

559 approximation does not have a significant impact on the predictions. How-
 560 ever, in this case, the roughest approach, consisting in the estimation of the
 561 derivatives by finite differences, gives the best performances.

562 6.2.2. Noisy spectra

563 This section studies the situation in which functional data observations
 564 are corrupted by noise. This is done by adding a noise to each spectrum of
 565 the Tecator dataset. More precisely, each spectrum has been corrupted by

$$X_i^b(t) = X_i(t) + \epsilon_{it} \quad (9)$$

566 where (ϵ_{it}) are i.i.d. Gaussian variables with standard deviation equal to
 567 either 0.01 (small noise) or to 0.2 (large noise). 10 observations of the data
 568 generated this way are given in Figure 2.

569 The same methodology as for the non noisy data has been applied to (X_i^b)
 570 to predict the fat content. The experiments have been restricted to the use of
 571 kernel ridge regression with a Gaussian kernel (according to the nonlinearity
 572 of the problem shown in the previous section). Results are summarized in
 573 Table 2 and Figure 3.

574 In addition, the results can be ranked this way:

Noise with sd equal to 0.01

$$S2 < S1 < IS1 \leq O < FD1 < IS2 \leq FD2$$

Noise with sd equal to 0.2

$$S1 < O < S2 < FD1 < IS1 < IS2 \leq FD2$$

575 where $<$ corresponds to a significant difference (for a paired Student test
 576 with level 1%).

577 The first conclusion of these experiments is that, even though the deriva-
 578 tives are the relevant predictors, their performances are strongly affected by
 579 the noise (compared to the ones of the original data: note that the average
 580 M.S.E. reported in Table 1 are more 10 times lower than the best ones from
 581 Table 2 and that, in the best cases, R^2 is slightly greater than 50% for the
 582 most noisy dataset). In particular, using interpolating splines or finite differ-
 583 ence derivatives leads to highly deteriorated performances. In this situation,
 584 the approach proposed in the paper is particularly useful and helps to keep

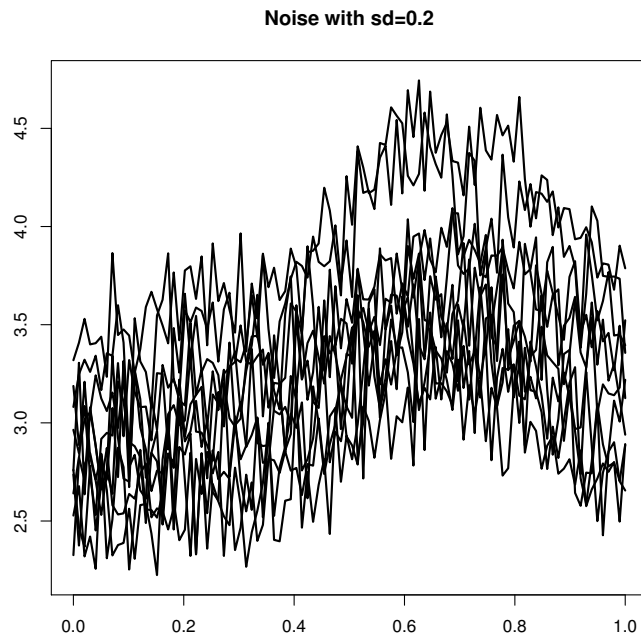
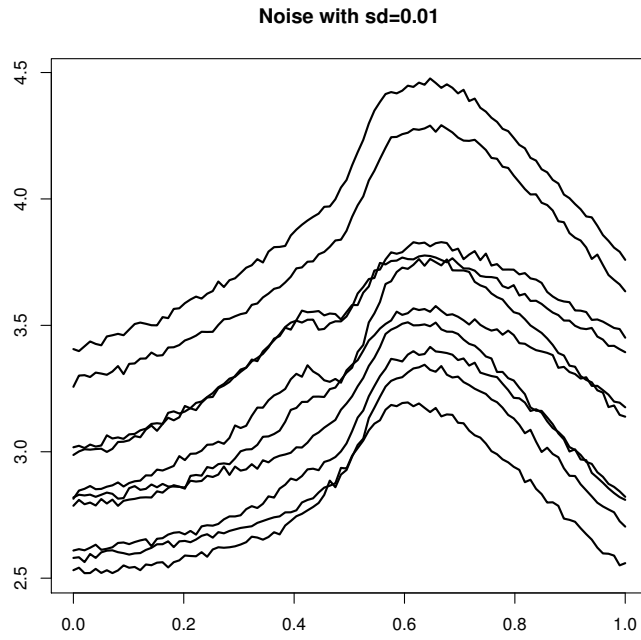


Figure 2: 10 observations of the noisy data generated from the Tecator spectra as in Equation 9

Noise	Data	Average M.S.E. and SD	Average R^2
sd = 0.01	O	13.3 (13.5)	93.5%
	S1	7.45 (1.5)	96.4%
	IS1	12.72 (2.2)	93.8%
	FD1	20.03 (2.8)	90.3%
	S2	6.83 (1.4)	96.7%
	IS2	31.23 (5.9)	84.9%
	FD2	31.10 (5.9)	84.9%
sd = 0.2	O	87.9 (13.9)	57.4%
	S1	85.0 (12.5)	58.8%
	IS1	210.1 (36.1)	-1.9%
	FD1	209.1 (33.0)	-1.4%
	S2	95.9 (12.8)	53.5%
	IS2	213.7 (33.1)	-3.6%
	FD2	235.1 (222.7)	-14.0%

Table 2: Summary of the performances of the chosen models on the test set (fat Tecator regression problem) with noisy spectra.

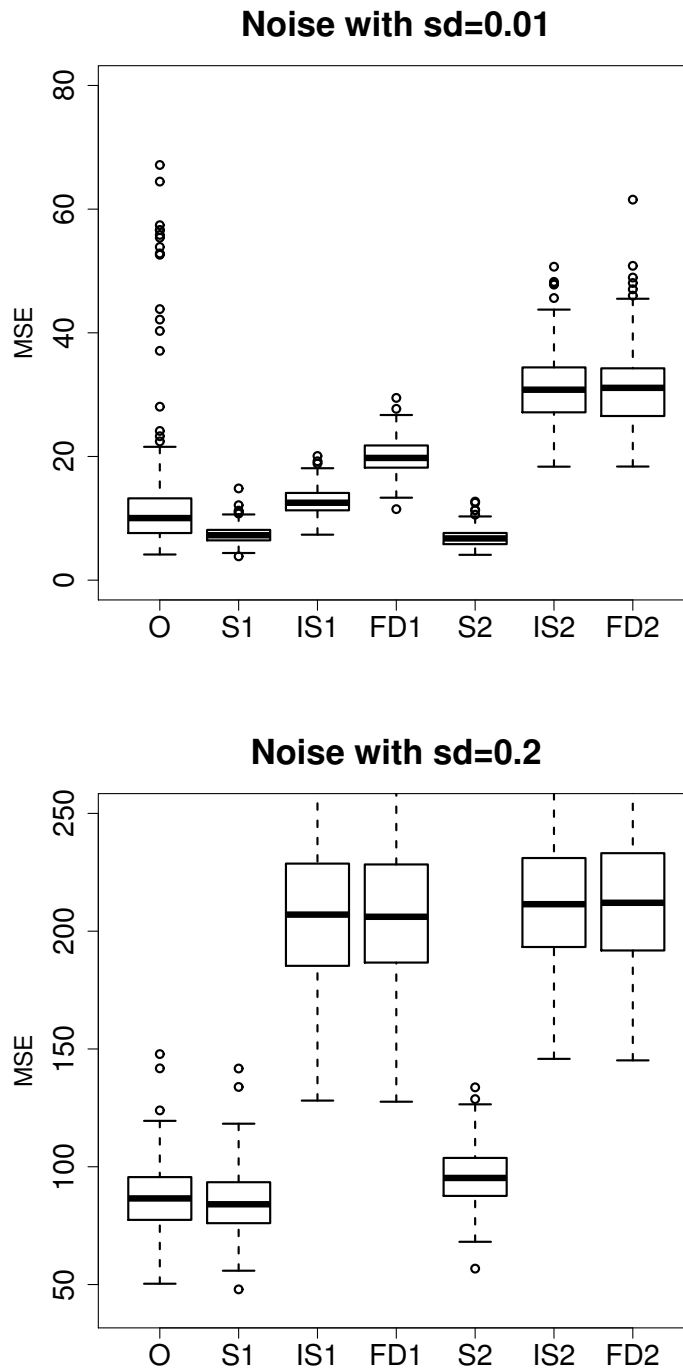


Figure 3: Mean squared errors boxplot for the noisy fat Tecator regression problem with Gaussian kernel (the worst test samples for S and FD have been removed for a sake of clarity)

585 better performances than with the original data. Indeed, the differences of
586 the smoothing splines approach with the original data is still significant (for
587 both derivatives in the “small noise” case and for the first order derivative
588 in the “high noise” case), even though, the most noisy the data are, the
589 most difficult it is to estimate the derivatives in an accurate way. That is,
590 except for smoothing spline derivatives, the estimation of the derivatives for
591 the most noisy dataset is so bad that it leads to negative R^2 when used in
592 the regression task.

593 *6.2.3. Fat content classification*

594 In this section, the fat content regression problem is transformed into a
595 classification problem. To avoid imbalance in class sizes, the median value
596 of the fat in the dataset is used as the splitting criterion: the first class
597 consists in 119 samples with strictly less than 13.5 % of fat, while the second
598 class contains the other 121 samples with a fat content equal or higher than
599 13.5 %.

600 As in previous sections, the analysis is conducted on 250 random splits of
601 the dataset into 160 learning spectra and 80 test spectra. We used stratified
602 sampling: the test set contains 40 examples from each class. The 4 fold
603 cross-validation used to select the parameters of the models on the learning
604 set is also stratified with roughly 20 examples of each class in each fold.

605 The performance index is the mis-classification rate (MCR) on the test
606 set, reported in percentage and averaged over the 250 random splits. Results
607 are summarized in Table 3. As in the previous sections, both the model
608 and the preprocessing have some influence on the results. In particular,
609 using derivatives always improves the classification accuracy while the actual
610 method used to compute those derivatives has no particular influence on the
611 results. Additionally, using interpolation splines leads, in this particular
612 problem, to results that are exactly identical to the ones obtained with the
613 smoothing splines: they are not reported in Table 3.

614 More precisely, for the three models (linear SVM, Gaussian SVM and
615 KNN), differences in mis-classification rates between the smoothing spline
616 preprocessing and the finite differences calculation is never significant, ac-
617 cording to a Student test with level 1 %. Additionally while the actual aver-
618 age mis-classification rates might seem quite different, the large variability of
619 the results (shown by the standard deviations) leads to significant differences
620 only for the most obvious cases. In particular, SVM models using derivatives
621 (of order one or two) are indistinguishable one from another using a Student

Method	Data	Average MCR	SD of MCR
Linear SVM	O	1.41	1.55
	S1	0.73	1.15
	FD1	0.74	1.15
	S2	0.94	1.27
	FD2	0.92	1.23
Gaussian SVM	O	3.39	2.57
	S1	0.97	1.41
	FD1	0.98	1.42
	S2	0.99	2.00
	FD2	0.97	1.27
KNN	O	22.0	5.02
	S1	6.67	2.55
	FD1	6.57	2.55
	S2	1.93	1.65
	FD2	1.93	1.63

Table 3: Summary of the performances of the chosen models on the test set (Tecator fat classification problem). See Table 1 for notations. MCR stands for mis-classification rate, SD for standard deviation.

622 test with level 1 %: all methods with less than 1 % of mean mis-classification
623 rate perform essentially identically. Other differences are significant: for in-
624 stance the linear SVM used on raw data performs significantly worse than
625 any SVM model used on derivatives.

626 It should be noted that the classification task studied in the present sec-
627 tion is obviously simpler than the regression task from which it is derived.
628 This explains the very good predictive performances obtained by simple mod-
629 els such as a linear SVM, especially with the proper preprocessing.

630 6.3. Yellow-berry dataset

631 The goal of the last experiment is to predict the presence of yellow-berry in
632 durum wheat (*Triticum durum*) kernels via a near infrared spectral analysis
633 (see Figure 4). Yellow-berry is a defect of the durum wheat seeds that reduces
634 the quality of the flour produced from affected wheat. The traditional way
635 to assess the occurrence of yellow-berry is by visual analysis of a sample of
636 the seed stock. In the current application, a quality measure related to the
637 occurrence of yellow-berry is predicted from the spectrum of the seed.

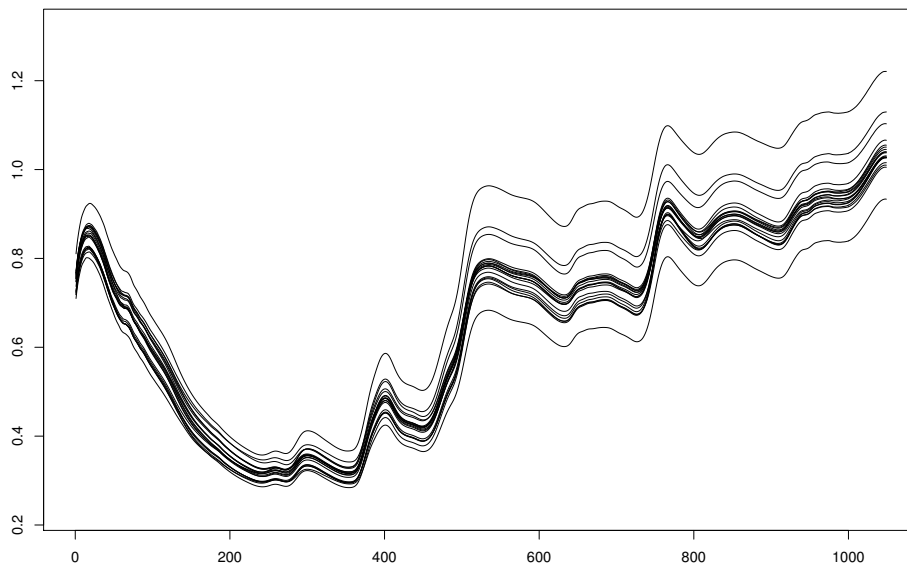


Figure 4: 20 observations of NIR spectra of durum wheat

638 The dataset consists in 953 spectra sampled at 1049 wavelengths uni-
639 formly spaced in the range 400–2498 nm. The dataset is split randomly into
640 600 learning spectra and 353 test spectra. Comparatively to the Tecator
641 dataset, the variability of the results is smaller in the present case. We used
642 therefore 50 random splits rather than 250 in the previous section.

643 The regression models were build via a Kernel Ridge Regression approach
644 using a linear kernel and a Gaussian kernel. In both cases, the regularization
645 parameter of the model is optimized by a leave-one-out approach. In addi-
646 tion, the width parameter of the Gaussian kernel is optimized via the same
647 procedure at the same time.

648 The performance index is the mean squared error (M.S.E.). As a refer-
649 ence, the target variable has a variance of 0.508. Results are summarized in
Table 4 and Figure 5.

Kernel and Data	Average M.S.E.	Standard deviation	Average R^2
Linear-O	0.122	$8.77 \cdot 10^{-3}$	76.1%
Linear-S1	0.138	$9.53 \cdot 10^{-3}$	73.0%
Linear-S2	0.122	$8.41 \cdot 10^{-3}$	76.1%
Gaussian-O	0.110	$20.2 \cdot 10^{-3}$	78.5%
Gaussian-S1	0.0978	$7.92 \cdot 10^{-3}$	80.9%
Gaussian-S2	0.0944	$8.35 \cdot 10^{-3}$	81.5%

Table 4: Summary of the performances of the chosen models on the test set (durum wheat regression problem)

650
651 As in the previous section, we can rank the methods in increasing order
652 of modelling error, we obtain the following result:

$$\text{G-S2} < \text{G-S1} < \text{G-O} < \text{L-O} \leq \text{L-S2} < \text{L-S1},$$

653 where G stands for Gaussian kernel and L for linear kernel (hence G-S2 stands
654 for kernel ridge regression with gaussian kernel and smoothing splines with
655 order 2 derivatives); $<$ corresponds to a significant difference (for a paired
656 Student test with level 1%) and \leq to a non significant one. For this appli-
657 cation, there is a significant gain in using a non linear model (the Gaussian
658 kernel). In addition, the use of derivatives leads to less contrasted perfor-
659 mances that the ones obtained in the previous section but it still improves
660 the quality of the non linear model in a significant way. In term of normal-
661 ized mean squared error (mean squared error divided by the variance of the

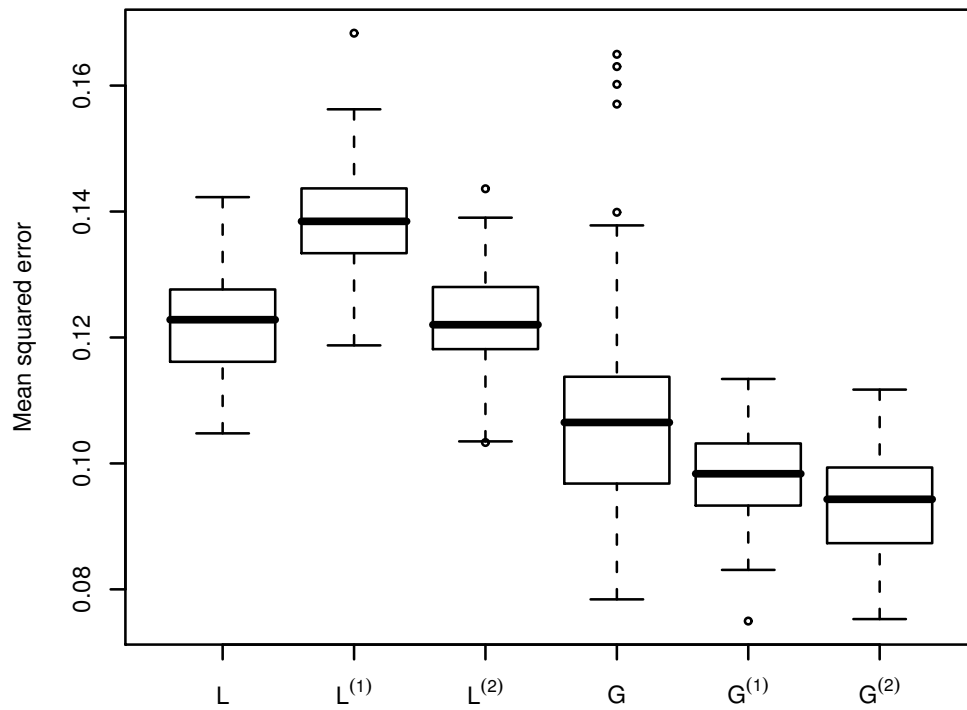


Figure 5: Mean squared error boxplots for the “durum wheat” regression problem (see Table 4 for the full names of the regression models)

662 target variable), using a non linear model with the second derivatives of the
663 spectra corresponds to an average gain of more than 5% (i.e., a reduction of
664 the normalised mean squared error from 24% for the standard linear model
665 to 18.6%).

666 7. Conclusion

667 In this paper we proposed a theoretical analysis of a common practice that
668 consists in using derivatives in classification or regression problems when the
669 predictors are curves. Our method relies on smoothing splines reconstruction
670 of the functions which are known only via a discrete deterministic sampling.
671 The method is proved to be consistent for very general classifiers or regres-
672 sion schemes: it reaches asymptotically the best risk that could have been
673 obtained by constructing a regression/classification model on the true ran-
674 dom functions.

675 We have validated the approach by combining it with nonparametric re-
676 gression and classification algorithms to study two real-world spectrometric
677 datasets. The results obtained in these applications confirm once again that
678 relying on derivatives can improve the quality of predictive models compared
679 to a direct use of the sampled functions. The way the derivatives are esti-
680 mated does not have a strong impact on the performances except when the
681 data are noisy. In this case, the use of smoothing splines is quite relevant.

682 In the future, several issues could be addressed. An important practical
683 problem is the choice of the best order of the derivative, m . We consider
684 that a model selection approach relying on a penalized error loss could be
685 used, as is done, in e.g., Rossi and Villa (2006), to select the dimension of
686 truncated basis representation for functional data. Note that in practice,
687 such parameter selection method could lead to select $m = 0$ and therefore to
688 automatically exclude derivative calculation when it is not needed. This will
689 extend the application range of the proposed model.

690 A second important point to study is the convergence rate for the method.
691 It would be very convenient for instance, to be able to relate the size of
692 the sampling grid to the number of functions. But, this latter issue would
693 require the use of additional assumptions on the smoothness of the regression
694 function whereas the result presented in this paper, even if more limited, only
695 needs mild conditions.

696 8. Acknowledgement

697 We thank Cécile Levasseur and Sylvain Coulomb (École d'Ingénieurs de
698 Purpan, EIP, Toulouse, France) for sharing the interesting problem presented
699 in Section 6.3.

700 We also thank Philippe Besse (Institut de Mathématiques de Toulouse,
701 Université de Toulouse, France) for helpfull discussions and suggestions.

702 Finally, we thank the anonymous reviewers for their valuable comments
703 and suggestions that helped to improve the quality of the paper.

704 References

705 Bahlmann, C., Burkhardt, H., 2004. The writer independent online hand-
706 writing recognition system *frog on hand* and cluster generative statistical
707 dynamic time warping. IEEE Transactions on Pattern Analysis and Ma-
708 chine Intelligence 26, 299–310.

709 Berlinet, A., Biau, G., Rouvière, L., 2008. Functional supervised classifica-
710 tion with wavelets. Annales de l'ISUP 52, 61–80.

711 Berlinet, A., Thomas-Agnan, C., 2004. Reproducing Kernel Hilbert Spaces
712 in Probability and Statistics. Kluwer Academic Publisher.

713 Besse, P., Ramsay, J., 1986. Principal component analysis of sampled curves.
714 Psychometrika 51, 285–311.

715 Biau, G., Bunea, F., Wegkamp, M., 2005. Functional classification in Hilbert
716 spaces. IEEE Transactions on Information Theory 51, 2163–2172.

717 Cardot, H., Ferraty, F., Sarda, P., 1999. Functional linear model. Statistics
718 and Probability Letters 45, 11–22.

719 Cawley, G., Talbot, N., 2004. Fast exact leave-one-out cross-validation of
720 sparse least-squares support vector machines. Neural Networks 17, 1467–
721 1475.

722 Cox, D., 1984. Multivariate smoothing splines functions. SIAM Journal on
723 Numerical Analysis 21, 789–813.

724 Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions.
725 Numerische Mathematik 31, 377–403.

- 726 Dauxois, J., Pousse, A., 1976. Les analyses factorielles en calcul des proba-
727 bilités et en statistique : essai d'étude synthétique. Thèse d'État. Univer-
728 sité Toulouse III.
- 729 Dejean, S., Martin, P., Baccini, A., Besse, P., 2007. Clustering time-series
730 gene expression data using smoothing spline derivatives. EURASIP Jour-
731 nal on Bioinformatics and Systems Biology 2007, Article ID70561.
- 732 Deville, J., 1974. Méthodes statistiques et numériques de l'analyse har-
733 monique. Annales de l'INSEE 15, 3–97.
- 734 Devroye, L., Györfi, L., 1985. Nonparametric Density Estimation: the L_1
735 view. John Wiley, New York.
- 736 Devroye, L., Györfi, L., Krzyżak, A., Lugosi, G., 1994. On the strong uni-
737 versal consistency of nearest neighbor regression function estimates. The
738 Annals of Statistics 22, 1371–1385.
- 739 Devroye, L., Györfi, L., Lugosi, G., 1996. A Probabilistic Theory for Pattern
740 Recognition. Springer-Verlag, New York.
- 741 Devroye, L., Krzyżak, A., 1989. An equivalence theorem for l_1 convergence
742 of the kernel regression estimate. Journal of Statistical Planning and In-
743 ference 23, 71–82.
- 744 Faragó, T., Györfi, L., 1975. On the continuity of the error distortion func-
745 tion for multiple-hypothesis decisions. IEEE Transactions on Information
746 Theory 21, 458–460.
- 747 Ferraty, F., Vieu, P., 2003. Curves discrimination: a non parametric ap-
748 proach. Computational and Statistical Data Analysis 44, 161–173.
- 749 Ferraty, F., Vieu, P., 2006. NonParametric Functional Data Analysis.
750 Springer.
- 751 Ferré, L., Yao, A., 2003. Functional sliced inverse regression analysis. Statis-
752 tics 37, 475–488.
- 753 Györfi, L., Kohler, M., Krzyżak, A., Walk, H., 2002. A Distribution-Free
754 Theory of Nonparametric Regression. Springer, New York.

- 755 Heckman, N., Ramsay, J., 2000. Penalized regression with model-based
756 penalties. *The Canadian Journal of Statistics* 28, 241–258.
- 757 James, G., 2002. Generalized linear models with functional predictor vari-
758 ables. *Journal of the Royal Statistical Society Series B* 64, 411–432.
- 759 James, G., Hastie, T., 2001. Functional linear discriminant analysis for ir-
760 regularly sampled curves. *Journal of the Royal Statistical Society, Series*
761 *B* 63, 533–550.
- 762 James, G., Hastie, T., Sugar, C., 2000. Principal component models for
763 sparse functional data. *Biometrika* 87, 587–602.
- 764 James, G., Silverman, B., 2005. Functional adaptive model estimation. *Jour-
765 nal of the American Statistical Association* 100, 565–576.
- 766 Kallenberg, O., 1997. *Foundations of Modern Probability. Probability and
767 its Applications*, Springer.
- 768 Kimeldorf, G., Wahba, G., 1971. Some results on Tchebycheffian spline
769 functions. *Journal of Mathematical Analysis and Applications* 33, 82–95.
- 770 Laloë, T., 2008. A k-nearest neighbor approach for functional regression.
771 *Statistics and Probability Letters* 78, 1189–1193.
- 772 Lugosi, G., Zeger, K., 1990. Nonparametric estimation via empirical risk
773 minimization. *IEEE Transaction on Information Theory* 41, 677–687.
- 774 Mas, A., Pumo, B., 2009. Functional linear regression with derivatives. *Jour-
775 nal of Nonparametric Statistics* 21, 19–40. Submitted: under revision.
776 Available at <http://www.math.univ-montp2.fr/~mas/FLRD.pdf>.
- 777 Pollard, D., 2002. *A User’s Guide to Measure Theoretic Probability*. Cam-
778 bridge University Press, Cambridge.
- 779 Ragozin, D., 1983. Error bounds for derivative estimation based on spline
780 smoothing of exact or noisy data. *Journal of Approximation Theory* 37,
781 335–355.
- 782 Ramsay, J., Dalzell, C., 1991. Some tools for functional data analysis (with
783 discussion). *Journal of the Royal Statistical Society. Series B. Statistical
784 Methodology* 53, 539–572.

- 785 Ramsay, J., Silverman, B., 1997. *Functional Data Analysis*. Springer Verlag,
786 New York.
- 787 Ramsay, J., Silverman, B., 2002. *Applied Functional Data Analysis*. Springer
788 Verlag.
- 789 Rossi, F., Conan-Guez, B., 2005. Functional multi-layer perceptron: a non-
790 linear tool for functional data analysis. *Neural Networks* 18, 45–60.
- 791 Rossi, F., Conan-Guez, B., 2006. Theoretical properties of projection based
792 multilayer perceptrons with functional inputs. *Neural Processing Letters*
793 23, 55–70.
- 794 Rossi, F., Delannay, N., Conan-Guez, B., Verleysen, M., 2005. Representa-
795 tion of functional data in neural networks. *Neurocomputing* 64, 183–210.
- 796 Rossi, F., Villa, N., 2006. Support vector machine for functional data classi-
797 fication. *Neurocomputing* 69, 730–742.
- 798 Saunders, G., Gammerman, A., Vovk, V., 1998. Ridge regression learning
799 algorithm in dual variables, in: *Proceedings of the Fifteenth International*
800 *Conference on Machine Learning (ICML'98)*, Madison, Wisconsin, USA.
801 pp. 515–521.
- 802 Shawe-Taylor, J., Cristianini, N., 2004. *Kernel methods for pattern analysis*.
803 Cambridge University Press, Cambridge, UK.
- 804 Steinwart, I., 2002. Support vector machines are universally consistent. *Jour-
805 nal of Complexity* 18, 768–791.
- 806 Steinwart, I., Christmann, A., 2008. *Support Vector Machines*. Information
807 Science and Statistics, Springer.
- 808 Thodberg, H., 1996. A review of bayesian neural network with an application
809 to near infrared spectroscopy. *IEEE Transaction on Neural Networks* 7,
810 56–72.
- 811 Utreras, F., 1981. Optimal smoothing of noisy data using spline functions.
812 *SIAM Journal on Scientific Computing* 2, 153–163.
- 813 Utreras, F., 1988. Boundary effects on convergence rates for Tikhonov regu-
814 larization. *Journal of Approximation Theory* 54, 235–249.

- 815 Wahba, G., 1990. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
816
- 817 Williams, B., Toussaint, M., Storkey, A., 2006. Extracting motion primitives from natural handwriting data, in: In Proceedings of the International
818 Conference on Artificial Neural Networks (ICANN).
819
- 820 Zhao, L., 1987. Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *Journal of Multivariate Analysis* 21,
821 168–178.
822

823 9. Proofs

824 9.1. Theorem 1

825 In the original theorem (Lemma 3.1) in Kimeldorf and Wahba (1971),
826 one has to verify that $(k_0(t_l, \cdot))_l$ spans \mathcal{H}_0^m and that $(k_1(t_l, \cdot))_l$ are linearly
827 independent. These are consequences of Assumption (A1).

828 First, $k_0(s, t) = \sum_{i,j=0}^{m-1} \tilde{B}_{ij}^{(-1)} s^i t^j$ where $\tilde{B} = (\tilde{B}_{i,j}^{(-1)})_{i,j}$ is the inverse of $(\sum_{l=1}^m B^l s^i B^l t^j)_{i,j}$ (see Heckman and Ramsay (2000)). Then
829 $(k_0(t_1, s), \dots, k_0(t_{|\tau_d|}, s)) = (1, s, \dots, s^{m-1}) \tilde{B} [V_{m-1}(t_1, \dots, t_{|\tau_d|})]^T$ where
830 $V_{m-1}(t_1, \dots, t_{|\tau_d|})$ is the Vandermonde matrix with $m - 1$ columns and $|\tau_d|$
831 rows associated to values $t_1, \dots, t_{|\tau_d|}$. If the $(t_l)_l$ are distinct, this matrix is
832 of full rank.
833

834 Moreover the reproducing property shows that $\sum_{l=1}^{|\tau_d|} a_l k_1(t_l, \cdot) \equiv 0$ implies
835 $\sum_{l=1}^{|\tau_d|} a_l f(t_l) \equiv 0$ for all $f \in \mathcal{H}_1^m$. Hence, $\mathcal{H}_1^m = \text{Ker}(B^T, \sum_{l=1}^{|\tau_d|} a_l \zeta_l)^T$
836 where ζ_l denotes the linear form $h \in \mathcal{H}^m \rightarrow h(t_l)$. As the co-dimension of
837 \mathcal{H}_1^m is $\dim \mathcal{H}_0^m = m$ and as, by Assumption (A1), B is linearly independent
838 of $\sum_{l=1}^{|\tau_d|} a_l \zeta_l$, we thus have $\sum_{l=1}^{|\tau_d|} a_l \zeta_l \equiv 0$ (or $\text{codim Ker}(B^T, \sum_{l=1}^{|\tau_d|} a_l \zeta_l)^T =$
839 $\dim \text{Im}(B^T, \sum_{l=1}^{|\tau_d|} a_l \zeta_l)$ would be $m+1$). Thus, we obtain that $\sum_{l=1}^{|\tau_d|} a_l f(t_l) \equiv$
840 0 for all f in \mathcal{H}^m and, as (t_l) are distinct, that $a_l = 0$ for all l , leading to the
841 independence conclusion for the $(k_1(t_l, \cdot))_l$.

842 Finally, we prove that $\mathcal{S}_{\lambda, \tau_d}$ is of full rank. Indeed, if $\mathcal{S}_{\lambda, \tau_d} \mathbf{x}^{\tau_d} = 0$,
843 $\omega^T M_0 \mathbf{x}^{\tau_d} = 0$ and $\eta^T M_1 \mathbf{x}^{\tau_d} = 0$. As $(\omega_k)_k$ is a basis of \mathcal{H}_0^m , $\omega^T M_0 \mathbf{x}^{\tau_d} = 0$
844 implies $M_0 \mathbf{x}^{\tau_d} = 0$ and therefore $M_1 = (K_1 + \lambda I_d)^{-1}$. As shown above,
845 the $(k_1(t_l, \cdot))_l$ are linearly independent and therefore $\eta M_1 \mathbf{x}^{\tau_d} = 0$ implies
846 $M_1 \mathbf{x}^{\tau_d} = 0$, which in turns leads to $\mathbf{x}^{\tau_d} = 0$ via the simplified formula for M_1 .

847 *9.2. Corollary 1*

848 We give only the proof for the classification case, the regression case is
849 identical.

850 According to Theorem 1, there is a full rank linear mapping from $\mathbb{R}^{|\tau_d|}$
851 to \mathcal{H}^m , $\mathcal{S}_{\lambda, \tau_d}$, such that for any function $x \in \mathcal{H}^m$, $\widehat{x}_{\lambda, \tau_d} = \mathcal{S}_{\lambda, \tau_d} \mathbf{x}^{\tau_d}$. Let
852 us denote $\mathcal{I}_{\lambda, \tau_d}$ the image of $\mathbb{R}^{|\tau_d|}$ by $\mathcal{S}_{\lambda, \tau_d}$, $\mathbf{P}_{\lambda, \tau_d}$ the orthogonal projection
853 from \mathcal{H}^m to $\mathcal{I}_{\lambda, \tau_d}$ and $\mathcal{S}_{\lambda, \tau_d}^{-1}$ the inverse of $\mathcal{S}_{\lambda, \tau_d}$ on $\mathcal{I}_{\lambda, \tau_d}$. Obviously, we have
854 $\mathcal{S}_{\lambda, \tau_d}^{-1} \circ \mathbf{P}_{\lambda, \tau_d}(\widehat{x}_{\lambda, \tau_d}) = \mathbf{x}^{\tau_d}$.

855 Let ψ be a measurable function from $\mathbb{R}^{|\tau_d|}$ to $\{-1, 1\}$. Then ζ_ψ de-
856 fined on \mathcal{H}^m by $\zeta_\psi(u) = \psi(\mathcal{S}_{\lambda, \tau_d}^{-1} \circ \mathbf{P}_{\lambda, \tau_d}(u))$ is a measurable function from
857 \mathcal{H}^m to $\{-1, 1\}$ (because $\mathcal{S}_{\lambda, \tau_d}^{-1}$ and $\mathbf{P}_{\lambda, \tau_d}$ are both continuous). Then for
858 any measurable ψ , $\inf_{\phi: \mathcal{H}^m \rightarrow \{-1, 1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) \leq \mathbb{P}(\zeta_\psi(\widehat{X}_{\lambda, \tau_d}) \neq Y) =$
859 $\mathbb{P}(\psi(\mathbf{X}^{\tau_d}) \neq Y)$, and therefore

$$\inf_{\phi: \mathcal{H}^m \rightarrow \{-1, 1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y) \leq \inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1, 1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y). \quad (10)$$

860 Conversely, let ψ be a measurable function from \mathcal{H}^m to $\{-1, 1\}$. Then ζ_ψ de-
861 fined on $\mathbb{R}^{|\tau_d|}$ by $\zeta_\psi(\mathbf{u}) = \psi(\mathcal{S}_{\lambda, \tau_d}(\mathbf{u}))$, is measurable. Then for any measurable
862 ψ , $\inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1, 1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y) \leq \mathbb{P}(\zeta_\psi(\mathbf{X}^{\tau_d}) \neq Y) = \mathbb{P}(\psi(\widehat{X}_{\lambda, \tau_d}) \neq Y)$,
863 and therefore

$$\inf_{\phi: \mathbb{R}^{|\tau_d|} \rightarrow \{-1, 1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y) \leq \inf_{\phi: \mathcal{H}^m \rightarrow \{-1, 1\}} \mathbb{P}(\phi(\widehat{X}_{\lambda, \tau_d}) \neq Y). \quad (11)$$

864 The combination of equations (10) and (11) gives equality (4).

865 *9.3. Corollary 3*

866 **1. Suppose assumption (A4a) is fulfilled**

867 The proof is based on Theorem 1 in Faragó and Györfi (1975). This
868 theorem relates the Bayes risk of a classification problem based on
869 (X, Y) with the Bayes risk of the problem $(T_d(X), Y)$ where (T_d) is a
870 series of transformations on X .

871 More formally, for a pair of random variables (X, Y) , where X takes
872 values in \mathcal{X} , an arbitrary metric space, and Y in $\{-1, 1\}$, let us

873 denote for any series of functions T_d from \mathcal{X} to itself, $L^*(T_d) =$
874 $\inf_{\phi: \mathcal{X} \rightarrow \{-1,1\}} \mathbb{P}(\phi(T_d(X)) \neq Y)$. Theorem 1 from Faragó and Györfi
875 (1975) states that $\mathbb{E}(\delta(T_d(X), X)) \xrightarrow{d \rightarrow +\infty} 0$ implies $L^*(T_d) \xrightarrow{d \rightarrow +\infty} L^*$,
876 where δ denotes the metric on \mathcal{X} .

877 This can be applied to $\mathcal{X} = (\mathcal{H}^m, \langle \cdot, \cdot \rangle_{L^2})$ with $T_d(X) =$
878 $\widehat{X}_{\lambda_d, \tau_d} = S_{\lambda_d, \tau_d} \mathbf{X}^{\tau_d}$: under Assumptions (A1) and (A2), Theo-
879 rem 2 gives: $\|T_d(X) - X\|_{L^2}^2 \leq \left(A_{R,m} \lambda_d + B_{R,m} \frac{1}{|\tau_d|^{2m}} \right) \|D^m X\|_{L^2}^2$.
880 Taking the expectation of both sides gives $\mathbb{E}(\|T_d(X) - X\|_{L^2}^2) \leq$
881 $\left(A_{R,m} \lambda_d + B_{R,m} \frac{1}{|\tau_d|^{2m}} \right) \mathbb{E}(\|D^m X\|_{L^2}^2)$, using the fact that the constants
882 are independent of the function under analysis. Then under Assump-
883 tions (A4a) and (A3), $\mathbb{E}(\|T_d(X) - X\|_{L^2}^2) \xrightarrow{d \rightarrow +\infty} 0$. According to
884 Faragó and Györfi (1975), this implies $\lim_{d \rightarrow \infty} L_d^* = L^*$.

885 2. Suppose assumption (A4b) is fulfilled

886 The conclusion will follow both for classification case and for regression
887 case. The proof follows the general ideas of Biau et al. (2005); Rossi
888 and Conan-Guez (2006); Rossi and Villa (2006); Laloë (2008). Under
889 assumption (A1), by Theorem 1 and with an argument similar to those
890 developed in the proof of Corollary 1, $\sigma(\widehat{X}_{\lambda_d, \tau_d}) = \sigma(\{X(t)\}_{t \in \tau_d})$. From
891 assumption (A4b), $\sigma(\{X(t)\}_{t \in \tau_d})$ is clearly a filtration. Moreover, as
892 $\mathbb{E}(Y)$ and thus $\mathbb{E}(Y^2)$ are finite, $\mathbb{E}(Y|\widehat{X}_{\lambda_d, \tau_d})$ is a uniformly bounded
893 martingale for this filtration (see Lemma 35 of Pollard (2002)). This
894 martingale converges in L^1 -norm to $\mathbb{E}(Y|\sigma(\cup_d \sigma(\widehat{X}_{\lambda_d, \tau_d})))$; we have

- 895 • $\sigma(\cup_d \sigma(\widehat{X}_{\lambda_d, \tau_d})) \subset \sigma(X)$ as $\widehat{X}_{\lambda_d, \tau_d}$ is a function of X (via Theo-
896 rem 1);
- 897 • by Theorem 2, $\widehat{X}_{\lambda_d, \tau_d} \xrightarrow{d \rightarrow +\infty, \text{ surely}} X$ in L^2 which proves that X
898 is $\sigma(\cup_d \sigma(\widehat{X}_{\lambda_d, \tau_d}))$ -measurable.

899 Finally, $\mathbb{E}(Y|\sigma(\cup_d \sigma(\widehat{X}_{\lambda_d, \tau_d}))) = \mathbb{E}(Y|X)$ and
900 $\mathbb{E}(Y|\widehat{X}_{\lambda_d, \tau_d}) \xrightarrow{d \rightarrow +\infty, L^1} \mathbb{E}(Y|X)$.

901 The conclusion follows from the fact that:

902 (a) *binary classification case:* the bound $L_d^* - L^* \leq$
 903 $2\mathbb{E}\left(\left|\mathbb{E}\left(Y|\widehat{X}_{\lambda_d,\tau_d}\right) - \mathbb{E}(Y|X)\right|\right)$ (see Theorem 2.2 of Devroye
 904 et al. (1996)) concludes the proof;

905 (b) *regression case:* as $\mathbb{E}(Y^2)$ is finite, $\mathbb{E}\left(\mathbb{E}\left(Y|\widehat{X}_{\lambda_d,\tau_d}\right)^2\right)$ is also fi-
 906 nite and the convergence also happens for the quadratic norm (see
 907 Corollary 6.22 in Kallenberg (1997)), i.e.,

$$\lim_{d \rightarrow +\infty} \mathbb{E}\left(\left(\mathbb{E}(Y|X) - \mathbb{E}\left(Y|\widehat{X}_{\lambda_d,\tau_d}\right)\right)^2\right) = 0$$

908 Hence, as $L_d^* - L^* = \mathbb{E}\left(\left(\mathbb{E}(Y|X) - \mathbb{E}\left(Y|\widehat{X}_{\lambda_d,\tau_d}\right)\right)^2\right)$, the con-
 909 clusion follows.

910 9.4. Theorem 3

911 We have

$$L(\phi_{n,d}) - L^* = L\phi_{n,\tau_d} - L_d^* + L_d^* - L^*. \quad (12)$$

912 Let ϵ be a positive real. By Corollary 3, it exists $d_0 \in \mathbb{N}^*$ such that, for all
 913 $d \geq d_0$,

$$L_d^* - L^* \leq \epsilon. \quad (13)$$

914 Moreover, as shown in Corollary 1 and as $\mathbf{Q}_{\lambda_d,\tau_d}$ is invertible, we have
 915 in the binary classification case: $L_d^* = \inf_{\phi:\mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{X}^{\tau_d}) \neq Y) =$
 916 $\inf_{\phi:\mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{Q}_{\lambda_d,\tau_d}\mathbf{X}^{\tau_d}) \neq Y)$, and in the regression case: $L_d^* =$
 917 $\inf_{\phi:\mathbb{R}^{|\tau_d|} \rightarrow \mathbb{R}} \mathbb{E}\left([\phi(\mathbf{X}^{\tau_d}) - Y]^2\right) = \inf_{\phi:\mathbb{R}^{|\tau_d|} \rightarrow \mathbb{R}} \mathbb{E}\left([\phi(\mathbf{Q}_{\lambda_d,\tau_d}\mathbf{X}^{\tau_d}) - Y]^2\right)$. By hy-
 918 pothesis, for any fixed d , ϕ_{n,τ_d} is consistent, that is

$$\lim_{n \rightarrow +\infty} \mathbb{E}(L(\phi_{n,\tau_d})) = \inf_{\phi:\mathbb{R}^{|\tau_d|} \rightarrow \{-1,1\}} \mathbb{P}(\phi(\mathbf{Q}_{\lambda_d,\tau_d}\mathbf{X}^{\tau_d}) \neq Y),$$

919 in the classification case and

$$\lim_{n \rightarrow +\infty} \mathbb{E}(L(\phi_{n,\tau_d})) = \inf_{\phi:\mathbb{R}^{|\tau_d|} \rightarrow \mathbb{R}} \mathbb{E}\left([\phi(\mathbf{Q}_{\lambda_d,\tau_d}\mathbf{X}^{\tau_d}) - Y]^2\right),$$

920 in the regression case, and therefore for any fixed d_0 ,
 921 $\lim_{n \rightarrow +\infty} \mathbb{E}\left(L(\phi_{n,\tau_{d_0}})\right) = L_{d_0}^*$. Combined with equations (12) and
 922 (13), this concludes the proof.