Consistency of Derivative Based Functional Classifiers on Sampled Data

Fabrice Rossi[1] and Nathalie Villa[2]

1– Projet AxIS, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay Cedex – France

2– Institut de Mathématiques de Toulouse, Université Toulouse III
118 route de Narbonne, 31062 Toulouse Cedex 9 – France

Abstract. In some applications, especially spectrometric ones, curve classifiers achieve better performances if they work on the m-order derivatives of their inputs. This paper proposes a smoothing spline based approach that give a strong theoretical background to this common practice.

1 Introduction

Spectrometric data are one particular case of functional data [11]: while each spectrum could be considered as a high dimensional vector, it has been shown that in some situations, a more appropriate model is to view it as a sampled function [1, 9]. One of the first benefits of this functional approach was to solve the difficulties faced by linear methods when confronted to the highly correlated variables resulting from a fine grid sampling of smooth functions (see [11] for a comprehensive presentation of linear methods for functional data).

On a theoretical point of view, Functional Data Analysis (FDA) departs from standard multivariate data analysis in one crucial point: the *real* data are not the finite dimensional vector representations but the underlying functions. For instance, in the particular case of spectral data, FDA focuses on the underlying spectra, not on their sampling via the spectrometer. As a consequence, standard multivariate consistency theorems cannot be directly applied to FDA. While the functional approach works quite well in practice (see, e.g., [7, 13]), empirical evidences of success are not sufficient to guarantee its soundness.

Several recent works tackle this problem, mainly by proving that some algorithms are consistent in the sense that they asymptotically estimate $\mathbb{E}(Y|X)$ (where X and Y are respectively a functional valued and a real valued random variable) from a learning set, or that they reach the Bayes risk in the limit. A comprehensive presentation of the capabilities of non parametric kernel based regression estimators for functional predictors can be found in [8]. In the case of classification, recent results include [4] that proposes a universal consistency result for k-nearest neighbors classifier and [14] which extends those results to Support Vector Machine (SVM). However, none of those consistency results covers some important aspects of the FDA methodology, namely the use of functional pre-processing. In spectrometric applications that deal with smooth spectra (e.g., near infrared spectroscopy), it is quite common, for instance, to replace functions by their first or second derivatives, as this allows to focus on curvature rather than on the actual values taken by the functions. This improves

prediction performances in certain situations (see e.g. [7, 13, 14]). Another limitation is that most of the theoretical results assume perfect knowledge of the observed curves and do not take into account sampling (see [12, 2] for examples of results on sampled functions).

This paper addresses both limitations via smooth spline representations of the sampled functions. The remainder of this paper is organized as follows: Section 2 introduces the general setting and outlines the proposed solution. Section 3 gives details about smoothing splines while Section 4 gives the main consistency result.

2 Setup and notation

We consider a binary classification problem[1] given by the pair of random variables (X, Y), where X takes values in a functional space $\mathcal{X}$, and Y in $\{-1, 1\}$. A learning set $S_{n,d}$ is constructed via n i.i.d. copies of (X, Y), the $\{(X_i, Y_i)\}_{i=1}^n$ and via a non-random sampling grid[2] of finite length $|\tau_d|$, $\tau_d = (t_l)_{l=1}^{|\tau_d|}$. It consists in the n pairs $(X_i^{\tau_d}, Y_i)$, where $X_i^{\tau_d} = \left(X_i(t_1), \ldots, X_i\left(t_{|\tau_d|}\right)\right)^T \in \mathbb{R}^{|\tau_d|}$. From S_{n,τ_d}, one builds a classifier ϕ_{n,τ_d} whose mis-classification probability is given by

$$L(\phi_{n,\tau_d}) = P\left(\phi_{n,\tau_d}\left(X^{\tau_d}\right) \neq Y | S_{n,\tau_d}\right).$$

The classifier is **consistent** if $L(\phi_{n,\tau_d})$ asymptotically reaches the Bayes risk for the *original functional problem*

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

Obviously, one cannot hope to find a consistent classifier without some regularity assumptions on the functions. Indeed if X takes arbitrary values in $L^2([0,1])$ then the actual value of $X(t_l)$ is not precisely defined. To solve this problem, [12] works on continuous functions while [2] uses more technical but related regularity assumptions.

The present paper targets situations in which the derivatives of the functions convey more information than the functions themselves. It is therefore natural to assume that the functional space $\mathcal{X}$ contains only differentiable functions. More precisely, for $m > \frac{3}{2}$, $\mathcal{X}$ is the Sobolev space $\mathcal{H}^m$ of functions from $L^2([0,1])$ for which $D^j h$ exists in the weak sense for all $j \leq m$ and such that $D^m h \in L^2([0,1])$, where $D^j h$ is the j-order derivative of h (also denoted $h^{(j)}$). Any real interval can be used instead of $[0,1]$.

The method proposed in this paper relies on smoothing spline representations of the unobserved functions. More precisely, let us consider $x \in \mathcal{H}^m$ sampled on the aforementioned grid. A smoothing spline estimate of x is the solution of

$$\widehat{x}_{\lambda,\tau_d} = \arg\min_{h \in \mathcal{H}^m} \frac{1}{|\tau_d|} \sum_{l=1}^{|\tau_d|} (x(t_l) - h(t_l))^2 + \lambda \int_0^1 (h^{(m)}(t))^2 dt, \qquad (1)$$

[1] extensions to more than two classes and to regression are straightforward.

[2] the sampling points depend on d and should therefore be denoted t_l^d, but this would clutter the notations.

where $\lambda > 0$ is a regularization parameter that balances interpolation errors and smoothness (measured by the L^2 norm of the m-order derivative of the estimate). The goal of this paper is to show that, under reasonable hypothesis, a classifier built on $\widehat{x}^{(m)}_{\lambda,\tau_d}$ is a consistent classifier. The proof is based on two steps. The first step constructs an inner product in $\mathbb{R}^{|\tau_d|}$ which can be used to approximate inner products performed on $\widehat{x}^{(m)}_{\lambda,\tau_d}$ using the sampled function only. The second step shows that the Bayes risk associated to $(\widehat{X}_{\lambda,\tau_d}, Y)$ converges to the Bayes risk of the original pair (X, Y), where $\widehat{X}_{\lambda,\tau_d}$ is the solution of equation (1) for the sampled function $X^{\tau_d} = \big(X(t_1), \ldots, X(t_{|\tau_d|})\big)^T$.

Those results stand as a consequence of several major properties of the smoothing splines: they asymptotically estimate the underlying function that generates the sampled curve, there is a one-to-one mapping between $\widehat{x}_{\lambda,\tau_d}$ and $\mathbf{x} = (x(t_1), \ldots, x(t_{|\tau_d|}))^T$ and the L^2-norm of the derivatives of $\widehat{x}_{\lambda,\tau_d}$ can be deduced from the norm of $\mathbf{x}$.

3 Smoothing splines and differentiation kernels

3.1 RKHS and smoothing splines

As the goal is to work on $\widehat{x}^{(m)}_{\lambda,\tau_d}$, it might seem natural to consider $\mathcal{H}^m$ with the metric induced by the inner product $(u, v) \mapsto \int_0^1 u^{(m)}(t)v^{(m)}(t)dt$. However, a slightly different structure is needed to ensure consistency. It is obtained by decomposing $\mathcal{H}^m$ into $\mathcal{H}^m = \mathcal{H}^m_0 \oplus \mathcal{H}^m_1$ [10], where $\mathcal{H}^m_0 = \operatorname{Ker} D^m = \mathbb{P}^{m-1}$ (the space of polynomial functions of degree less or equal to $m-1$) and $\mathcal{H}^m_1$ is an infinite dimensional subspace of $\mathcal{H}^m$ defined via m boundary conditions. The boundary conditions are given by a full rank linear operator from $\mathcal{H}^m$ to $\mathbb{R}^m$, denoted $B = (B^j)_{j=1}^m$, such that $\operatorname{Ker} B \cap \mathbb{P}^{m-1} = \{0\}$. Then,

$$\langle u, v\rangle^m_1 = \langle D^m u, D^m v\rangle_{L^2} = \int_0^1 u^{(m)}(t)v^{(m)}(t)dt$$

is an inner product on $\mathcal{H}^m_1$. Moreover, $\langle u, v\rangle^m_0 = \sum_{j=1}^m B^j u B^j v$ is an inner product on $\mathcal{H}^m_0$. Then a combined inner product on $\mathcal{H}^m$ is given by

$$\langle u, v\rangle_{\mathcal{H}^m} = \int_0^1 u^{(m)}(t)v^{(m)}(t)dt + \sum_{j=1}^m B^j u B^j v. \tag{2}$$

Equipped with $\langle ., .\rangle_{\mathcal{H}^m}$, $\mathcal{H}^m$ is a Reproducing Kernel Hilbert Space (RKHS, see e.g. [3]). The metric induced by $\langle ., .\rangle_{\mathcal{H}^m}$ consists in comparing functions based almost only on their m-order derivatives, up to a correction based on the boundary conditions. The most classical boundary conditions are given by $B^j h = h^{(j)}(0)$, $0 \leq j \leq m-1$, but others can be used [3] depending on the application.

3.2 Splines and sampled functions

Now, we need to link $\langle \widehat{u}_{\lambda,\tau_d}, \widehat{v}_{\lambda,\tau_d} \rangle_{\mathcal{H}^m}$ with $\mathbf{u} = (u(t_1), \ldots, u(t_{|\tau_d|}))^T$ and $\mathbf{v} = (v(t_1), \ldots, v(t_{|\tau_d|}))^T$. This can be done via a theorem from [10]. Compatibility assumptions between the sampling grid $\tau_d = (t_l)_{l=1}^{|\tau_d|}$ and the boundary conditions operator B are needed:

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

1. $|\tau_d| \geq m - 1$
2. *sampling points are distinct with* $0 \leq t_1 < \cdots < t_{|\tau_d|} \leq 1$
3. *the m boundary conditions B^j are linearly independent from the $|\tau_d|$ linear forms $h \mapsto h(t_l)$, for $l = 1, \ldots, |\tau_d|$ (defined on $\mathcal{H}^m$)*

Then, using a theorem from [10], one can show[3] the following corollary:

Corollary 1. *Under Assumption 1 there is a one-to-one bi-continuous mapping between $\mathbf{u}$ and $\widehat{u}_{\lambda,\tau_d}$. In addition, there is a symmetric and positive definite matrix $\mathbf{M}_{\tau_d}$ such that for any $\mathbf{u} = (u(t_1), \ldots, u(t_{|\tau_d|}))^T$ and $\mathbf{v} = (v(t_1), \ldots, v(t_{|\tau_d|}))^T$ in $\mathbb{R}^{|\tau_d|}$,*

$$\langle \widehat{u}_{\lambda,\tau_d}, \widehat{v}_{\lambda,\tau_d} \rangle_{\mathcal{H}^m} = \mathbf{u}^T \mathbf{M}_{\tau_d} \mathbf{v}. \tag{3}$$

This corollary defines an inner product on $\mathbb{R}^d$ which is equivalent to the one chosen on $\langle ., . \rangle_{\mathcal{H}^m}$. It is therefore possible to work implicitly on the m-order derivatives of the spline representation of sampled functions only by replacing the canonical Euclidean inner product by the one associated to $\mathbf{M}_{\tau_d}$. In practice, this is done simply by multiplying the sampled functions by $\mathbf{Q}_{\tau_d}$, the Cholesky decomposition of $\mathbf{M}_{\tau_d}$ ($\mathbf{Q}_{\tau_d}^T \mathbf{Q}_{\tau_d} = \mathbf{M}_{\tau_d}$) prior to submitting the obtained vectors to any consistent classification method (see Section 4.3).

4 Consistency

4.1 Spline approximation

From the sampled random function $X(t_1), \ldots, X(t_{|\tau_d|})$, one can reconstruct at best $\widehat{X}_{\lambda,\tau_d}$. To ensure consistency, $\widehat{X}_{\lambda,\tau_d}$ must be showed to converge to X. This problem has been studied extensively (see, e.g., [5]). Convergence is linked to a well chosen sampling grid, as a badly designed one would not convey in $\mathbf{x}$ enough information to recover x. As general hypothesis on the sampling grids are very technical, we limit ourselves in the present paper to a simple particular case:

Assumption 2. *For each d, the sampling grid $\tau_d = (t_l)_{l=1}^{|\tau_d|}$ is uniformly spaced.*

Then we have:

[3] Proofs are omitted dues to space constraints.

Theorem 1 ([5]). *Under Assumption 2, for* $\lambda_{|\tau_d|} = \mathcal{O}\left(|\tau_d|^{-2m/(2m+1)}\right)$,

$$\|\widehat{x}_{\lambda_{|\tau_d|},\tau_d} - x\|_{L^2}^2 = \mathcal{O}\left(|\tau_d|^{-2m/(2m+1)}\right).$$

Similar results are also available for the derivatives up to order m (included).

4.2 Conditional expectation approximation

Theorem 1 can be used to relate the Bayes risk for the classification problem (X, Y) to the one for $(\widehat{X}_{\lambda_{\tau_d},\tau_d}, Y)$ when $|\tau_d|$ goes to infinity, i.e. L^* to

$$L^*_{\mathcal{H}^m,\tau_d} = \inf_{\phi:\, \mathcal{H}^m \to \{-1,1\}} P(\phi(\widehat{X}_{\lambda_{\tau_d},\tau_d}) \neq Y).$$

We have the following corollary:

Corollary 2. *Under Assumptions 1 and 2, if*

1. $\mathbb{E}(\|D^m X\|_{L^2}^2)$ *is finite,*
2. ***or*** $\tau_d \subset \tau_{d+1}$,

then

$$\lim_{|\tau_d|\to\infty} L^*_{\mathcal{H}^m,\tau_d} = L^*.$$

Each condition corresponds to a specific proof. Condition 1 uses a general result from [6] and a more precise version of Theorem 1 (which relates the convergence speed with the L^2 norm of the m-order derivative of the functions). Condition 2 uses the martingale approach proposed in [4] together with Theorem 1 (this generalizes the results from [16]).

4.3 Wrapping up the results

Let us now consider any consistent classification algorithm for standard multivariate data, such as e.g., Support Vector Machines [15] or Multi-Layer Perceptrons [17]. Let us denote ϕ_{n,τ_d} the classifier obtained with this algorithm on the learning set $(\mathbf{Q}_{\tau_d} X_i^{\tau_d}, Y_i)_{i=1}^n$. We have the following theorem:

Theorem 2. *Under the assumptions of Corollary 2 and any additional assumptions needed by the multivariate algorithm,*

$$\lim_{|\tau_d|\to\infty} \lim_{n\to+\infty} \mathbb{E}\left(L^*(\phi_{n,\tau_d})\right) = L^*.$$

The proof is based on the following steps. First, as $\mathbf{Q}_{\tau_d}$ is a one-to-one continuous mapping, a consistent classifier on $(\mathbf{Q}_{\tau_d} X_i^{\tau_d}, Y_i)_{i=1}^n$ is also a consistent classifier on $(X_i^{\tau_d}, Y_i)_{i=1}^n$. As the chosen algorithm is consistent, $\mathbb{E}\left(L^*(\phi_{n,\tau_d})\right)$ converges with n to $L^*_{\tau_d} = \inf_{\phi:\, \mathbb{R}^d \to \{-1,1\}} P(\phi(X^{\tau_d}) \neq Y)$. According to Corollary 1, the mapping between X^{τ_d} and $\widehat{X}_{\lambda_{\tau_d},\tau_d}$ is also one-to-one and bi-continuous. Therefore $L^*_{\tau_d} = L^*_{\mathcal{H}^m,\tau_d}$. Corollary 2 gives the conclusion.

In addition to being consistent, this scheme is also "derivative based". Indeed, as shown by Corollary 2

$$\|\mathbf{Q}_{\tau_d}(\mathbf{u}-\mathbf{v})\|^2_{R^{|\tau_d|}} = \|\widehat{u}_{\lambda,\tau_d} - \widehat{v}_{\lambda,\tau_d}\|^2_{\mathcal{H}^m}\,.$$

Therefore, a classifier constructed on the $\mathbf{Q}_{\tau_d} X_i^{\tau_d}$ compares the underlying functions via the metric induced by the inner product of equation (2), that is mainly via the m-order derivatives of those functions, up to the boundary conditions. It gives therefore a theoretical background to the common practice of using derivatives for some spectrometric problems [7, 13, 14] and more generally in functional data analysis.

References

[1] B. K. Alsberg. Representation of spectra by continuous functions. *Journal of Chemometrics*, 7:177–193, 1993.

[2] A. Berlinet, G. Biau, and L. Rouvière. Functional classification with wavelets. *submitted*, 2006.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publisher, 2004.

[4] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172, 2005.

[5] D. Cox. Multivariate smoothing splines functions. *SIAM Journal on Numerical Analysis*, 21:789–813, 1984.

[6] T. Faragó and L. Györfi. On the continuity of the error distortion function for multiple-hypothesis decisions. *IEEE Transactions on Information Theory*, 21(4):458–460, July 1975.

[7] F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17(4), 2002.

[8] F. Ferraty and P. Vieu. *NonParametric Functional Data Analysis*. Springer, 2006.

[9] I. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148, 1993.

[10] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[11] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer Verlag, June 1997.

[12] F. Rossi and B. Conan-Guez. Functional multi-layer perceptron: a nonlinear tool for functional data analysis. *Neural Networks*, 18(1):45–60, January 2005.

[13] F. Rossi, N. Delannay, B. Conan-Guez, and M. Verleysen. Representation of functional data in neural networks. *Neurocomputing*, 64:183–210, March 2005.

[14] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7–9):730–742, March 2006.

[15] I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18(3):768–791, September 2002.

[16] N. Villa and F. Rossi. Un résultat de consistance pour des svm fonctionnels par interpolation spline. *Comptes Rendus Mathématiques*, 343(8):555–560, Octobre 2006.

[17] H. White. Connectionist nonparametric regression: Mutilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.