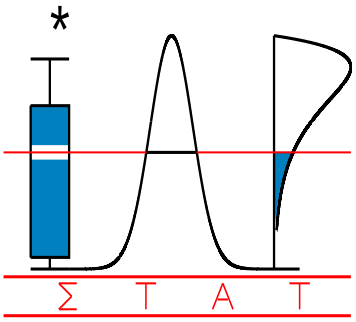
T E C H N I C A L

R E P O R T

11022

Using Bagadis in nonparametric functional data analysis: predicting from curves with sharp local features

TIMMERMANS, C., DELSOL, L. and R. von SACHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Using Bagidis in nonparametric functional data analysis: predicting from curves with sharp local features

Catherine Timmermans[a,*], Laurent Delsol[b], Rainer von Sachs[a]

[a]*Institute of Statistics, Biostatistics and Actuarial Sciences - Université catholique de Louvain - Voie du Roman Pays, 20 - BE-1348 Louvain-la-Neuve - Belgium.*
[b]*Laboratoire MAPMO, Université d'Orléans, UFR Sciences- Bâtiment de mathématiques - Rue de Chartres, B.P. 6759 - FR-45067 Orléans cedex 2 - France.*

Abstract

Our goal is to predict a scalar value or a group membership from the discretized observation of curves with sharp local local features that might vary both vertically and horizontally. To this aim, we propose to combine the use of the non parametric functional regression estimator developed by Ferraty and Vieu (2006) [1] with the Bagidis semimetric developed by Timmermans and von Sachs (2010) [2] in view of efficiently measuring dissimilarities between curves with sharp patterns. This association reveals powerful. Under quite general conditions, we obtain the rate of convergence of the nonparametric regression estimator in this case, as a function of the parameters of the Bagidis semimetric. We propose to optimize those parameters using a cross-validation procedure, and show the optimality of the selected vector. This last result has a larger scope and concerns the optimization of any vector parameter characterizing a semimetric used in this context. The performances of our methodology are assessed on simulated and real data examples. Results are shown superior than those obtained using competing semimetrics as soon as the variations of the significant sharp patterns in the curves have an horizontal component.

Keywords: functional data, nonparametric regression, semimetric, wavelet, misalignment, cross-validation

1. Introduction

Modern datasets often provide sets of points corresponding to discretized curves, typically time series or spectra. In this framework, the suitable information unit is the underlying curve instead of the vectorial quantity encoding the series - we refer to it as to a *functional data*. Functional statistical methods aim at taking this feature into account when extracting the information content of a dataset (see Ramsay and Silverman [3, 4], Bosq [5], Ferraty and Vieu [1], Ferraty and Romain [6], e.g.). Amongst them, nonparametric methods often rely on the availability of a suitable metric or semimetric for measuring differences amongst the curves of the dataset.

In this framework, the Bagidis semimetric has been introduced in Timmermans and von Sachs [2] as a highly adaptive wavelet-based tool for measuring dissimilarities between functional data. Its main originality is to be based upon the expansion of each series of a dataset into a *different* wavelet basis, one that is particularly suited for its hierarchical description. Measuring dissimilarities in such a way implies comparing not only the projections of the series onto the bases, as usual, but also the bases themselves. Because of this specificity, the semimetric is named Bagidis, which means *BAses GIving DIStances.* As a consequence of this feature, the Bagidis semimetric has the ability to capture the variations of patterns occurring in series along both the vertical and the horizontal axis. This property makes the semimetric particularly powerful when dealing with curves that might be affected simultaneously by horizontal shifts and vertical amplifications.

*Catherine Timmermans. Address: Institute of Statistics, Biostatistics and Actuarial Sciences - Université catholique de Louvain - Voie du Roman Pays, 20 - BE-1348 Louvain-la-Neuve - Belgium. Phone: 00 32 (0)10 47 42 81. Fax: 00 32 (0)10 47 30 32.

Email addresses: catherine.timmermans@uclouvain.be (Catherine Timmermans), laurent.delsol@univ-orleans.fr (Laurent Delsol), rainer.vonsachs@uclouvain.be (Rainer von Sachs)

Nonparametric functional data analysis techniques have been widely described in Ferraty and Vieu [1]. They include a set of prediction techniques that do not require to make an hypothesis on the form of the prediction operator - only smoothness hypotheses are made - and are able to efficiently deal with functional data provided it is used together with a semimetric able to extract the relevant features of the curves and satisfying some theoretical properties.

The purpose of the present study is to illustrate how we can advantageously make use of the Bagidis semimetric in the context of nonparametric functional prediction when the curves we are predicting from are characterized by some horizontally- and vertically-varying sharp local patterns. Our goal is also to show how this good behaviour is theoretically supported. Simulated examples are shown as well as a real data example, with spectrometric H-NMR data.

This paper is organized as follows. The statistical framework of our study is first described in Section 2: the notion of functional models for prediction is introduced, before focusing on nonparametric functional prediction; the Bagidis semimetric, which is the tool we propose to use in this setting, is then presented. Then, our two main theoretical results are stated in Section 3. The first one gives the rate of convergence of a nonparametric functional prediction using the Bagidis semimetric, as a function of its parametrization. A practical cross-validation approach for optimizing the parametrization of Bagidis in this context is deduced therefrom and theoretically validated. This second result has a more general scope as it theoretically supports for a cross-validated optimization of the parameters of any projection-based semimetric. Finally, simulated and real data examples are investigated in Section 4, so as to illustrate nonparametric prediction with Bagidis in action.

2. Statistical framework

This paper proposes to bring together two advanced statistical tools, in view of providing a new, efficient, way to predict from curves with sharp local patterns. Those tools are, on the one hand, the non parametric functional regression estimator provided by Ferraty and Vieu [1] for obtaining predictions from functional data and, on the other hand, the Bagidis semimetric developed by Timmermans and von Sachs [2] for comparing curves with sharp local patterns. As a support for our present work, those two tools are successively described in this Section, after a general presentation of the stakes and challenges of functional prediction.

2.1. Functional models for prediction

The general goal of a regression model is to link two random variables; the first one is a *response variable* $\boldsymbol{Y}$ which we are interested to explain; the second one is an *explanatory variable* $\boldsymbol{\chi}$ which is believed to be able to inform us about the response. Practically, this requires to estimate an unknown link operator r, based on some known pairs (χ_i, Y_i) - those pairs are called the *training set* - so as to be able to predict $\boldsymbol{Y}$ for any new χ using this link.

About the response variable. Regarding the response variable, two cases are most commonly encountered:

- **the response $\boldsymbol{Y}$ is a real measurement** whose we aim at predicting the conditional mean value for any given value of $\boldsymbol{\chi}$. In that case, the link function r between the two variables is an operator defined as the conditional expectation of the response variable, given the explanatory variable:

$$r(\boldsymbol{\chi}) = \mathbb{E}(\boldsymbol{Y}|\boldsymbol{\chi}),$$

  which can be equivalently rewritten as

$$\boldsymbol{Y} = r(\boldsymbol{\chi}) + \boldsymbol{\epsilon},$$

  where $\boldsymbol{\epsilon}$ is an error term such that $\mathbb{E}(\boldsymbol{\epsilon}|\boldsymbol{\chi}) = 0$. This is called a *regression model*. For any given value χ of $\boldsymbol{\chi}$, the associated scalar value Y is thus estimated as

$$\hat{Y} = \hat{r}(\chi),$$

  for an estimator $\hat{r}$ of r.

- **the response $\boldsymbol{Y}$ is a class membership** that we have to determine for any given value of $\boldsymbol{\chi}$. This is called a *discrimination model*. The link function we are looking at in that case is the probability of being member of a given class g given a value of $\boldsymbol{\chi}$:

$$r_g(\boldsymbol{\chi}) = P(\boldsymbol{Y} = g|\boldsymbol{\chi}) = \mathbb{E}(\mathbf{1}_{[\boldsymbol{Y}=g]}|\boldsymbol{\chi}),$$

  for each class index $g \in \{g_i\}_{i=1...G}$, with G the number of classes and $\mathbf{1}_{[\boldsymbol{Y}=g]} = 1$ if $\boldsymbol{Y} = g$ and 0 otherwise. For any given value χ of $\boldsymbol{\chi}$, the associated class membership Y is thus estimated as

$$\hat{Y} = \underset{g\in\{g_i\}_{i=1...G}}{\arg\max}(\hat{r}_g(\chi)),$$

  according to the Bayes rule, for an estimator $\hat{r}_g$ of r_g, $g \in \{g_i\}_{i=1...G}$.

As we can see, both regression and discrimination problems imply the estimation of an operator r or r_g that is defined as a conditional expectation. Consequently, the same statistical tools for its estimation may be involved in both cases.

About the explanatory variable. Classically, the explanatory variable has been a scalar or small-dimensional vector variable, and ways to estimate the operator r (or r_g) in that case have been known for a long time. New challenges and opportunities have appeared with *functional* explanatory variables taking their values in some function space $\mathcal{F}$ - i.e. explanatory variables being actually curves, typically spectra or functions of time. Of course, from a practical point of view, a curve is necessarily observed as a discretized spectrum or a time series, that is to say a vector. Nevertheless, the classical multivariate regression framework is often not convenient anymore. There are two reasons for that:

- **The use of the complete information at hand.** The knowledge that there is a dynamic process underlying the sampled data is an information that could be exploited for an optimal estimation of the prediction model.
- **The "curse of dimensionality".** To render the functional nature of the explanatory variable, it is often useful to make use of either a fine discretization, either a long period of data collection. As a consequence, the vectors representing the curves are rather large, leading to the need of a very large set of observations to correctly estimate the prediction model. In the nonparametric framework, this "curse of dimensionality" has been mathematically stated by Stone [7]:

  Theorem. *Consider a p times differentiable unknown regression function r of a N – dimensional variable, and $\hat{r}$ a nonparametric estimator of r based on a training set of size n. Then the optimal rate of convergence of $\hat{r}$ to r is*

$$\left(\frac{\log n}{n}\right)^{\frac{p}{2p+N}}. \tag{2.1}$$

To the best of our knowledge, the only way to predict from a curve nonparametrically while avoiding this curse of dimensionality consists in assuming that relevant information of reduced dimension can be extracted from the curves, the price to pay associated with this assumption being found in regularity assumptions on the regression operator. The knowledge of the functional nature of the data might help to efficiently extract this reduced dimensional information.

About the estimation of the link operator r, in case of functional data. Several ways have been recently proposed to take into account the functional nature of the explanatory variable in regression or discrimination problems. Works of Ramsay and Silverman [4], as well as Crambes et al. [8], Ramsay and Dalzell [9] and Cai and Hall [10], for instance, are devoted to *parametric* functional modelling and focus on linear models. Beside, other kinds of parametric or semiparametric regression models have been considered in Sood et al. [11], Ait-Saidi et al. [12] and Aneiros-Pérez and Vieu [13], for instance. On the other hand, researchers such as Ferraty and Vieu [1] explore *nonparametric* functional prediction models - i.e. techniques allowing to dispose of the need to make an hypothesis on the form of the regression operator. Only smoothness hypotheses are required. Our work places itself in this nonparametric context.

2.2. Estimation in a nonparametric functional prediction model

Suppose we have got a set of curves $\{\chi_i\}_{i=1..N}$ and associated scalar values $\{Y_i\}_{i=1..N}$, and we are looking for a prediction model

$$r(\chi) = \mathbb{E}(\boldsymbol{Y}|\chi),$$

for which we do not assume a particular parametric form, but only some regularity conditions. One of the basic assumptions underlying the concept of modelling is that similar values of χ correspond to similar value of Y. We need thus ways to quantify the similarity of the explanatory curves. Following Ferraty and Vieu [1], we will say that d is a semimetric on some space $\mathcal{F}$ as soon as

- $\forall \chi \in \mathcal{F}, \quad d(\chi, \chi) = 0,$
- $\forall \chi_i, \chi_j, \chi_k \in \mathcal{F}, \quad d(\chi_i, \chi_k) \leq d(\chi_i, \chi_j) + d(\chi_j, \chi_k).$

A semimetric is thus defined in the same way as a distance, except that $d(\chi_i, \chi_j) = 0 \not\Rightarrow \chi_i = \chi_j$, and distances are particular cases of semimetrics. Semimetrics allow for measuring dissimilarities between curves through a reduced number of components. A well-suited semimetric may thus be a tool for extracting the relevant features of a set of curves.

A semimetric-based Nadaraya-Watson regression estimator. Given this definition, Ferraty and Vieu [1] have proposed an extended Nadaraya-Watson estimator, that is able to deal with functional data provided we have a well-suited semimetric d such that some theoretical properties are satisfied. This estimator has the form:

$$\hat{r}(\chi) = \frac{\sum_{i=1}^{n} Y_i K(\frac{d(\chi,\chi_i)}{h})}{\sum_{i=1}^{n} K(\frac{d(\chi,\chi_i)}{h})}, \qquad \text{for regression problems,} \tag{2.2}$$

and, as a particular case,

$$\hat{r}_g(\chi) = \frac{\sum_{i=1}^{n} \mathbf{1}_{[Y_i=g]} K(\frac{d(\chi,\chi_i)}{h})}{\sum_{i=1}^{n} K(\frac{d(\chi,\chi_i)}{h})}, \qquad \text{for classification problems,} \tag{2.3}$$

where K is an asymmetric bounded kernel, n is the number of independent pairs (χ_i, Y_i) in the training set, d is a semimetric and h is the bandwidth.

The choice of the semimetric. It is clear that the predictive qualities of the estimated regression or discrimination model depends on the features extraction capacities of the chosen semimetric d. It also relies on the regularity of the regression operator r with respect to d. Definitely, it is clear that the choice of the semimetric is crucial and must be related to the particular features on the functional dataset at hand. Commonly used families of semimetrics [1] are :

- The derivative-based family of semimetrics d_q^{deriv} based on the derivatives of order q of the curves :

$$d_q^{deriv}(\chi_i, \chi) = \sqrt{\int (\hat{\chi}_i^{(q)}(t) - \hat{\chi}^{(q)}(t))^2 dt}, \tag{2.4}$$

  where $\hat{\chi}_i^{(q)}(t)$, $\hat{\chi}^{(q)}(t)$ are the estimations of the q^{th} derivative of χ_i and χ , respectively, at abscissa t, and where the integral has to be numerically estimated by a sum. Estimating the derivatives usually rely on a smoothing of the data. This family includes the Euclidean distance L_2 between the smoothed curves as a particular case, with $q = 0$; it will be denoted d_0^{deriv}. A contrario, the classical vectorial euclidean L_2 distance between the unsmoothed observations will be shortly referred to as d^{L_2}.

- The PCA-based family of semimetrics d_q^{PCA} based on a certain number q of principal components of the dataset:

$$d_q^{PCA}(\chi_i, \chi) = \sqrt{\sum_{k=1}^{q} (\int (\chi_i(t) - \chi(t)) \hat{v}_k(t) dt)^2}, \tag{2.5}$$

  where $\hat{v}_k$, $k = 1..q$ is the k^{th} estimated eigenfunction of the principal component analysis, and where the integral has to be numerically estimated by a sum.

- The *hshift* semimetric d^{hshift} that realigns curves before computing d_0^{deriv} distances between them.

Those semimetrics have been shown useful in various problems [1, 14, 15]. However, they happen to fail when dealing with curves with sharp local features that might not be well aligned from one curve to another one, as discussed by Timmermans and von Sachs [2]. Besides, the computation of those semimetrics relies on a smoothing of the data, which is generally problematic for curves with abrupt patterns. This difficulty with curves with sharp patterns will be illustrated in subsequent examples of regression problems, in Section 4. However, such curves with sharp patterns happen to be dealt with in a large variety of scientific area (spectrometric curves, time series ...), so that it is worth thinking of the use of another, better adapted, semimetric for those data.

2.3. Bagidis*, a semimetric for comparing curves with sharp horizontally- and vertically-varying local features*

The Bagidis semimetric d_p^B has been introduced by Timmermans and von Sachs [2] so as to measure differences between regularly discretized curves that are characterized by some sharp local features. It is a functional data-driven and wavelet-based measure that is highly adaptive to the curves being considered. It has been proved to be a semimetric by Timmermans and von Sachs [2]. Key ideas are as follows.

Looking for a hierarchical description of the patterns of the series. We consider series observed on a regular grid $\mathbb{N}_{[1;N]}$. When we evaluate dissimilarities between series visually, we intuitively investigate first the global shapes of the series for estimating their resemblance, before refining the analysis by comparing the smaller features of the series. In other words our comparison is based upon a hierarchical comprehension of the curves. This visual approach inspired us to define our semimetric: we expand each series in a (different, series-adapted) basis that describes its features hierarchically, in the sense that the first basis vectors carry the main features of the series while subsequent basis vectors support less significant patterns; afterwards, we compare both the bases and the expansions of the series onto those bases, rank by rank, according to the hierarchy.

Expanding each series of the dataset in the Unbalanced Haar Wavelet Basis *that is best suited for the hierarchical description of its shape.* The family of *Unbalanced Haar Wavelet Bases* has been introduced by Girardi and Sweldens [16]. It consists in orthonormal bases that are made of one constant vector and a set of Haar-like (i.e. *up-and-down* shaped) orthonormal wavelets whose discontinuity point (hereafter the breakpoint) between the positive and negative parts is not necessarily located at the middle of its support. The *Bottom Up Unbalanced Haar Wavelet Transform* (Buuhwt), an algorithm that was developed by Fryzlewicz [17], allows for selecting amongst this family of bases the best basis for describing a given series hierarchically. Beside this hierarchical organization, the selected basis inherits the good capacity of Haar wavelets to efficiently capture sharp patterns.

We denote the expansion of a series χ_i in that basis as

$$\chi_i = \sum_{k=0}^{N-1} d_i^k \boldsymbol{\psi}_i^k, \tag{2.6}$$

where the coefficients d_i^k (hereafter the *detail* coefficients) are the projections of the series χ_i on the corresponding basis vectors $\boldsymbol{\psi}_i^k$ and where the set of vectors $\{\boldsymbol{\psi}_i^k\}_{k=0...N-1}$ is the Unbalanced Haar wavelet basis that is best suited to the series χ_i, as obtained using the Buuhwt algorithm. Besides, we denote b_i^k, the breakpoint of the wavelet $\boldsymbol{\psi}_i^k$, at every rank $k \neq 0$.

Defining a semimetric by taking advantage of the hierarchy of those expansions . As shown by Fryzlewicz [17], the ordered set of breakpoints $\{b_i^k\}_{k=1...N-1}$ determines the basis $\{\boldsymbol{\psi}_i^k\}_{k=0...N-1}$ uniquely. As a consequence, the set of points

$$\{z_i^k\}_{k=1...N-1} = \{(b_i^k, d_i^k)\}_{k=1...N-1} \tag{2.7}$$

determines the shape of the series χ_i uniquely - i.e. it determines the series on the grid $\mathbb{N}_{[1;N]}$, except for a change of the mean level of the series, that is encoded by the additional coefficient d_i^0. It is the signature of the series in the

breakpoints-details plane. Given that, and with the definition $b_i^0 = 0$ for each curve, we define the Bagidis semimetric as a 2-norm (weighted) distance in the *breakpoints-details* plane:

$$d^B_{w_k}(\chi_1, \chi_2) = \sum_{k=0}^{N-1} w_k \left\| z_1^k - z_2^k \right\|_2 = \sum_{k=0}^{N-1} w_k \left(\left| b_1^k - b_2^k \right|^2 + \left| d_1^k - d_2^k \right|^2 \right)^{1/2}$$

where w_k, $k = 0 \dots N-1$, are well suited weights. As such, this semimetric takes advantage of the hierarchy of the well adapted unbalanced Haar wavelet bases: breakpoints and details of similar rank k in the hierarchical description of each series are compared to each other, and the resulting differences can be weighted according to that rank. As the breakpoints point to level changes in the series, the term $\left| b_1^k - b_2^k \right|$ can be interpreted as a measure of the difference of location of the features, along the horizontal axis. Being a difference of the projections of the series onto wavelets that encode level changes, the term $\left| d_1^k - d_2^k \right|$ can be interpreted as a measure of the differences of the amplitudes of the features, along the vertical axis. At rank $k = 0$, $\left| b_1^0 - b_2^0 \right|$ vanishes and $\left| d_1^0 - d_2^0 \right|$ measures the difference between the means of the curves.

Investigating the balance between breakpoints and details differences. We introduce an extension of the Bagidis semimetric as follows:

$$d^B_{w_k,\lambda}(\chi_1, \chi_2) = \sum_{k=0}^{N-1} w_k \left(\lambda \left| b_1^k - b_2^k \right|^2 + (1-\lambda) \left| d_1^k - d_2^k \right|^2 \right)^{1/2} \tag{2.8}$$

with $\lambda \in [0; 1]$. This parameter λ actually defines a scaling in the *breakpoints-details* plane, and hence in the original units of the problem. Setting λ at its extreme values 0 or 1 allows to investigate the contributions of the breakpoints differences and details differences separately. Moreover, the presence of this parameter allows the semimetric to be robust with respect to scaling effects: if λ is optimized according to a given criteria (such as the mean square error of a prediction model), the relative dissimilarities between the series of a dataset will remain the same, whatever the scales of measurements along the horizontal and vertical axes, so that the predictive qualities of the model will not be affected by such a change in the units of measurements. This variant (2.8) of the Bagidis semimetric is the one that we will use throughout this paper. For the sake of simplicity, we will simply denote it by d^B.

Choosing the weights. In a prediction setting, weights should ideally be positive at rank k if that rank carries information for discriminating the series, and 0 otherwise. In such a way, the weights could act as a filter that extract the part of the distances between the curves that carries relevant features. This paper will illustrate and validate that the weights could easily be obtained, altogether with the balance parameter and the bandwidth, using a cross-validation procedure across a set of possible values, in the framework of non parametric functional prediction.

3. The main results

This Section states the two main theoretical results of this paper, that support for the use of Bagidis in nonparametric functional regression. First, we obtain the rate of convergence of the nonparametric regression estimator (2.2) used with the Bagidis semimetric under suitable conditions. We see that this rate of convergence is related to the sparsity of the weight function that parametrizes the Bagidis semimetric in equation (2.8). Consequently, we propose to use a cross-validation procedure so as to optimize this weight function, as well as the balance parameter λ in equation (2.8) and the bandwidth h in equation (2.2). Our second result is the asymptotic optimality of this method.

3.1. Rate of convergence of estimator (2.2) *when used together with* Bagidis.

Our first result is the rate of pointwise convergence of the nonparametric functional regression estimator (2.2) (and (2.3) as a particular case) used together with the Bagidis semimetric. Under quite general conditions, we show that we can reach the rate of convergence

$$\left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+K}},$$

with β being a Lipzschitz parameter quantifying the smoothness of r, n the number of curves in the training set and K the number of non-zero weights w in the Bagidis semimetric (2.8). This rate of convergence is to be compared with the rate of convergence (2.1) for a nonparametric multivariate regression directly based on a N-dimensional variable:

$$\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+N}},$$

with p the order of differentiability of r and N the length of the discretized curve. This indicates that we can reach a quite good rate of convergence with $K \ll N$, provided we can restrict ourselves to a sparse enough weight function while satisfying regularity conditions on r. This may happen if the number of significant features in the curves of the dataset is not too large.

Our result relies on the following set of assumptions and definitions.

About the random curve $\boldsymbol{\chi}$. The random curve $\boldsymbol{\chi}$ is observed on a regular grid $\mathbb{N}_{[1;N]}$. It is defined as a function of the breakpoints and details parameters (2.7) defining its Buuhwt expansion (2.6):

$$\boldsymbol{\chi} = \chi(\boldsymbol{b}, \boldsymbol{d}), \text{ with } (\boldsymbol{b}, \boldsymbol{d}) \in \mathbb{N}^N_{[0;N]} \times \mathbb{R}^N. \tag{3.1}$$

Let us recall that the random signature $(\boldsymbol{b}, \boldsymbol{d})$ characterizes the random curve $\boldsymbol{\chi}$ uniquely on the grid $\mathbb{N}_{[1;N]}$.

About the point of prediction χ. We denote by

$$\chi = \chi(b, d) \tag{3.2}$$

the fixed curve for which we want to obtain a prediction. This curve is uniquely related to the fixed point (b, d) in the breakpoint-detail plane.

About the response $\boldsymbol{Y}$. We assume that $\boldsymbol{Y}$ is a scalar variable and that there exists $\sigma_m(.)$ continuous at χ such that

$$\forall m \geq 2,\ \mathbb{E}(|\boldsymbol{Y}|^m | \boldsymbol{\chi} = \chi) \leq \sigma_m(\chi). \tag{3.3}$$

About the dataset. We assume to have n independent observations

$$(\chi_i, Y_i)_{i=1\ldots n} \tag{3.4}$$

of the random pair $(\boldsymbol{\chi}, \boldsymbol{Y})$.

About the Bagidis semimetric. We denote by d^B the Bagidis semimetric (2.8) with given balance parameter λ and weight function $w = \{w_k\}_{k=0\ldots N-1}$. We denote by $\mathcal{K}$ the set of indexes of non-zero components in w, and K the cardinality of this set. We assume that the non-zero weights are strictly positive:

$$\forall k \in \mathcal{K}, \quad w_k > 0. \tag{3.5}$$

About the probability distribution of $\boldsymbol{\chi}$. We denote by $f^{\mathcal{K}}_{\boldsymbol{d}|b}(\boldsymbol{d})$ the conditional density function of $\boldsymbol{d}$ given $\boldsymbol{b} = b$, restricted on the (b^k, d^k) such that $k \in \mathcal{K}$. We assume that, at the fixed point (b, d), $f^{\mathcal{K}}_{\boldsymbol{d}|b}(\boldsymbol{d})$ is strictly positive and continuous with respect to d^{Bagidis} : for all ϵ positive, there exists δ_ϵ positive such that

$$d^B(\chi(b, d), \chi(b, \boldsymbol{d})) \leq \delta_\epsilon \text{ implies } |f^{\mathcal{K}}_{d|b}(d) - f^{\mathcal{K}}_{\boldsymbol{d}|b}(\boldsymbol{d})| \leq \epsilon. \tag{3.6}$$

We also assume that the curves $\chi(b, d)$ and $\chi(b, \boldsymbol{d})$ can have the same breakpoints for all ranks $k \in \mathcal{K}$ with a non-vanishing probability:

$$P(\forall k \in \mathcal{K},\ \boldsymbol{b}^k = b^k) > 0. \tag{3.7}$$

This is possible because the breakpoints take their values on a finite grid of values. According to Ferraty and Vieu [1], we define the small ball probability of $\boldsymbol{\chi}$ around χ as

$$\phi_{d,\chi}(h) = P(\boldsymbol{\chi} \in B_d(\chi, h)),$$

where $B_d(\chi, h)$ is the ball of radius h centered on χ and defined according to the semimetric d. We assume that

$$\forall \epsilon > 0,\ \phi_{d^B,\chi}(\epsilon) > 0 \tag{3.8}$$

About the regression operator. We assume that there exists β positive such that

$$r \in Lip_{\mathcal{F},\beta} \equiv \left\{ f : \mathcal{F} \to \mathbb{R},\ \exists C \in \mathbb{R}_0^+,\ \forall \chi' \in \mathcal{F},\ |f(\chi) - f(\chi')| < C\, d^B(\chi,\chi')^\beta \right\} \tag{3.9}$$

About the kernel. We assume that K is a kernel function from $\mathbb{R}$ to $\mathbb{R}^+$ such that $\int K = 1$. We assume that there exists positive constants C and C' such that

$$C\phi_{d^B,\chi}(h) \leq \mathbb{E}\Big(K\Big(\frac{d^B(\chi,\boldsymbol{\chi})}{h}\Big)\Big) \leq C'\phi_{d^B,\chi}(h) \tag{3.10}$$

About the bandwidth. The bandwidth h is chosen according to a positive sequence h_n related to n in such a way that

$$\lim_{n\to\infty} h_n = 0 \ \text{ and } \ \lim_{n\to\infty} \frac{\log n}{n\phi_{d^B,\chi}(h_n)} = 0. \tag{3.11}$$

About the type of convergence. Following Ferraty and Vieu [1], we consider almost complete convergence. One says that the stochastic sequence $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ converges almost completely to the real random variable $\boldsymbol{X}$ if and only if for all ϵ positive, we have

$$\sum_{n\in\mathbb{N}} P(|\boldsymbol{X}_n - \boldsymbol{X}| > \epsilon) < \infty,$$

and we denote it by $\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}$ *a.co*. Moreover, one says that the rate of almost complete convergence of $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ to $\boldsymbol{X}$ is u_n if and only if there exists ϵ_0 positive such that

$$\sum_{n\in\mathbb{N}} P(|\boldsymbol{X}_n - \boldsymbol{X}| > \epsilon_0 u_n) < \infty,$$

and we write $\boldsymbol{X}_n - \boldsymbol{X} = O_{a.co}(u_n)$. A direct application of the Borel-Cantelli Lemma allows to prove that Ferraty and Vieu [1] that almost complete convergence implies almost sure convergence and convergence in probability, and that

$$\boldsymbol{X}_n - \boldsymbol{X} = O_{a.co}(u_n) \text{ implies } \boldsymbol{X}_n - \boldsymbol{X} = O_{a.s.}(u_n) \text{ and } \boldsymbol{X}_n - \boldsymbol{X} = O_p(u_n).$$

We note that most of those conditions are very general and are not specific to the use of (2.2) with Bagidis. Only conditions (3.5), (3.6) and (3.7) are specific to the Bagidis semimetric and to the expansion of the curves in the *breakpoints-details* plane. Given those assumptions, our result is stated as follows.

Theorem 1. *Given assumptions* (3.1) *to* (3.11)*, the functional kernel regression estimate* (2.2) *used together with the* Bagidis *semimetric is such that*

$$\hat{r}(\chi) - r(\chi) = O(h^\beta) + O_{a.co.}\left(\sqrt{\frac{\log n}{n\, h^K}}\right).$$

In particular, it can reach the rate of pointwise almost complete convergence

$$\hat{r}(\chi) - r(\chi) = O_{a.co.}\left(\Big(\frac{\log n}{n}\Big)^{\frac{\beta}{2\beta+K}}\right).$$

The proof of this Theorem is given in AppendixA. An idea is as follows.

It is shown in Ferraty and Vieu [1] that the rate of convergence of estimator (2.2) is linked with the semimetric through the behaviour of the small ball probability $\phi_{d,\chi}(h)$ about 0: the higher $\phi_{d,\chi}(h)$, the faster the rate of convergence. This function $\phi_{d,\chi}(h)$ measures the concentration of the functional variable $\boldsymbol{\chi}$, according to the topology defined by the semimetric. Given this, our proof relies on two steps. First, we determine the behaviour of the small ball probability around 0, for the Bagidis semimetric, as a function of the number of non-zero weights in its definition. We see that there exists $C > 0$ such that $\phi_{d^B,\chi}(h) \sim C\, h^K$, when h tends to zero. This means that $\boldsymbol{\chi}$ is fractal of order K around χ. Then, we make use of a result of Ferraty and Vieu [1] that gives the rate of convergence of estimator (2.2) as a function of the small ball probability of fractal random variables.

The key step in this proof is thus the study of $\phi_{d^B,\chi}(h)$, the probability for a curve $\boldsymbol{\chi}$ to be in a small ball of radius h around χ, according to the semimetric d^B. As soon as we consider curves $\boldsymbol{\chi}$ at distances h smaller than the minimum

of the weights for $k \in \mathcal{K}$ (which is strictly positive because of condition (3.5)), it is clear that their breakpoints $\boldsymbol{b}^{\boldsymbol{k}}$ must be the same as the breakpoints b^k of the curve χ at which we want to predict, for all $k \in \mathcal{K}$. Indeed, if this were not true, there would exist at least one $k^* \in \mathcal{K}$ such that $|b^{k^*} - \boldsymbol{b}^{k^*}| \geq 1$, as the step of the grid is 1, so that the distance between the curves would be strictly greater than

$$w_{k^*}|b^{k^*} - \boldsymbol{b}^{k^*}| \geq w_{k^*} \geq \min_{k \in \mathcal{K}} w_k,$$

which leads to a contradiction for h small enough. This is the reason for condition (3.7). It is also the reason why our proof requires only the continuity of the conditional density function $f_{|b}$ of $(\boldsymbol{b}, \boldsymbol{d})$ given $\boldsymbol{b}$ at (b, d) (equation (3.6)). Finally, the reason for the cardinality K of $\mathcal{K}$ to appear in the behavior of the small ball probability and hence in the rate of convergence is that the volume of a ball of radius h and dimension K is proportional to h^K.

In the second step of our analysis, we refer to Theorem 6.11 in Ferraty and Vieu [1], that states that

$$\hat{r}(\chi) - r(\chi) = O(h^\beta) + O_{a.co.}\left(\sqrt{\frac{\log n}{n\,\phi_{\chi,d}(h)}}\right). \tag{3.12}$$

This Theorem relies on conditions (3.3), (3.8), (3.9), (3.10) and (3.11) As in the classical multivariate setting, the first component on the right hand side of equation (3.12) is related to the bias of the estimate and depends only on the smoothness of the operator r. This component is controlled through condition (3.11, *left*). Similarly, the second component on the right hand side of equation (3.12) is related to the variance of the estimate. It is controlled through condition (3.11, *right*). In the case of a fractal type variable, with $\phi_{d^B,\chi}(h) \sim C\,h^K$ when h tends to zero, this term becomes

$$O_{a.co.}\left(\sqrt{\frac{\log n}{n\,h^K}}\right),$$

which proves the first part of Theorem 1. The second part of the Theorem is then proved by choosing the bandwidth used in (3.12) as

$$h \sim C\left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+K}}.$$

We note that condition (3.11, *right*) is automatically satisfied in that case.

We finally mention here two possible ways to generalize Theorem 1. First, we could combine our result about the fractality of the random variable χ with results of Ferraty et al. [18], so as to obtain the rate of uniform almost complete convergence of the functional kernel regression estimate (2.2) used with the Bagidis semimetric. Second, we could consider the use of the Bagidis semimetric in a more general nonparametric regression setting, with a functional response variable, in the framework provided by Ferraty et al. [19].

3.2. Asymptotic optimality of a cross-validated choice of the parameters of the semimetric and the bandwidth

As a second result, we propose, and theoretically support, the selection of a relevant weight function by using a leave-one-out cross-validation procedure, with a mean square error minimization criterion .

The cross-validated leave-one-out mean square error minimization criterion in nonparametric functional regression

We consider the estimator given by equation (2.2). The bandwidth h has to be specified in this expression. Moreover, in case d designates the Bagidis semimetric, it is parametrized by the balance parameter λ and the weight function w. Consequently, $\hat{r}$ relies on a vectorial parameter $H = (h, \lambda, w) \in \mathbb{R}^{N+2}$. We propose to chose this vectorial parameter H amongst a set $\mathcal{H}_n$ of possibilities by using a cross-validation procedure with a leave-one-out mean square error (MSE) criterion.

This cross-validation based approach for optimizing $\hat{r}$ generalizes the ideas of Rachdi and Vieu [20] and Ait-Saidi et al. [12]. Rachdi and Vieu [20] use a leave-one-out cross-validated MSE minimizer for choosing the bandwidth h, once a semimetric has been fully specified. They have shown the optimality of this procedure. For our purpose, not only the bandwidth but also the parameters λ and w defining the semimetric within a given family of semimetrics have to be optimized. A leave-one-out cross-validated selection of a parameter specifying a semimetric used within $\hat{r}$ has

been proved asymptotically optimal in the particular case of a single functional index model [12]. In what follows, we generalize those results to families of semimetrics determined by a vectorial parameter. This more general framework includes the optimization of $\hat{r}$ used not only with the Bagidis semimetric, but also with any kind of projection-based semimetric for which we aim at selecting the components (or the number of first components) upon which we project the series for their comparison. Given the importance of the class of projection-based semimetric (see Ferraty and Vieu [1, Chapter 13]), the opportunity to derive such a general result is clear.

Notations and main ideas

We denote by $\hat{r}_H(\chi)$ the regression operator estimator $\hat{r}(\chi)$ used with the fixed parameter H. For the sake of simplicity, we denote

$$\Delta_i(\chi) = K\Big(\frac{d(\chi,\chi_i)}{h}\Big) \quad \text{and} \quad K_H(\chi,\chi_i) = \frac{\Delta_i(\chi)}{\mathbb{E}\Big(\Delta_i(\chi)\Big)},$$

with d, a semimetric parametrized by $H\backslash h$. Consequently, $\hat{r}_H(\chi)$ is denoted by

$$\hat{r}_H(\chi) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i K_H(\chi,\chi_i)}{\frac{1}{n}\sum_{i=1}^{n} K_H(\chi,\chi_i)}. \tag{3.13}$$

Our criterion for measuring the quality of $\hat{r}_H(\chi)$ is the Mean Integrated Square Error (MISE) defined as follows:

$$MISE(H) \equiv MISE(\hat{r}_H, r) = \mathbb{E}\Big(\int (\hat{r}_H(\chi) - r(\chi))^2\, W(\chi)\, dP_\chi(\chi)\Big), \tag{3.14}$$

where $W(\chi)$ is a non negative weight function and P_χ is the probability distribution measure of the functional variable χ. We define $\mathcal{H}_n$, a set of possible values for the parameter H, with the cardinality of $\mathcal{H}_n$ increasing with the sample size. We aim at selecting

$$H^* = \arg\min_{H\in\mathcal{H}_n} MISE(H).$$

However H^* cannot be obtained as r is unknown in expression (3.14). Consequently, we propose to estimate H^* by H^{CV}, defined as follows:

$$H^{CV} = \arg\min_{H\in\mathcal{H}_n} CV(H),$$

where $CV(H)$ is the leave-one-out cross-validated criterion defined by

$$CV(H) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{r}_H^{-j}(\chi_j))^2 W(\chi_j),$$

with

$$\hat{r}_H^{-j}(\chi) = \frac{\sum_{i=1, i\neq j}^{n} Y_i K_H(\chi,\chi_i)}{\sum_{i=1, i\neq j}^{n} K_H(\chi,\chi_i)}.$$

$\hat{r}_H^{-j}(\chi_j)$ is thus the prediction associated to χ_j based upon the regression estimator (3.13) applied to our dataset whose the pair (χ_j, Y_j) has been excluded, with the parameter H. $CV(H)$ is thus an estimation of $MISE(H)$ calculated over the dataset.

The main result

The main result of this Section relies on the following conditions.

About the dataset. We assume to have n independent observations

$$(\chi_i, Y_i)_{i=1\dots n} \tag{3.15}$$

of the random variable (χ, Y).

About the kernel. The kernel K is bounded with compact support $[0;1]$, Lipszchitz on $\mathbb{R}^+$, and we have that for all $j = 1, 2\dots$, there exist $C_{1j}, C_{2j} > 0$ such that for all $H \in \mathcal{H}_n$, there exists $0 < \Phi_H \leq 1$ so that

$$\forall \chi, \chi_i \in \mathcal{W}, \quad C_{1,j}\Phi_H \leq \mathbb{E}\Big(K^j\Big(\frac{d(\chi,\chi_i)}{h}\Big)\Big) \leq C_{2,j}\Phi_H. \tag{3.16}$$

About the probability distribution of χ. We have that

$$\exists \gamma > 0,\ \exists C_1 > 0 \text{ such that } \sup_{H\in\mathcal{H}_n} \Phi_H \leq C_1 n^{-\gamma}, \tag{3.17}$$

and

$$\exists \delta > 0,\ \exists C_2 > 0 \text{ such that } \inf_{H\in\mathcal{H}_n} n\Phi_H \geq C_2 n^{\delta}. \tag{3.18}$$

About the weight function . The weight function $W(.)$ is non negative, of compact support $\mathcal{W} \subset \mathcal{F}$, bounded by some positive constant C_W, and such that

$$0 < \int W(\chi) dP_\chi(\chi). \tag{3.19}$$

The interior of $\mathcal{W}$ is non-empty and we have, for all $H \in \mathcal{H}_n$,

$$\mathcal{W} \subset \bigcup_{k=1}^{d_n} B(c_k, r_n), \tag{3.20}$$

where $B(c_k, r_n)$ are balls of $\mathcal{F}$, of center c_k and radius $r_n = o\Big(\inf_{H\in\mathcal{H}_n} h\Phi_H\Big)$, with $d_n \leq Cn^\eta$, $\eta > 0$.

About the conditional distribution of the errors. The conditional mean of the errors is zero:

$$\mathbb{E}(\boldsymbol{\epsilon}|\chi) = 0. \tag{3.21}$$

The conditional variance of the error is positive and there exists $\sigma_0 > 0$ such that

$$\mathbb{E}(\boldsymbol{\epsilon}^2|\chi) \geq \sigma_0^2. \tag{3.22}$$

About the regression operator. The regression operator r is bounded by some positive constant C_r. We introduce the following definitions of the bias and the integrated square bias:

$$B(\chi) = \mathbb{E}((Y_i - r(\chi))K_H(\chi,\chi_i)|\chi) \quad \text{and} \quad b_H = \int B^2(\chi)W(\chi)dP_\chi(\chi).$$

There exists a positive constant C_B such that for all $H \in \mathcal{H}_n$,

$$\forall \chi, \chi' \in \mathcal{F} \text{ such that } d(\chi,\chi') \leq h, \text{ we have } |r(\chi) - r(\chi')| \leq C_B b_H^{\frac{1}{2}}. \tag{3.23}$$

This property is satisfied as soon as there exists constants $C,\ C' > 0$ such that $\forall \chi, \chi' \in \mathcal{F}$ such that $d(\chi,\chi') \leq h$, there exists $\beta_H \geq 0$ so that

$$|r(\chi) - r(\chi')| \leq C' h^{\beta_H} \quad \text{and} \quad \int B^2(\chi)W(\chi)dP_\chi(\chi) \geq C' h^{2\beta_H}.$$

About the set of parameters. The cardinality of the set of parameters $\mathcal{H}_n$ is increasing at most algebraically fast:

$$\exists \alpha > 0,\ \exists C > 0 \text{ such that } \#\mathcal{H}_n \leq C n^{\alpha}. \tag{3.24}$$

About the conditional moments of Y. We assume that Y is such that

$$\forall k = 1, 2, \ldots\ \exists C_k > 0 \text{ such that } \mathbb{E}(|Y|^k | \chi) \leq C_k, \tag{3.25}$$

and

$$\forall \chi, \chi_i \in \mathcal{W},\ \forall k, l = 1, 2, \ldots\ \exists C_{kl} > 0 \text{ such that } \mathbb{E}\Big(|Y|^k K^l(\frac{d(\chi, \chi_i)}{h})\Big) \leq C_{kl}\Phi_H. \tag{3.26}$$

This last property is valid as soon as conditions (3.16) and (3.25) are satisfied.

Theorem 2. *Assuming conditions* (3.15) *to* (3.26)*, we have*

$$\frac{MISE(H^{CV})}{MISE(H^*)} \longrightarrow 1 \quad a.s.$$

This Theorem states the asymptotic optimality of the cross-validated choice H^{CV} amongst the set $\mathcal{H}_N$. It is proved in AppendixA. An idea of the proof is as follows.

In a view to prove Theorem 2, we need to introduce the quantity

$$MISE^*(H) = \int \mathbb{E}\Big((\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi))^2\Big) W(\chi) dP_\chi(\chi) = \int \mathbb{E}\Big((\frac{1}{n}\sum_{i=1}^{n} \delta_{i\chi})^2\Big) W(\chi) dP_\chi(\chi),$$

with

$$\hat{r}_{1H}(\chi) = \frac{1}{n}\sum_{i=1}^{n} K_H(\chi, \chi_i), \quad \hat{r}_{2H}(\chi) = \frac{1}{n}\sum_{i=1}^{n} Y_i K_H(\chi, \chi_i) \text{ and } \delta_{i\chi} = (Y_i - r(\chi))K_H(\chi, \chi_i).$$

Two results are then needed for this quantity $MISE^*(H)$:

$$\sup_{H \in \mathcal{H}_n} \left| \frac{MISE(H) - MISE^*(H)}{MISE^*(H)} \right| = o_{a.s.}(1) \tag{3.27}$$

$$\frac{MISE^*(H^{CV})}{MISE^*(H^*)} \longrightarrow 1 \quad a.s. \tag{3.28}$$

The proof is then rather short: first, (3.27) is used to deal with $MISE^*(H)$ instead of $MISE(H)$, then $\left|\frac{MISE^*(H^{CV}) - MISE^*(H^{CV})}{MISE^*(H^*)}\right|$ is bounded above by a sequence tending to zero when n tends to infinity, because of (3.28). The proof of (3.27) is the purpose of Lemma 13. It requires that

$$\forall H \in \mathcal{H}_n,\ \exists C, C' > 0, \text{ such that } \quad \frac{C}{n\Phi_H} + \frac{n-1}{n} b_H \leq MISE^*(H).$$

This bound from below of $MISE^*(H)$ is shown valid by Lemma 5. The proof of (3.28) is given by Lemma 14. This Lemma relies on the following inequality:

$$\left|\frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)}\right|\Big(1 - T_\alpha - T_\beta - T_\gamma\Big) \leq \Big(T_\alpha + T_\beta + T_\gamma\Big)\frac{2}{1 - T_\alpha} + 2T_\alpha,$$

with

$$T_\alpha = \sup_{H \in \mathcal{H}_n}\left|\frac{MISE^*(H) - ASE(H)}{MISE^*(H)}\right|,\ T_\beta = \sup_{H \in \mathcal{H}_n}\left|\frac{ASE(H) - \widetilde{ASE}(H)}{MISE^*(H)}\right| \text{ and } T_\gamma = 2\sup_{H \in \mathcal{H}_n}\left|\frac{CT(H)}{MISE^*(H)}\right|, \tag{3.29}$$

and the following definitions of the Average Square Error

$$ASE(H) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_H(\chi_i) - r(\chi_i)\right)^2 W(\chi_i),$$

of the Average Square Error of the leave-one-out predictor

$$\widetilde{ASE}(H) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{r}_H^{-j}(\chi_j) - r(\chi_j)\right)^2 W(\chi_j),$$

and with the following definition of $CT(H)$:

$$CT(H) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - r(\chi_j)\right)\left(\hat{r}_H^{-j}(\chi_j) - r(\chi_j)\right) = \frac{1}{n}\sum_{j=1}^{n}\epsilon_j\hat{\epsilon}_j^{-j}.$$

It is thus necessary to show that T_α, T_β and T_γ converge to zero almost surely when n goes to infinity. Those terms are dealt with separately through Lemmas 10, 11 and 12 respectively. Lemma 6 validates a condition that is needed for the proofs of those three Lemmas. Lemmas 8 and 9 provide with decompositions that have a role in the proof of Lemma 10. Those decompositions allow to highlight terms that have a form similar to the ones identified by Marron and Hardle [21] for a similar problem in the multivariate setting. Those terms can be shown to converge to 0 in a way similar to the one proposed by those authors, but with conditions adapted to our functional setting. The purpose of Lemma 7 is to show that the conditions for our Theorem are enough to obtain conditions that play the same role as the ones used by Marron and Hardle [21] in their proof.

Although our proofs follow similar ideas as the one given in Rachdi and Vieu [20] and Ait-Saidi et al. [12], some major differences occur. As the parameter H that we cross-validate makes changes in the semimetric, it sounds not natural to assume that the function $\hat{r}_H$ is always regular with respect to the semimetric. Opposite to Rachdi and Vieu [20] and Ait-Saidi et al. [12], we can thus not make use of this condition anymore. This translates into the fact that the Lipschitz parameter β_H might be equal to zero in condition (3.23). As a consequence, the bias of the estimate might not necessarily go to zero, for some choice of the parameters. This translates into the need to use a more precise inferior bound

$$\frac{C}{n\Phi_H} + \frac{n-1}{n}b_H$$

for the $MISE^*(H)$ (Lemma 5), instead of $\frac{C}{n\Phi_H}$ in Rachdi and Vieu [20] and Ait-Saidi et al. [12]. In particular, the proof of the convergence of T_γ in Lemma 12 require a much more careful treatment.

4. Applications and discussions

In this Section, we assess the performances of Bagidis in nonparametric functional prediction using estimator (2.2), as compared with results relying upon the usual semimetrics defined in Subsection 2.2. We also illustrate the efficiency of the cross-validation procedure described in Subsection 3.2 for selecting the weights of the Bagidis semimetric in this context. Simulated and real data examples involving curves with sharp patterns are studied therefore.

We observe on simulated examples that Bagidis shows prediction performances highly superior to competitors as soon as the model involves curves whose significant variations of sharp local patterns have an horizontal component. In case the model involves a sharp pattern whose variation in amplitude is significant but which remains well-aligned across the dataset, the Pca-based semimetric is best, but performances of Bagidis achieve nearly the same order of magnitude when the noise on the curves is not too high. The cross-validated selection of the non-zero weights of the Bagidis semimetric, and hence of the significant ranks in the Buuhwt expansion (2.6) of the curves, proves very efficient on those examples, with very few selections of unsignificant ranks. This also holds when the significant sharp pattern to capture is a secondary pattern and is thus not encoded in the first ranks of the Buuhwt expansion. Such a cross-validated selection allows to further improve the predicting performances of Bagidis as compared with a non-optimized version of Bagidis and with competing semimetrics.

A real data example involving H-NMR serum spectra is then studied. The goal is to discriminate healthy patients from patients suffering from a given illness, according to the composition of their blood serum. Again, Bagidis proves highly efficient for that purpose.

Those analyses are performed using the R software [22], and using the R implementation of estimators (2.2) and (2.3), provided by Ferraty and Vieu [1], slightly adapted for their use together with the Bagidis semimetric.

4.1. A systematic simulated study of the prediction capacities of Bagidis *for datasets of curves having horizontally shifted and/or vertically amplified sharp patterns*

We investigate the potential of using the Bagidis semimetric in regression by studying simulated datasets involving curves having a single significant sharp pattern that is either horizontally shifted across the dataset, or vertically amplified, or both simultaneously. The related responses in those regression problems derive from the amplitude and/or location of that sharp pattern. Those examples allow for diagnosing the ability of Bagidis to deal with different kinds of differences amongst the curves of a dataset. We emphasize the fact that the method does not make use of the prior information of the nature of the variation amongst the curves. One of our goals in this work is precisely for our method to automatically adapt to this nature, through an optimal choice of the balance parameter λ, of the weights w, and of the bandwidth h.

Definition of the simulated models. The models we study are as follows. First, we investigate how Bagidis handles horizontal shifts and vertical amplifications of patterns separately, through the analysis of curves generated from the following models:

- **Model 1: an *up-and-down* horizontally shifted pattern is related to its delay.** The first example involves series of length 21, being zero-valued except for the presence of an *up-and-down* pattern $(10, -10)$ that is horizontally shifted from one series to the next one. Each series is related to the delay at which the *up-and-down* pattern occurs. This is illustrated at Figure 1 *(top, left).*

- **Model 2: an *up-and-down* vertically amplified pattern is related to its height.** The second example involves series of length 21 being zero-valued except for an *up-and-down* pattern located at abscissas $(10, 11)$, that is more or less amplified from one series to the next one, from amplitude 1 to amplitude 20. Each series is associated with the height of the *up-and-down* pattern. This is illustrated at Figure 2 *(top, left).*

We then study a model that combines horizontal shifts and vertical amplifications of sharp patterns:

- **Model 3: an amplified and shifted *up-and-down* pattern is related to a value depending on both its height and delay**. We consider series of length 21, being zero-valued except for the presence of an *up-and-down* pattern $(1, -1)$. That pattern appears after a certain delay and is affected by a certain multiplicative amplification factor, both being randomly generated in $1 \dots 20$. Sample curves generated according to this model are illustrated in Figure 3 (*top, left*). The responses associated with those curves are the sum of the delay and the amplitude.

The last model we consider involves two sharp patterns, the main one being non-informative, the secondary one being the only one whose variation carries significant information:

- **Model 4: an horizontally shifted secondary *up-and-down* pattern is related to its delay.** We consider series of length 21, being zero-valued except for the presence of a main *up-and-down* pattern (-20,20) located at abscissas (10,11), as well as the presence of a secondary *up-and-down* pattern (-10,10) that is horizontally shifted along the series (it is thus possibly combined to the main pattern, for certain delays). Response values are defined as the delay at which this secondary pattern occurs. This is illustrated at Figure 4 (*top, left*).

The simulated series we generate according to those four models are affected by a Gaussian noise with standard deviation σ_X, with σ_X taking its values in (0.25, 0.5, 1, 2, 3) - depending on the simulation, and the responses are affected by a Gaussian noise with standard deviation $\sigma_Y = 1$. The related signal-to-noise ratio are provided in Table 1.

σ_χ:	**0.25**	**0.5**	**1**	**2**	**3**
$\frac{s}{\sigma_\chi}$	4	2	1	0.5	0.3
$\frac{\bar{sd}(\chi)}{\sigma_\chi}$ for **Model 1: shifted patterns**	12.8	6.4	3.2	1.6	1.1
$\frac{\bar{sd}(\chi)}{\sigma_\chi}$ for **Model 2: amplified patterns**	13.2	66.6	3.3	1.6	1.1
$\frac{\bar{sd}(\chi)}{\sigma_\chi}$ for **Model 3: randomly shifted and amplified patterns**	13.2	66.6	3.3	1.6	1.1
$\frac{\bar{sd}(\chi)}{\sigma_\chi}$ for **Model 4: secondary shifted patterns**	28	14	7	3.5	2.3

Table 1: **Signal-to-noise ratio for the simulation study of Subsection 4.1.** $\frac{s}{\sigma_\chi}$ is the ratio of the smallest difference (vertically or horizontally) between the model curves s and the standard deviation of the noise applied to the curves σ_χ. It is the same for all models, as s is always fixed to 1. $\frac{\bar{sd}(\chi)}{\sigma_\chi}$ is the ratio of the standard deviation of the model curves, averaged on a sample set of curves, $\bar{sd}(\chi)$ and the standard deviation of the noise applied to the curves σ_χ. Although common, those values $\frac{\bar{sd}(\chi)}{\sigma_\chi}$ have to be taken with caution in our study, as $\bar{sd}(\chi)$ includes thus variations of the curves that are either significant or non-significant (Model 4) and do not take horizontal shifts into account (Models 1, 3 and 4) .

Description of the analysis. The following test is performed T times, for each model and each value of σ_χ. We generate M noisy pairs $(\chi_i, Y_i)_{i=1\dots M}$ according to the chosen model, each model value of the delay and/or height having the same probability to appear in the dataset. Then, we randomly select n pairs out of those M, and use them as a training set to calibrate the regression model. Using the model for predicting the responses associated with the $M-n$ remaining series and comparing it with their "true" simulated noisy response, we calculate the associated mean square error of prediction (MSE).

The performances obtained using the Bagidis semimetric with estimator (2.2) on those problems are compared with the one we obtain using the functional PCA-based semimetric d_q^{PCA} with various number of principal components, the derivative-based semimetric d_q^{deriv} with various order of derivation (including no derivation) and the *hshift* semimetric d^{hshift}. The use of a vectorial L_2-distance d^{L_2} as a semimetric is also considered. Besides, a *no effect* prediction is provided - i.e. a prediction by the mean of the response values of the training set, which acts as a benchmark for the performances.

Our analysis actually proceeds into two steps.

- **Step 1: Studying the performances of Bagidis in nonparametric functional regression as a function of the balance parameter λ, with a prior, sub-optimal, choice of the weights and a cross-validated-choice of the bandwidth h.** In order to get a first insight into the behaviour of the Bagidis-based regression estimator as a function of λ, only an adaptation of the bandwidth is considered, and Bagidis is used with a prior weight function defined as

$$w_0 = 0;\ w_k = \frac{\log(N+1-k)}{\log(N+1)} \text{ for } k = 1 \dots N-1, \tag{4.1}$$

  as proposed in Timmermans and von Sachs [2]. This allows to associate a large weight to the comparison of features encoded at the first rank of the hierarchy, and a decreasing weight to the smaller features at the end of the hierarchy, which is empirically what we expect for relatively sparse noisy curves. Values from λ from 0 to 1 with a step of 0.1 are tested. The bandwidth h is optimized through a set of values defined as a sequence of 20 equispaced values from the quantile 0.05 to the quantile 0.5 of the observed distances between the curves, which is the default behaviour of the R function provided by Ferraty and Vieu [1]. With this first step, we investigate thus how those "sub-optimal" versions of the Bagidis semimetric behave compared with "classical" semimetrics, depending on the value of λ, which gives a first idea of the potential of our method. It also allows to identify the best competitors of Bagidis in each setting.
  This analysis is performed with $T = 100$ and $M = 60$, $n = 45$ for **Model 1**, **Model 2** and **Model 4**. Because of the more important complexity of **Model 3**, a larger training set ($M = 180$, $n = 160$) is to be used if we wish to achieve an explained percentage of the no-effect MSE that is about 90 for $\sigma_\chi = 0.25$, as for the other models. Smaller size of the training sets (e.g. $M = 60$, $n = 45$ as for the other models) leads to the same relative performances of the semimetrics as those presented here-above, but with a systematically higher MSE.

- **Step 2: Optimizing the weights, the balance parameter and the bandwidth in nonparametric regression using a *leave-one-out* cross-validation procedure.** As theoretically supported by Theorem 1, having a sparse

weight function will significantly improve the rate of convergence the estimator, and hence the performance of Bagidis compared to competitors. With this second step, we illustrate how we can further improve the predicting performances of Bagidis on the above-defined models by optimizing the weights w, the balance parameter λ as well as the smoothing parameter h. Practically, this is done using a cross-validation procedure and a *leave-one-out* mean square error criterion, as suggested by Theorem 2. The set of parameters $H = (w, \lambda, h)$ over which we optimize the *leave-one-out* MSE criterion is defined as follows: values of λ are tested from 0 to 1 with a step of 0.1; h is allowed for taking its values in a sequence of 20 equispaced values from the quantile 0.05 to the quantile 0.5 of the observed distances between the curves; values of the weights w_k are only allowed for being 1 or 0 - i.e each rank k can be *activated* or *unactivated* in the semimetric. However, not all the possible combinations of weights are actually tested, as a *forward selection approach* is favoured in order to reduce the optimization time [23]. This means we first compute the *leave-one-out* MSE on the training set for any possible combination of λ and h, for each of the possible single activated weights. The rank k^* whose activation leads to the smaller *leave-one-out* MSE is selected. If the best so-obtained *leave-one-out* MSE is strictly smaller than a *"no effect" leave-one-out* MSE, the weight w_{k^*} is set to 1. We then do the same for selecting another activated rank in the weight function. This procedure is iterated while the resulting *leave-one-out* MSE decreases. The best set of parameters $H^{Opt} = (w^{k^{Opt}}, \lambda^{Opt}, h^{Opt})$ is thus selected as the minimizer of our criterion amongst the tested sets of parameters. Note here that such a *forward selection procedure* is a very common approach for selecting among a large set of parameters (see Guyon and Elisseeff [23], for instance). Once the optimal set of parameters H^{Opt} is selected, the mean square errors of prediction is evaluated on the validation set, using the optimal so-parametrized predictor.
This analysis is performed with $T = 30$ and M and n defined as for Step 1.

Presentation and discussion of the results. Resulting distributions of the MSE obtained at **Step 1** for each of the tested semimetrics are presented in Figures 1 to 4, for each model and each value of σ_χ. When interpreting those results, we have to keep in mind that the Bagidis results are sub-optimal here, as the parametrization of the semimetric is not optimized in this first analysis. Summary results extracted from those graphs about the MSE obtained using Bagidis and using its best competitor semimetric are shown in Table 2. Resulting distribution of the MSE obtained at **Step 2,** when the parametrization of Bagidis is fully optimized using a cross-validation procedure, are then summarized in Table 3. MSE distributions obtained with the optimized Bagidis semimetric for each model are confronted with the best competitor MSE distributions on the same model obtained by the analysis of **Step 1**, this best competitor distribution being either Bagidis with unoptimized weights and with the best expected value of λ, or another semimetric -typically the Pca.

Analysis of Model 1: Capturing the location of an horizontally shifted sharp pattern. A look at Figure 1 and Table 2 (*row 1*) tells us that, as expected, the Bagidis semimetric leads to excellent performances compared to all competitors for dealing with the **Model 1: shifted patterns,** as soon as $\lambda > 0$ - i.e. as soon as the differences between the breakpoints are taken into account in equation (2.8). We observe that the sensitivity to the choice of the parameter λ increases with σ_χ, and $\lambda = 1$ is most systematically favored in this case. This is not surprising as we know, by construction of the model, that only the breakpoints (solely captured with $\lambda = 1$) carry significant information. Only in the least noisy case $\sigma_\chi = 0.25$, d_0^{deriv} performs better than Bagidis. In that case, the bandwidth is actually selected so small that quasi-only similarly aligned curves define the predictor. On the contrary, Bagidis is able to detect the closeness of neighbour shifted curves for building the predictor. This leads to the fact that Bagidis-based model explains a significant part of the no-effect MSE, up to $\sigma_\chi = 3$. The additionally explained percentage of the no-effect MSE that is explained by Bagidis as compared with the best competitor reaches 39.75 when $\sigma_\chi = 3$. At that level of noise, the percentage of explanation offered by the sub-optimal Bagidis-based model is more than twice better than the one achieved by its best competitor. Further non illustrated studies show that this advantage of the Bagidis semimetric remains up to a noise level $\sigma_\chi = 6$, where no model is able to do significantly better than the no-effect MSE. Moreover, as could have been expected, increasing σ_Y increases the MSE whatever the semimetric, but does not affect their relative performances.
Optimizing the parameters of the Bagidis allows to further improve our prediction performances. As can be seen from Table 3 *(row 1)*, a significant percentage (> 80%) of the no-effect MSE is now explained even in the most noisy illustrated setting with $\sigma_\chi = 3$. We observe a gain of 7 to 10 % of explanation of Y, as compared with the Bagidis

Model 1: Shifted Patterns, $\sigma_Y = 1$

	No-effect MSE	**BAGIDIS**			**Competitor**			**Comparison of performances**	
σ_X		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	36.09	**0.2**	3.34	90.74	**Deriv-0**	2.14	94.06	-3.32	0.96
0.5	35.54	**0.2**	3.45	90.29	**PCA-11**	5.23	85.28	5.01	1.06
1	34.80	**0.5**	4.08	88.28	**PCA-11**	7.46	78.56	9.72	1.12
2	34.77	**0.8**	4.62	86.71	**PCA-21**	13.92	59.96	26.75	1.45
3	35.59	**1**	8.83	75.15	**PCA-21**	22.99	35.40	39.75	2.12

Model 2: Amplified Patterns, $\sigma_Y = 1$

	No-effect MSE	**BAGIDIS**			**Competitor**			**Comparison of performances**	
σ_X		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	36.96	**0**	1.46	96.04	**PCA-11**	1.41	96.18	-0.14	1.00
0.5	35.20	**0**	1.56	95.57	**PCA-7**	1.43	95.94	-0.37	1.00
1	36.26	**0**	2.77	92.36	**PCA-5**	1.98	94.54	-2.18	0.98
2	37.17	**0**	7.65	79.42	**PCA-1**	3.83	89.97	-10.55	0.88
3	35.86	**0.1**	11.18	68.82	**PCA-1**	5.57	84.47	-15.65	0.81

Model 3:Random Amplification and Location of the Patterns, $\sigma_Y = 1$

	No-effect MSE	**BAGIDIS**			**Competitor**			**Comparison of performances**	
σ_X		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	82.16	**0.2**	4.91	94.02	**PCA-12**	12.30	85.03	8.99	1.11
0.5	82.57	**0.2**	6.98	91.5	**PCA-13**	14.59	82.33	9.17	1.11
1	70.55	**0.3**	11.35	83.91	**PCA-19**	19.38	72.53	11.38	1.16
2	70.14	**0.3**	18.55	73.55	**PCA-15**	22.40	68.06	5.49	1.08
3	67.29	**0.3**	26.93	59.98	**hshift**	36.11	46.34	13.64	1.29

Model 4:Second Order Shifted Patterns, $\sigma_Y = 1$

	No-effect MSE	**BAGIDIS**			**Competitor**			**Comparison of performances**	
σ_X		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	35.65	**0.6**	7.04	80.25	**Deriv-0**	2.08	94.16	-13.91	0.85
0.5	35.13	**0.6**	7.118	79.56	**Deriv-0**	6.94	80.24	-0.68	0.99
1	34.29	**0.6**	6.91	79.85	**PCA-11**	8.86	74.16	5.69	1.08
2	35.54	**0.9**	8.55	79.94	**PCA-21**	12.49	64.85	11.09	1.17
3	36.04	**0.9**	12.36	65.70	**PCA-21**	23.63	34.43	31.27	1.91

Table 2: **Summary analysis of the examples of Step 1 analysis in Subsection 4.1.** The *percentage of no-effect explained by the model* is calculated as $100(1 - \frac{\text{Mean MSE}}{\text{Mean no-effect MSE}})$. The difference between this percentage for the BAGIDIS semimetric and for its best competitor is given as the *Additionaly explained percentage of no-effect MSE*. This percentage is thus negative in case the competitor semimetric performs better than BAGIDIS. The last column of the table is the ratio of the percentages of no-effect MSE explained using BAGIDIS and using its best competitor. BAGIDIS is superior as soon as this ratio exceeds 1.

Model 1: Shifted Patterns, $\sigma_Y = 1$

	Optimized Bagidis				**Competitor**			**Comparison of performances**	
σ_X	Mean number of activated weights	Most activated weights	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	**1.4**	**1 and 2**	1.48	95.11	**Bagidis-1**	4.16	85.99	9.13	1.11
0.5	**1.2**	**1 and 2**	1.62	94.92	**Bagidis-1**	4.48	85.73	9.19	1.11
1	**1.5**	**1 and 2**	1.46	95.34	**Bagidis-1**	4.01	86.37	8.97	1.10
2	**1.5**	**1 and 2**	2.48	92.76	**Bagidis-1**	4.89	85.16	7.60	1.09
3	**2.3**	**1 and 2**	5.77	80.86	**Bagidis-1**	8.16	73.75	7.11	1.10

Model 2: Amplified Patterns, $\sigma_Y = 1$

	Optimized Bagidis				**Competitor**			**Comparison of performances**	
σ_X	Mean number of activated weights	Most activated weights	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	**4.6**	**2 and 3**	1.47	95.44	**pca-1**	1.49	95.36	0.07	1.00
0.5	**4.1**	**2 and 3**	1.56	95.08	**pca-1**	1.46	95.41	-0.33	1.00
1	**4.1**	**2 and 3**	2.57	91.70	**pca-1**	1.92	93.65	-1.95	0.98
2	**4.3**	**2 and 3**	5.56	80.34	**pca-1**	3.65	87.11	-6.77	0.92
3	**3.2**	**2 and 3**	9.93	68.40	**pca-1**	5.82	81.68	-13.28	0.84

Model 3: Randomly Shifted and Amplified Patterns, $\sigma_Y = 1$

	Optimized Bagidis				**Competitor**			**Comparison of performances**	
σ_X	Mean number of activated weights	Most activated weights	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	**2.0**	**2 and 1**	2.37	96.19	**Bagidis-0.25**	5.94	90.58	5.61	1.06
0.5	**2.6**	**2 and 1**	5.72	91.58	**Bagidis-0.25**	8.91	87.06	4.52	1.05
1	**2.4**	**2 and 1**	11.41	80.67	**Bagidis-0.25**	11.95	80.61	0.06	1.00
2	**2.7**	**2 and 1**	13.83	75.35	**Bagidis-0.25**	16.92	69.88	5.46	1.08
3	**2.6**	**2 and 1**	24.06	60.75	**Bagidis-0.25**	31.07	49.87	10.88	1.22

Model 4: Second order Shifted Patterns, $\sigma_Y = 1$

	Optimized Bagidis				**Competitor**			**Comparison of performances**	
σ_X	Mean number of activated weights	Most activated weights	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionaly explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	**1.5**	**4**	2.60	90.27	**Bagidis-1**	11.01	66.68	25.45	1.38
					Deriv-0	2.21	93.31	-1.18	0.99
0.5	**1.5**	**4 and 5**	3.18	92.13	**Bagidis-1**	10.51	67.85	22.42	1.33
					Deriv-0	7.95	75.68	14.52	1.19
1	**1.7**	**4 and 5**	3.75	88.16	**Bagidis-1**	9.24	70.83	17.33	1.24
					Deriv-0	17.87	43.59	44.57	2.02
2	**2.7**	**4 and 5**	6.83	78.57	**Bagidis-1**	9.48	70.25	8.32	1.12
					Deriv-0	33.93	-6.46	78.57*	*
3	**3.0**	**4 and 5**	13.34	58.30	**Bagidis-1**	13.28	58.49	-0.19	1.00
					Deriv-0	35.34	-10.47	58.30*	*

Table 3: **Summary analysis of the examples of Step 2 analysis in Subsection 4.1.** The *percentage of no-effect explained by the model* is calculated as $100(1 - \frac{\text{Mean MSE}}{\text{Mean no-effect MSE}})$. The difference between this percentage for the Bagidis semimetric and for its best competitor is given as the *Additionaly explained percentage of no-effect MSE*. This percentage is thus negative in case the competitor semimetric performs better than Bagidis. The last column of the table is the ratio of the percentages of no-effect MSE explained using Bagidis and using its best competitor. Bagidis is superior as soon as this ratio exceeds 1. Values denoted by $*$ in the *Comparison of performances* for **Model 4** appear when the *percentage of no effect MSE explained using the competitor semimetric* is observed negative, indicating the the no-effect prediction is better - i.e. that the semimetric does not capture anything about the significant variations in the curves. In this case, the *additionaly explained percentage of no-effect MSE* is the actual *percentage of no-effect MSE explained by* Bagidis, and the *Ratio of explained percentages* is not computed.

semimetric used with its prior weight function and the optimal value $\lambda = 1$, that was noticed superior to all competitors at **Step 1** for $\sigma_\chi > 0.25$. Moreover, the Bagidis semimetric is now competitive as compared with d_0^{deriv} even in this situation of small noise $\sigma_\chi = 0.25$. The number of selected weights remains small in all examples. Ranks 1 and/or 2, the ones that carry significant, redundant, information about the shifted pattern, are selected most of the time. On average less than 2 non-zero weights are selected for $\sigma_\chi < 3$, as expected. In conclusion, the Bagidis semimetric is clearly better than competitor for capturing the shift of a sharp pattern, and its performances are further improved by optimizing its parameters.

Analysis of Model 2: Capturing the amplification of well-aligned sharp pattern. Not surprisingly, the **Model 2: amplified patterns** is best tackled by the functional PCA-based semimetric, as can be seen from Figure 2 and Table 2 *(row 2)*. Nevertheless, it is interesting to note that our sub-optimal Bagidis semimetric performs quite well too for $\sigma_\chi < 2$, with $\lambda = 0$ - i.e where only amplitude differences are taken into account. In those cases less than 2.18 percent of the no-effect MSE is additionally explained when using a Pca-based semimetric. The advantage of the Pca-based semimetric becomes really significant afterwards. However, it is interesting to note that Bagidis still significantly detects an effect of the amplification of the sharp pattern, up to a noise level $\sigma_\chi = 5$, what the derivative-based family of semimetric cannot do. Again, increasing σ_Y increases the MSE whatever the semimetric, but does not affect their relative performances.
As indicated by Table 3 *(row 2)*, optimizing the parameters of the Bagidis does not significantly allow to further improve our prediction performances. The number of selected weights is higher than 4 most of the time, indicating a certain number of spurious rank selection. In summary, the PCA-based semimetric is best in case of well aligned sharp patterns variations, but the Bagidis semimetric remains competitive if the noise on the curves is not too important

Analysis of Model 3: Capturing the height and delay of a randomly amplified and shifted sharp pattern. From Figure 3 and row 3 of Table 2, it is clear that even the sub-optimal form of Bagidis performs very well on **Model 3: randomly amplified and shifted pattern**, and significantly better than competitors in every illustrated case. For upper values of σ_χ ($\sigma_\chi > 4$), no model is able to do significantly better than the no effect MSE. As expected, an intermediate value of λ seems to be the best choice as both differences in the localizations and in the amplitudes are informative for the prediction. $\lambda = 0.2$ or $\lambda = 0.3$ seem to be favoured. This can be interpreted because 0.25 is the ratio of the squared range for the breakpoints over the square range of the details (the square has to be taken because we use $p = 2$ in equation (2.8)). Again, further non illustrated studies show that increasing σ_Y increases the MSE whatever the semimetric, but does not affect their relative performances.
Optimizing the parameters of the Bagidis semimetric slightly improves our prediction performances, as compared with Bagidis semimetric used with its prior weight function and $\lambda = 0.25$. This smaller amelioration, as compared with the improvement achieved for Model 1, might be related to the fact that a larger training set is used here so that the noise that affects the prediction due to the presence of unsignificant ranks in the prior weights function is better averaged to zero. The difference is more marked for a high level of noise $\sigma_\chi = 3$, as seen from for Table 3 *(row 3)*. The number of selected weights remains small whatever the noise level. Ranks 1 and 2, the ones that carry significant, redundant, information about the shifted and amplified pattern, are essentially selected, and few spurious selections occur. In average about 2 non-zero weights are selected, as expected. This third example shows that the Bagidis semimetric is clearly superior to competitors and optimizing its parametrization might help to further improve its performances.

Analysis of Model 4: Capturing a secondary shifted sharp pattern. The definition of **Model 4: secondary shifted pattern** implies that the first ranks of the Bagidis semimetric should encode non significant information as they essentially compare the largest uninformative main pattern, while ranks 4 and 5 should be relevant for predicting the response. This last example aims at checking that this kind of behaviour is correctly handled by our optimization procedure. Results obtained at **Step 1** for this model can be found in Figure 4 and Table 2 (*row 4*). As for **Model 1: shifted pattern**, it shows that the sub-optimal non-optimized Bagidis semimetric has very good performances as compared with competitors, as soon as $\lambda > 0$, with the best λ significantly tending to 1 as $\sigma_\chi$increases. Again, and for the same reason as discussed for **Model 1**, d_0^{deriv} is best in case of a low level of noise $\sigma_\chi = 0.25$ and has equivalent performances for $\sigma_\chi = 0.5$. For higher noise levels $\sigma_\chi > 0.5$, the Bagidis semimetric, even with its sub-optimal parametrization, performs clearly better that competitors. Further non illustrated studies show that this advantage of

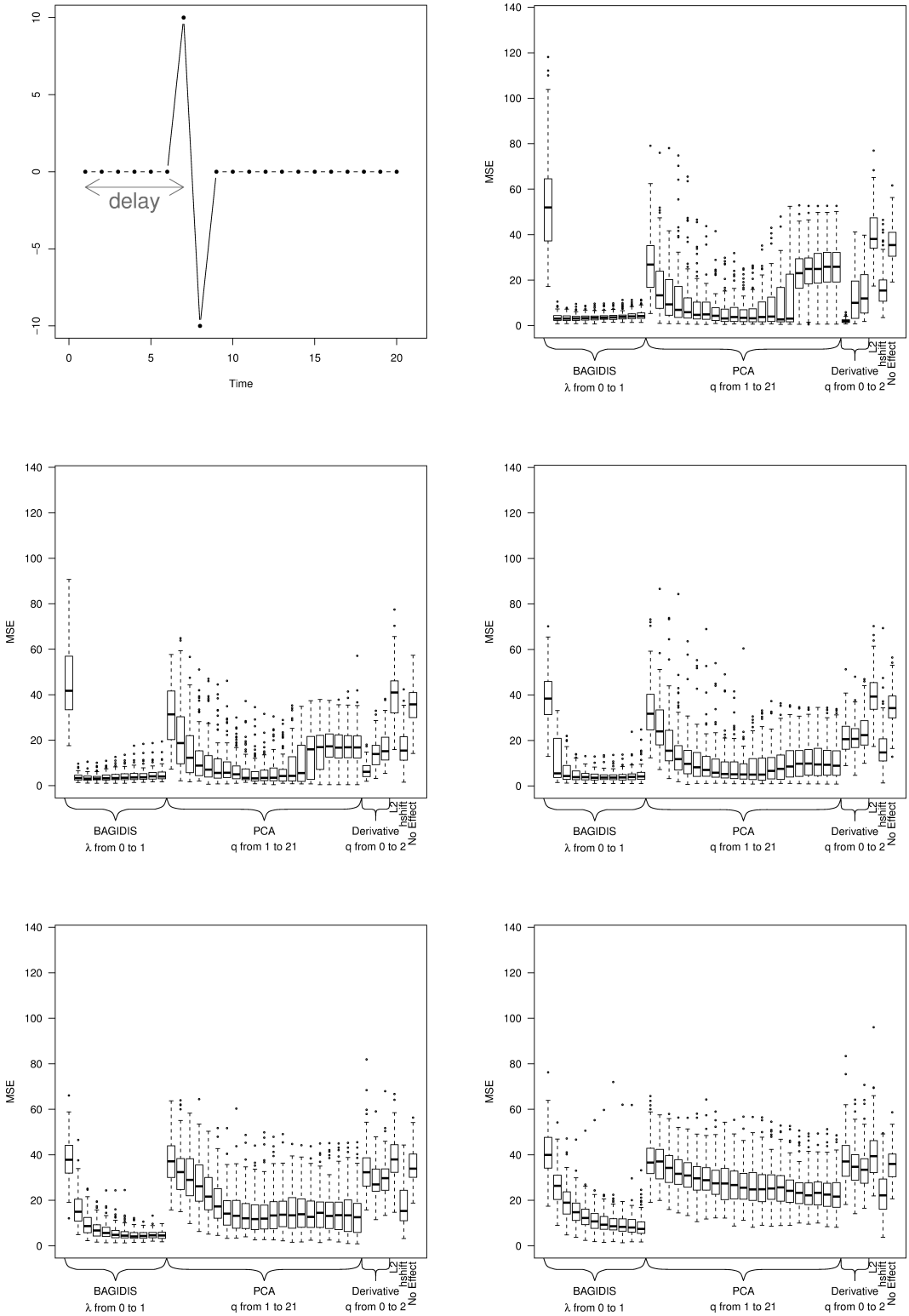


Figure 1: **Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to Model 1: shifted patterns.** *Top, left:* Schematic illustration of **Model 1: shifted patterns.** *Top, right to bottom right:* Boxplot representations of the observed MSE distributions, for data generated according to **Model 1,** with $\sigma_Y = 1$ and $\sigma_\chi = 0.25$, 0.5, 1, 2, and 3 respectively (from left to right and top to bottom).

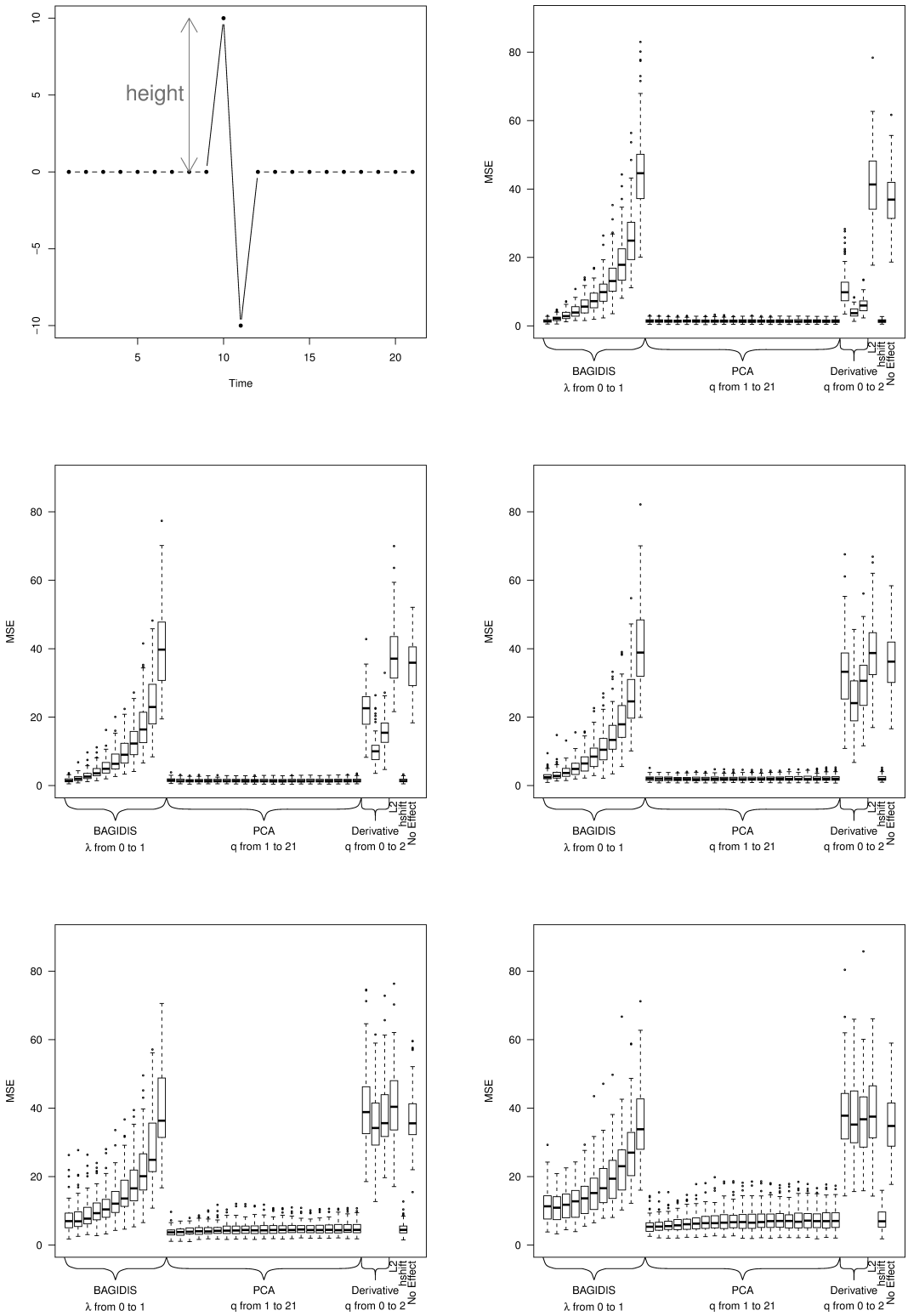


Figure 2: **Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to Model 2: amplified patterns.** *Top, left:* Schematic illustration of **Model 2: amplified patterns.** *Top, right to bottom right:* Boxplot representations of the observed MSE distributions, for data generated according to **Model 2,** with $\sigma_Y = 1$ and $\sigma_\chi = 0.25,\ 0.5,\ 1,\ 2,$ and 3 respectively (from left to right and top to bottom).

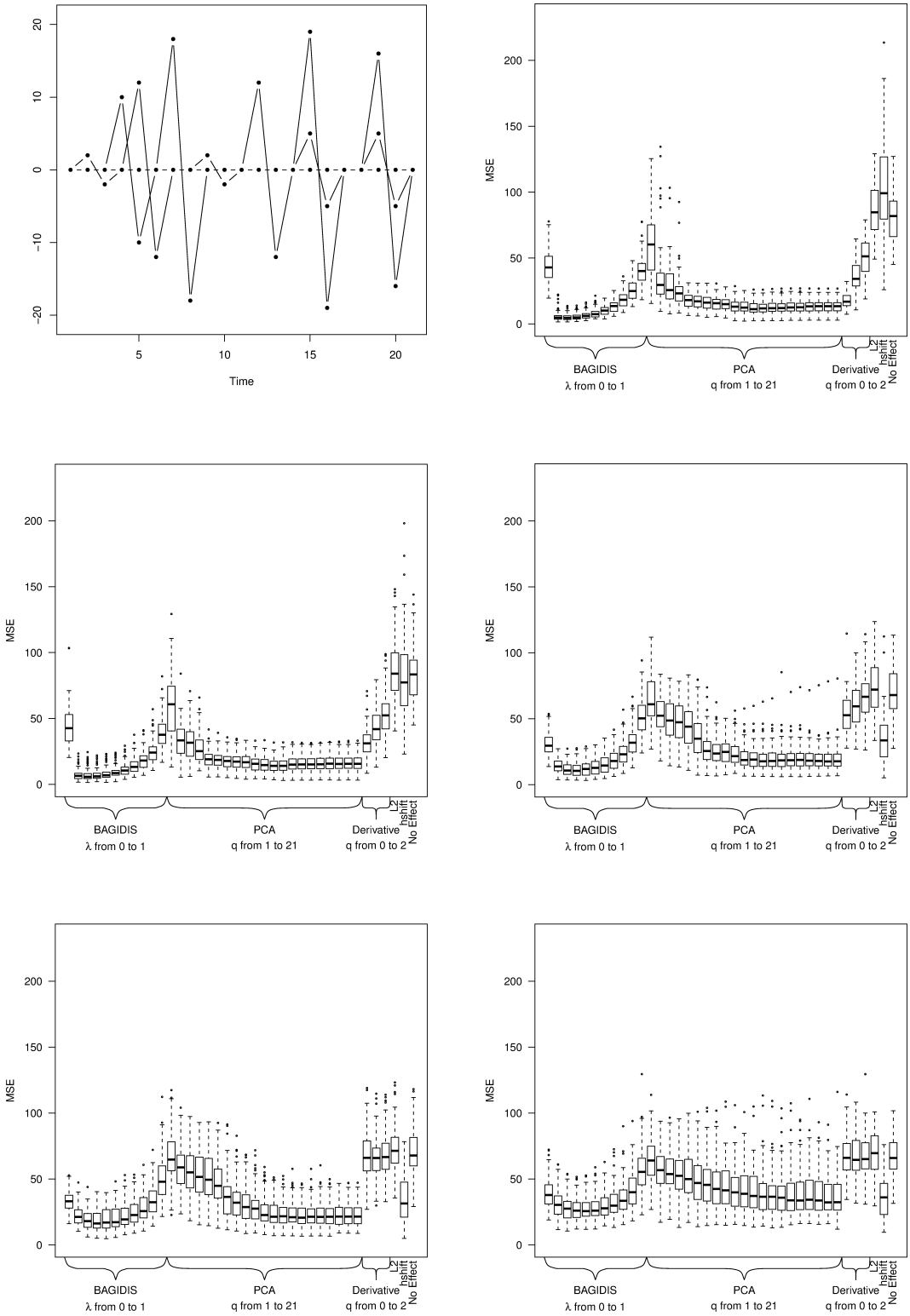


Figure 3: **Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to Model 3: randomly shifted and amplified patterns.** *Top, left:* Schematic illustration of sample curves generated from **Model 3: randomly shifted and amplified patterns.** *Top, right to bottom right:* Boxplot representations of the observed MSE distributions, for data generated according to **Model 3,** with $\sigma_Y = 1$ and $\sigma_\chi = 0.25,\ 0.5,\ 1,\ 2,$ and 3 respectively (from left to right and top to bottom).

the Bagidis semimetric remains up to a noise level $\sigma_\chi = 6$, where no model is able to do significantly better than the no-effect MSE.
Table 3 (*row 4*) shows that the cross-validated selection of the weights leads to a significant improvement of the performances of Bagidis, at least for $\sigma_\chi < 3$. This is probably due to the fact that the prior weight function was not really adapted to this example, as it gave a higher weights to the first unsignificant ranks. On the contrary, the optimization procedure efficiently selects ranks 4 and/or 5, the ones that carry significant information on the secondary shifted pattern, with very few spurious selection so that the number of selected weights remain small in every case. The advantage of d_0^{deriv} for small values of σ_χ disappears as soon as the selection procedure of the weights takes place, and results achieved with the optimized Bagidis semimetric are equivalent ($\sigma_\chi = 0.25$) or highly better ($\sigma_\chi > 0.25$) than with d_0^{deriv} .
In summary, the Bagidis semimetric is highly efficient for predicting from curves with a secondary shifted sharp pattern, and highly benefits from a cross-validated selection of the activated weights in this case. The significant ranks are directly selected with very few spurious selection.

Conclusion of the simulated study. Those simulated examples illustrates the potential of using Bagidis for nonparametric prediction of curves. Even used with its non-optimized prior weight function, performances of Bagidis shows superior performances compared to classical semimetric as soon as variations of sharp local patterns in curves have an horizontal component. Those performances may be further improved by a cross-validated selection of the parameters of Bagidis. The mean number of activated weights is then always small. This means we reach quite good rates of convergence in these examples. Moreover, the performances of Bagidis are equivalent to the ones of its best competitor, the Pca-based semimetric, in case no horizontal variation of the significant pattern occur, provided that the noise is not too high. It remains "acceptable" for higher noises, meaning that a prediction is still possible (which is not the case for the derivative-based semimetric for instance). This means that Bagidis could be used quite confidently on datasets with sharp patterns whose kind of variation might not be known in advance. A specificity of Bagidis is indeed that the semimetric can adapt itself to the kind of variation to detect in the dataset, through the optimization of the balance parameter λ.

4.2. Analysis of a real spectrometric dataset

This last Subsection presents a real data example of prediction from spectrometric curves (de Tullio, Frédérich and Lambert, Université de Liège). We consider 193 H-NMR serum spectra of length 600, as illustrated in Figure 5, 94 of which corresponding to patients suffering from a specific illness, the other ones corresponding to healthy patients. We aim at predicting from the spectrum if a patient is healthy or not. A training set of 150 spectra is randomly selected and a functional nonparametric discrimination model is adjusted, with various semimetrics. In each case, the number of misclassification observed on the remaining 43 spectra is recorded.

In order to avoid a confusion of the features in such long series, we make use of the Bagidis semimetric together with a sliding window of length 30 (as suggested in[2]). This allows for comparing the variations of one or few given peak(s) at a time. A specific R function for using estimator (2.3) in a discrimination setting has been provided by Ferraty and Vieu [1]. However, it makes use of a slightly different version of the non parametric estimator (2.3): a local bandwidth is used, which is defined through a number of nearest neighbour that have to be included in the support of the kernel function. Consequently, our good properties stated in Section 3 are not strictly proved in this case of a k-NN based estimator. However, we believe that the good convergence properties of the kernel-based *leave-one-out* MSE minimizer might be extended to a K-nn-based *leave-one-out* MSE minimizer, by generalizing our proofs to this case using similar arguments as those found in Burba et al. [24].

Our test for the prediction of the health status from the spectra is repeated 80 times, with different randomly selected training sets, using the sub-optimal Bagidis semimetric with its prior weight function, with $\lambda = 0.5$ and with a cross-validated bandwidth h, and with the competitor semimetrics identified in Subsection 2.2 and a cross-validated bandwidth. Results are summarized in Table 4.2, for Bagidis and its best competitor, being a PCA-based semimetric with at least 6 components. We observe that the non-optimized Bagidis obtains *no error* 10% more often than the PCA-based semimetric. Afterwards, we optimize the weights and the λ parameter of the Bagidis semimetric using a cross-validation procedure within the training set, and the resulting model is tested on the remaining 43 series. This test is repeated 18 times on different randomly selected training sets, and no prediction error occurs. At each repetition,

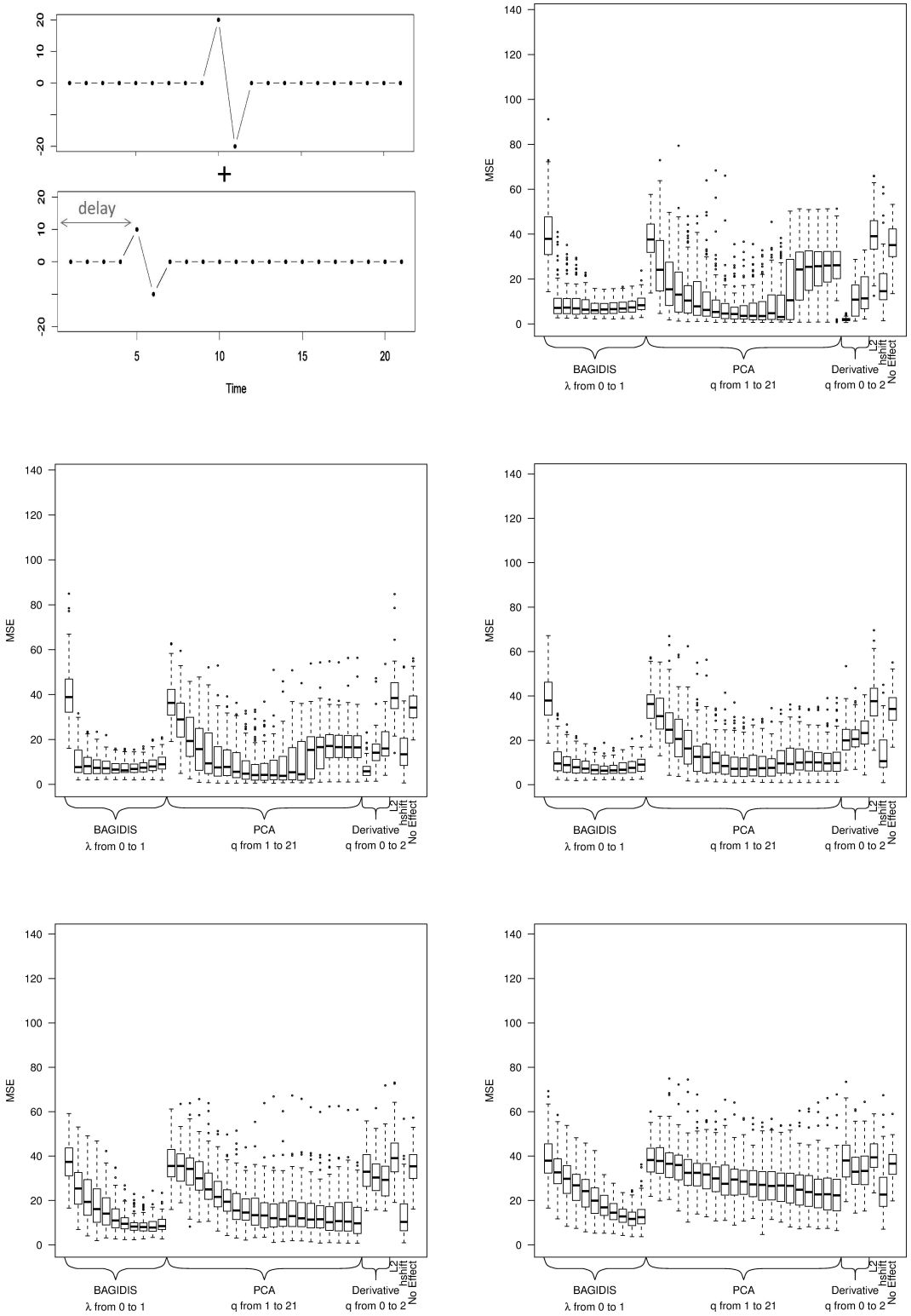


Figure 4: **Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to Model 4: second order shifted patterns.** *Top, left:* Schematic illustration of **Model 4: second order shifted patterns.***Top, right to bottom right:* Boxplot representations of the observed MSE distributions, for data generated according to **Model 4,** with $\sigma_Y = 1$ and $\sigma_\chi = 0.25,\ 0.5,\ 1,\ 2,$ and 3 respectively (from left to right and top to bottom).

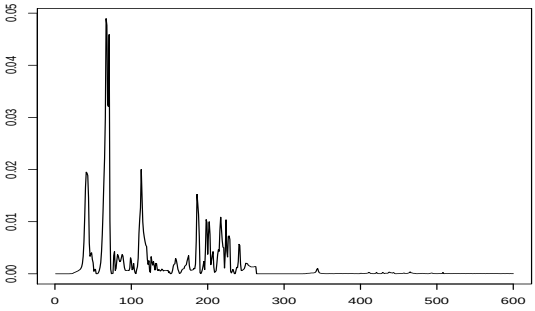


Figure 5: **An H-NMR serum spectra for a ill patient.**

	Occurrences of 0 error out of 43 predictions	Occurrences of 1 error out of 43 predictions
Pca-based semimetric	40 times out of 80	40 times out of 80
with $q \geq 6$	**50%**	**50%**
Non-optimized Bagidis semimetric	48 times out of 80	32 times out of 80
with prior weights and $\lambda = 0.5$	**60%**	**40%**
Optimized Bagidis semimetric	18 times out of 18	0 times out of 18
(1 non zero weight is selected)	**100%**	**0%**

Table 4: **Summary results for the prediction of the health status from the spectra.** Training sets of 150 curves are randomly selected. Predictions are obtained for the 43 remaining spectra and compared with the true health status of those 43 patients. The number of prediction errors is computed. The process is repeated several times for different randomly selected training sets.

only 1 non-zero weight is selected. We observe no prediction error in every case, indicating a risk of misclassification that is probably smaller than 0.05. This indicates a very good capacity of discriminating the serum spectra from ill and healthy patients.

Conclusion

The key idea of this paper is to combine the nonparametric functional framework provided by Ferraty and Vieu [1] with the highly adaptive Bagidis semimetric [2], from predicting from curves with sharp patterns. This association proves highly pertinent. Applications on simulated data have shown in a systematic way the ability of Bagidis to take into account both horizontal and vertical variations of the patterns, as well as its flexibility in the use of this information. Predictions using Bagidis appear to be clearly better than predictions using competing semimetrics, as soon as the variations of the significant sharp patterns have an horizontal component. Those performances concerns both high and relatively low signal-to-noise ratio, which makes the method really attractive. The method also proves really powerful for prediction based on H-NMR spectra, issued from biomedical research.

A theoretical support for those very good observed performances has also been provided in this paper. It was shown that a really competitive rate of convergence of the prediction estimator can be achieved, provided that the multidimensional parametrization of the Bagidis semimetric is sparse enough. It is also shown that this multidimensional parametrization can be chosen using a cross-validation procedure, with a mean-square-error minimization criterion. This method is proved to be asymptotically optimal, and is shown highly efficient on the proposed data analyses examples. The related theoretical results also support for a cross-validated choice of the multidimensional parametrization of others semimetrics, which opens a large scope of perspectives when using projection-based semimetrics in non parametric functional prediction, for instance.

Given all those elements, we think that the Bagidis semimetric really worth to have a place amongst the semimetrics used in nonparametric functional data analysis. Its automatic adaptivity to the nature of the variations of the

patterns in the curves, its ability to deal with horizontal shifts and its capacity to detect the signal in even quite noisy data make it a competitive tool for predicting from curves with sharp patterns.

Acknowledgements. The H-NMR database used in Subsection 4.2 was collected and preprocessed for a study lead by P. de Tullio, M. Frédérich and V. Lambert (Université de Liège). Their agreement for us to use this database is gratefully acknowledged. The name of the concerned illness remains temporarily confidential. We had useful discussions on this project with Frédéric Ferraty and Philippe Vieu, in particular during the Second International Workshop on Functional and Operatorial Statistics. Their comments and encouragements are gratefully acknowledged. The participation of Catherine Timmermans to this workshop has been partially supported by the Fond National de la Recherche Scientifique (Belgium). Part of this work was completed during a stay of Catherine Timmermans at Université d'Orléans-CNRS, that was funded by the Fond National de la Recherche Scientifique (Belgium). Part of this work was completed during a stay of Laurent Delsol at Université catholique de Louvain, that was funded by the Institut de Statistique, Biostatistique and sciences Actuarielles (Université catholique de Louvain). Financial support from the IAP research network grant P 06/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

AppendixA. Proofs

In what follows, we denote by C, C' positive constants, whose value might change from one line to another.

AppendixA.1. Proof of Theorem 1

Proof of Theorem 1 is a direct consequence of Lemma 3 and Lemma 4 above.

Lemma 3. *Assume conditions* (3.1), (3.5), (3.6) *and* (3.7) *to be satisfied. Then, the random variable* $\boldsymbol{\chi}$ *defined by* $(\boldsymbol{b}, \boldsymbol{d}) \in \mathbb{N}_{[0;N-1]} \times \mathbb{R}^N$ *is fractal of order K with respect to the* Bagidis *semimetric at point* $\chi \equiv (b, d)$*. This means that the small ball probability function* $\phi_{d^B,\chi}(.) = P(\boldsymbol{\chi} \in B_{d^B}(\chi, .))$ *of* $\boldsymbol{\chi}$ *about* χ *is such that there exists a positive constant C such that*

$$\phi_{d^B,\chi}(h) \sim C\, h^K, \quad \text{when } h \text{ tends to } 0.$$

Proof. We have

$$\begin{aligned} \phi_{d^B,\chi}(h) &= P(d^B(\boldsymbol{\chi},\chi) \leq h) \\ &= P(d^B(\boldsymbol{\chi},\chi) \leq h \cap \forall k \in \mathcal{K},\ \boldsymbol{b}^k = b^k) + P(d^B(\boldsymbol{\chi},\chi) \leq h \cap \exists k \in \mathcal{K},\ \boldsymbol{b}^k \neq b^k) \end{aligned} \tag{A.1}$$

If $\exists k \in \mathcal{K},\ \boldsymbol{b}^k \neq b^k$, it implies that $\exists\, k \in \mathcal{K}$ such that $|\boldsymbol{b}^k - b^k| \geq 1$, where 1 is the step of the grid $\mathbb{N}_{[0;N-1]}$ on which the curve is observed. In such a case, we have

$$\begin{aligned} d^B(\boldsymbol{\chi},\chi) &= \sum_{k\in\mathcal{K}} w_k \left(\lambda \left|b^k - \boldsymbol{b}^k\right|^p + \underbrace{(1-\lambda)\left|d^k - \boldsymbol{d}^k\right|^p}_{\geq 0} \right)^{1/p} \\ &\geq \sum_{k\in\mathcal{K}} w_k \lambda^{1/p} \left|b^k - \boldsymbol{b}^k\right| \\ &\geq \lambda^{1/p} \min_{k\in\mathcal{K}}(w_k). \end{aligned}$$

Consequently, the right hand side term of equation (A.1) has probability zero for $h < \lambda^{1/p}\min_{k\in\mathcal{K}}(w_k)$. Thus, when h tends to zero, we have

$$\begin{aligned} \phi_{d,\chi}(h) &= P(\underbrace{\sum_{\forall k\in\mathcal{K}} w_k (1-\lambda)^{1/p} |\boldsymbol{d}^k - d^k|}_{\equiv \|\boldsymbol{d}-d\|_{\lambda,w}} \leq h \quad \cap \forall k \in \mathcal{K},\ \boldsymbol{b}^k = b^k) \\ &= \underbrace{P(\forall k \in \mathcal{K},\ \boldsymbol{b}_k = b_k)}_{>0}.P(\| \boldsymbol{d} - d \|_{\lambda,w} \leq h \quad | \forall k \in \mathcal{K},\ \boldsymbol{b}_k = b_k) \end{aligned} \tag{A.2}$$

By definition, the last term of this expression is

$$P(\|\boldsymbol{d}-d\|_{\lambda,w}\leq h|\forall k\in\mathcal{K},\ \boldsymbol{b}^k=b^k)=\int_{B(d,h)_{\lambda,w}} f_{|b}(b,s)\,ds.$$

Then, by the continuity condition of the conditional density, we have $\forall\epsilon>0$, $\forall h<\min(\delta_\epsilon,\min(w_k))$,

$$\begin{aligned}
&\left\|\int_{B(d,h)_{\lambda,w}} f^{\mathcal{K}}_{d|b}(s)\,ds-\int_{B(d,h)_{\lambda,w}} f^{\mathcal{K}}_{d|b}(d)\,ds\right\| \\
=\ &\left\|\int_{B(d,h)_{\lambda,w}}\left\{f^{\mathcal{K}}_{d|b}(s)-f^{\mathcal{K}}_{d|b}(d)\right\}ds\right\| \\
\leq\ &\int_{B(d,h)_{\lambda,w}}\sup_{s\ t.q.\|s-d\|_{\lambda,w}\leq\delta_\epsilon}\left|f^{\mathcal{K}}_{d|b}(s)-f^{\mathcal{K}}_{d|b}(d)\right|ds \\
\leq\ &\int_{B(d,h)_{\lambda,w}}\epsilon\,ds \\
=\ &\epsilon.\underbrace{\int_{B(d,h)_{\lambda,w}} ds}_{V_{\lambda,w}(N;h)}
\end{aligned}$$

where $B(d,h)_{\lambda,w}$ is the N-dimensional ball of radius h, centered on d, at the sense of norm $\|.\|_{\lambda,w}$ and $V_{\lambda,w}(N;h)$ is the volume of this ball. Hence, we have

$$\forall\epsilon>0,\forall h<\min(\delta_\epsilon,\min(w_k)),\quad \left|\underbrace{\int_{B(d,h)_{\lambda,w}} f^{\mathcal{K}}_{d|b}(s)\,ds}_{P(\|\boldsymbol{d}-d\|_{\lambda,w}\leq h|\boldsymbol{b}=b)}-f_{|b}(b,d).\underbrace{\int_{B(d,h)_{\lambda,w}} ds}_{\equiv V_{\lambda,w}(N;h)}\right|\leq\epsilon V_{\lambda,w}(N;h).$$

As $V_{\lambda,w}(N;h)\ f^{\mathcal{K}}_{d|b}(d)>0$, this means

$$\forall\epsilon>0,\forall h<\min(\delta_\epsilon,\min(w_k)),\quad \|\frac{P(\|\boldsymbol{d}-d\|_{\lambda,w}\leq h|\boldsymbol{b}=b)}{V_{\lambda,w}(N;h)\ f^{\mathcal{K}}_{d|b}(d)}-1\|\leq\frac{\epsilon}{f_{|b}(b,d)}\leq C,$$

with $C>0$. Consequently, and because $V_{\lambda,w}(N;h)\sim C'.h^K$ with $C'>0$, we have

$$P(\|\boldsymbol{d}-d\|_{\lambda,w}\leq h|\forall k\in\mathcal{K},\ \boldsymbol{b}^k=b^k)=\underbrace{V_{\lambda,w}(N;h)}_{\sim C'.h^K}\underbrace{f^{\mathcal{K}}_{d|b}(d)}_{>0}\sim C''.h^K\quad\text{when } h \text{ tends to } 0,$$

with $C''>0$. Going back to equality (A.2), it results in

$$\phi_{d^B,\mathcal{X}}(h)=P(d^B(\mathcal{X},\chi)\leq h)=\underbrace{P(\forall k\in\mathcal{K}\ \boldsymbol{b}^k=b^k)}_{>0}\underbrace{P(\|\boldsymbol{d}-d\|_{\lambda,w}\leq h|\forall k\in\mathcal{K}\ \boldsymbol{b}^k=b^k)}_{\sim C''.h^K}\sim C''.h^K,$$

for h small enough. □

Lemma 4. *Assuming that there exists a finite constant $C>0$ such that*

$$\phi_{d,\mathcal{X}}(\epsilon)\sim C\epsilon^\tau\quad as\ \epsilon\to 0,$$

and under the conditions (3.9), (3.8), (3.11), (3.10) *and* (3.3), *the functional kernel regression estimate can reach the rate of convergence:*

$$\hat{r}(\chi)-r(\chi)=O_{a.co.}\Big((\frac{\log n}{n})^{\frac{\beta}{2\beta+\tau}}\Big)$$

Proof. This is proved in [1]. □

The proof of Theorem 1 is a direct consequence of Lemmas 3 and 4, for the semimetric $d = d^B$.

AppendixA.2. Proof of Theorem 2

In order to prove Theorem 2, we consider the following Lemmas.

Lemma 5. *Under conditions of Theorem 2, we have*

$$\exists C, C' > 0, \text{ such that } \quad \frac{C}{n\Phi_H} + \frac{n-1}{n} b_H \leq MISE^*(H).$$

Proof. We start with the following decomposition of $MISE^*(H)$, that holds because of condition (3.15):

$$\begin{aligned} MISE^*(H) &= \int \mathbb{E}((\frac{1}{n}\sum_{i=1}^{n}\delta_{i\chi})^2)\, W(\chi) dP_\chi(\chi) \\ &= \int \mathbb{E}(\frac{1}{n^2}\sum_{i=1}^{n}\delta_{i\chi}^2)\, W(\chi) dP_\chi(\chi) + \int \frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} \mathbb{E}(\delta_{i\chi}\delta_{j\chi}|\chi)\, W(\chi) dP_\chi(\chi) \\ &= \frac{1}{n}\int \mathbb{E}(\delta_{i\chi}^2)\, W(\chi) dP_\chi(\chi) + \frac{n-1}{n}\int \mathbb{E}^2(\delta_{i\chi}|\chi)\, W(\chi) dP_\chi(\chi) \\ &= R_3(H) + \frac{n-1}{n} b_H, \end{aligned} \tag{A.3}$$

with

$$R_3(H) = \frac{1}{n}\int \mathbb{E}(\delta_{i\chi}^2)\, W(\chi) dP_\chi(\chi). \tag{A.4}$$

It remains to bound $R_3(H)$ from below. We consider

$$\begin{aligned} \mathbb{E}(\delta_{ix}^2) &= \mathbb{E}\Big((Y_i - r(\chi_i) + r(\chi_i) - r(\chi)K_H(\chi,\chi_i))^2\Big) \\ &= \mathbb{E}\Big((\epsilon_i K_H(\chi,\chi_i) + (r(\chi_i) - r(\chi))K_H(\chi,\chi_i))^2\Big) \\ &= \mathbb{E}\Big(\epsilon K_H^2(\chi,\chi_i)\Big) + \mathbb{E}\Big((r(\chi_i) - r(\chi))^2 K_H^2(\chi,\chi_i)\Big) + 2\mathbb{E}\Big(\epsilon_i(r(\chi_i) - r(\chi))K_H(\chi,\chi_i)\Big). \end{aligned}$$

The last term of this equation is null, because of condition (3.21), and the second term is positive. Thus, using conditions (3.22) and (3.16), we have

$$\begin{aligned} \mathbb{E}(\delta_{i\chi}^2) &\geq \mathbb{E}(\epsilon_i^2 K_H^2(\chi,\chi_i)) = \mathbb{E}\Big(\mathbb{E}(\epsilon_i^2|\chi_i)\mathbb{E}(K_H^2(\chi,\chi_i)|\chi_i)\Big) \\ &\geq \sigma_0^2 \mathbb{E}\Big(\frac{K^2\Big(\frac{d(\chi,\chi_i)}{h}\Big)}{\mathbb{E}^2\Big(K\Big(\frac{d(\chi,\chi_i)}{h}\Big)\Big)}\Big)\Big) \\ &\geq \frac{\sigma_0^2}{\mathbb{E}^2\Big(K\Big(\frac{d(\chi,\chi_i)}{h}\Big)\Big)}\mathbb{E}\Big(K\Big(\frac{d(\chi,\chi_i)}{h}\Big)\Big) \\ &\geq C'\frac{\sigma_0^2}{\Phi_H}, \end{aligned}$$

for a certain $C' > 0$. Then, Lemma 5 follows from

$$R_3(H) \geq \frac{1}{n}\int C'\frac{\sigma_0^2}{\Phi_H} W(\chi) dP_x \geq \frac{C}{n\Phi_H},$$

where the last inequality holds because of condition (3.19). □

Lemma 6. *Under conditions of Theorem 2, we have*

$$\sup_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}}}\left|\hat{r}_{1H}(\chi)-1\right|\longrightarrow 0 \quad a.s.$$

and

$$\sup_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}}}\left|\hat{r}_{1H}^{-j}(\chi)-1\right|\longrightarrow 0 \quad a.s.$$

Proof. We denote by c_χ center that is the closest to χ in condition (3.20). Using condition (3.20), we observe that $\forall\epsilon>0$,

$$P\Big(\sup_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}}}|\hat{r}_{1H}(\chi)-1|>\epsilon\Big)\le P\Big(\sup_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}}}|\hat{r}_{1H}(\chi)-\hat{r}_{1H}(c_\chi)|>\frac{\epsilon}{2}\Big)+\sharp\mathcal{H}_n d_n\sup_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}\\ 1\le k\le d_n}}\Big(P(|(\hat{r}_{1H}(c_k)-1|>\frac{\epsilon}{2})\Big). \tag{A.5}$$

We first consider the first term on the right side of the inequality. We have

$$\begin{aligned}
&|\hat{r}_{1H}(\chi)-\hat{r}_{1H}(c_\chi)| \qquad\qquad (A.6)\\
=&\left|\frac{\sum_{i=1}^n\Delta_i(\chi)}{n\mathbb{E}\big(\Delta_i(\chi)\big)}-\frac{\sum_{i=1}^n\Delta_i(c_\chi)}{n\mathbb{E}\big(\Delta_i(c_\chi)\big)}\right|\\
=&\left|\frac{\sum_{i=1}^n\Delta_i(\chi)\mathbb{E}\big(\Delta_i(c_\chi)\big)-\sum_{i=1}^n\Delta_i(c_\chi)\mathbb{E}\big(\Delta_i(\chi)\big)}{n\mathbb{E}\big(\Delta_i(\chi)\big)\mathbb{E}\big(\Delta_i(c_\chi)\big)}\right|\\
=&\left|\frac{\sum_{i=1}^n\Big(\big(\Delta_i(\chi)-\Delta_i(c_\chi)\big)\mathbb{E}\big(\Delta_i(c_\chi)\big)+\Delta_i(c_\chi)\big(\mathbb{E}\big(\Delta_i(c_\chi)\big)-\mathbb{E}\big(\Delta_i(\chi)\big)\big)\Big)}{n\mathbb{E}(\Delta_i(\chi))\mathbb{E}(\Delta_i(c_\chi))}\right|.
\end{aligned}$$

Then, because the kernel is Lipschitz on $\mathbb{R}^+$ by condition (3.16), we know that

$$\left|\Delta_i(\chi)-\sum_{i=1}^n\Delta_i(c_\chi)\right|\le C\frac{d(\chi,c_\chi)}{h}\le C\frac{r_n}{h},$$

and

$$\left|\mathbb{E}\big(\Delta_i(c_\chi)-\Delta_i(\chi)\big)\right|\le\mathbb{E}\Big(\Big|\Delta_i(c_\chi)-\Delta_i(\chi)\Big|\mathbf{1}_{d(\chi_i,\chi)\le h\cup d(\chi_i,c_\chi)\le h}\Big)\le C\frac{r_n}{h}\Phi_H$$

Thus, equation (A.6) gives

$$\sup_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}}}|\hat{r}_{1H}(\chi)-\hat{r}_{1H}(c_\chi)|\le C\frac{r_n}{\inf\limits_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}}}n\Phi_H},$$

which tends to 0 because of condition (3.18), so that

$$P(\sup_{\substack{H\in\mathcal{H}_n\\ \chi\subset\mathcal{W}}}|\hat{r}_{1H}(\chi)-\hat{r}_{1H}(c_\chi)|>\frac{\epsilon}{2})=0 \tag{A.7}$$

for n large enough. We now consider the second term on the right side of inequality (A.5). We have

$$\hat{r}_{1H}(c_k)-1=\frac{1}{n}\sum_{j=1}^n(K_H(c_k,\chi_j)-1)=\frac{1}{n\mathbb{E}\big(\Delta_j(c_k)\big)}\sum_{j=1}U_j,$$

with

$$U_j = \Delta_j(c_k) - \mathbb{E}\Big(\Delta_j(c_k)\Big).$$

We would like to make use of a Bernstein inequality (see Van der Vaart and Wellner [25, Lemma 2.2.11], for instance), for the U_j. By construction, we have $\mathbb{E}(U_j) = 0$, and $|U_j|$ is bounded, as the kernel is bounded. Moreover, using condition (3.16) and the fact that $\Phi_H \leq 1$,

$$\mathbb{E}\left(\Big(\Delta_j(c_k) - \mathbb{E}\Big(\Delta_j(c_k)\Big)\Big)^2\right) = \mathbb{E}\Big(\Delta_j^2(c_k)\Big) - \mathbb{E}^2\Big(\Delta_j(c_k)\Big) \leq C_{2,2}\Phi_H + C_{2,1}^2\Phi_H^2 \leq C\Phi_H.$$

Finally , as U_j is bounded and for $m \geq 2$, we have

$$\mathbb{E}(|U_j|^m) = \mathbb{E}(|U_j|^2|U_j|^{m-2}) \leq \mathbb{E}(|U_j|^2 C^{m-2}) \leq C\Phi_H \leq \frac{m!}{2}C'\Phi_H.$$

This tells us that we are in the conditions of the Bernstein inequality, with $M = 1$, $v_i = C'\Phi_H$ and $v = \sum_{i=1}^n C'\Phi_H = nC'\Phi_H$ This ensures that, for all ϵ positive,

$$\begin{aligned}
P(|(\hat{r}_{1H}(c_k) - 1| > \epsilon) &= P\left(\left|\frac{1}{n\mathbb{E}\Big(\Delta_j(c_k)\Big)}\sum_{j=1}^n U_j\right| > \epsilon\right) \\
= P\left(\left|\sum_{j=1}^n U_j\right| > \epsilon n\mathbb{E}\Big(\Delta_j(c_k)\Big)\right) &\leq P\left(\left|\sum_{j=1}^n U_j\right| > \epsilon n C_{1,1}\Phi_H\right) \\
\leq 2\exp\left(-\frac{(\epsilon n C_1\Phi_H)^2}{2(nC'\Phi_H + \epsilon n C_1\Phi_H)}\right) &\leq 2\exp\left(-\frac{1}{2}\frac{\epsilon^2 C_{1,1}^2 n\Phi_H}{C' + \epsilon C_{1,1}}\right) \\
&= 2\exp(-C_\epsilon n\Phi_H)
\end{aligned}$$

where condition (3.16) is used. From this expression and from equations (A.5) and (A.7), and by using conditions (3.18), (3.20) and(3.24), it follows that $\forall\epsilon > 0$,

$$\begin{aligned}
0 &\leq \sum_{n=1}^\infty P\Big(\sup_{\substack{H\in\mathcal{H}_n \\ \chi\subset\mathcal{W}}} |\hat{r}_{1H}(\chi) - 1| > \epsilon\Big) \\
&\leq \sum_{n=1}^\infty P\Big(\sup_{\substack{H\in\mathcal{H}_n \\ \chi\subset\mathcal{W}}} |\hat{r}_{1H}(\chi) - \hat{r}_{1H}(c_\chi)| > \frac{\epsilon}{2}\Big) + \sum_{n=1}^\infty \sharp\mathcal{H}_n d_n \sum_{n=1}^\infty \sup_{\substack{H\in\mathcal{H}_n \\ \chi\subset\mathcal{W} \\ 1\leq k\leq d_n}} P(|\hat{r}_{1H}(c_k) - 1| > \frac{\epsilon}{2}) \\
&\leq \sum_{n=1}^\infty P\Big(C\frac{r_n}{\inf\limits_{\substack{H\in\mathcal{H}_n \\ \chi\subset\mathcal{W}}} n\Phi_H} > \frac{\epsilon}{2}\Big) + \sum_{n=1}^\infty n^{\alpha+\eta} \sup_{\substack{H\in\mathcal{H}_n \\ \chi\subset\mathcal{W} \\ 1\leq k\leq d_n}} 2\exp(-C_\epsilon n\Phi_H) \\
&\leq C + \sum_{n=1}^\infty n^{\alpha+\eta} 2\exp(-C_\epsilon \inf_{\substack{H\in\mathcal{H}_n \\ \chi\subset\mathcal{W} \\ 1\leq k\leq d_n}} (n\Phi_H)) \\
&\leq C + \sum_{n=1}^\infty n^{\alpha+\eta} 2\exp(-C'n^\delta) \leq C'' \qquad \text{(A.8)}
\end{aligned}$$

This ensures the uniform almost complete convergence of $\hat{r}_{1H}(\chi)$ to 1, itself implying the required almost sure convergence. The convergence to 1 of $\hat{r}_{1H}^{-j}(\chi)$ is shown using very similar ideas. In that case, (A.5) is replaced by

$$P\Big(\sup_{\substack{H\in\mathcal{H}_n \\ j=1\ldots n \\ \chi\subset\mathcal{W}}} |\hat{r}_{1H}^{-j}(\chi) - 1| > \epsilon\Big) \leq P\Big(\sup_{\substack{H\in\mathcal{H}_n \\ j=1\ldots n \\ \chi\subset\mathcal{W}}} |\hat{r}_{1H}^{-j}(\chi) - \hat{r}_{1H}^{-j}(c_\chi)| > \frac{\epsilon}{2}\Big) + \sharp\mathcal{H}_n d_n n \sup_{\substack{H\in\mathcal{H}_n \\ j=1\ldots n \\ \chi\subset\mathcal{W} \\ 1\leq k\leq d_n}} \Big(P(|(\hat{r}_{1H}^{-j}(c_k) - 1| > \frac{\epsilon}{2})\Big). \qquad \text{(A.9)}$$

□

Lemma 7. *We consider the quantity $\hat{g}_H(\chi) = \hat{r}_{2H}(\chi) - \hat{r}(\chi)\hat{r}_{1H}(\chi)$, and we note that $B(x) = \mathbb{E}(\hat{g}_H(\chi))$. Under conditions of Theorem 2, we have*

1. *The estimator $\hat{g}_H(\chi)$ has the form of a delta sequence:*

$$\hat{g}_H(\chi) = \frac{1}{n}\sum_{i=1}^{n} \delta_{i\chi}, \tag{A.10}$$

with $\delta_{i\chi} = (Y_i - r(\chi))K_H(\chi, \chi_i)$.

2. *For $k = 1, 2, \ldots$ there is a constant C_k so that for any $m = 2 \ldots 2k$, we have*

$$\left| \int \ldots \int \left[\prod_{i,i'=1}^{m} \delta_{i'i}^{\alpha_{i'i}} \right] \left[\prod_{i=1}^{m} W(\chi_i)^{\beta_i} \right] dP_\chi(\chi_1) dP_\chi(\chi_m) \right| \leq C_k \left(\frac{1}{\Phi_H} \right)^{k - \frac{m}{2}}, \tag{A.11}$$

where $\alpha_{i'i} = 0 \ldots k$ are subject to

$$\sum_{i,i'=1}^{m} \alpha_{i'i} = k, \tag{A.12}$$

and the restriction that

$$\text{for each } i = 1, \ldots, m, \text{ there is an } i' \neq i \text{ so that either } \alpha_{i'i} \text{ or } \alpha_{ii'} \text{ is non zero}, \tag{A.13}$$

and where $\beta_i = 0, 1$, with $\beta_i = 1$ any time an $\alpha_{i'i} \geq 1$, and with $W(\chi_i)^{\beta_i}$ taken to be 1 when $W(\chi_i) = \beta_i = 0$.

3. *The quantity $\tilde{\delta}_{ij} = \int \delta_{j\chi}\delta_{i\chi} W(\chi) dP(\chi)$ is such that for $k = 1, 2, \ldots$ there is a constant C_k so that for any $m = 2 \ldots 2k$, we have*

$$\left| \int \ldots \int \left[\prod_{i,i'=1}^{m} \widetilde{\delta}_{i'i}^{\alpha_{i'i}} \right] dP_\chi(\chi_1) dP_\chi(\chi_m) \right| \leq C_k \left(\frac{1}{\Phi_H} \right)^{k - \frac{m}{2}}, \tag{A.14}$$

where $\alpha_{i'i} = 0 \ldots k$ are subject to

$$\sum_{i,i'=1}^{m} \alpha_{i'i} = k, \tag{A.15}$$

and the restriction that

$$\text{for each } i = 1 \ldots m, \text{ there is an } i' \neq i \text{ so that either } \alpha_{ii'} \text{ or } \alpha_{ii'} \text{ is non zero}. \tag{A.16}$$

4. *There exists a constant $C > 0$ such that*

$$\int \int \widetilde{\delta}_{ij} dP_\chi(\chi_i) dP_\chi(\chi_j) \leq C. \tag{A.17}$$

5. *There exists a constant $C' > 0$ such that*

$$\int \widetilde{\delta}_{\chi\chi} dP_\chi(\chi) \geq \frac{C}{\Phi_H}. \tag{A.18}$$

6. *There exists $\xi > 0$ so that for $k = 1, 2, \ldots$ there is a constant $C_k > 0$ such that*

$$\int B(\chi)^{2k} W(\chi) dP_\chi(\chi) \leq C_k b_H. \tag{A.19}$$

Proof. The proof of (A.10) is trivial by definition of $\hat{g}_H(\chi)$. The proof of (A.11) is as follows. Because of condition (3.16), we have

$$\prod_{i,i'=1}^{m} \mathbb{E}^{\alpha_{i'i}} \left(K\left(\frac{d(\chi, \chi_i)}{h} \right) \right) = \mathbb{E}^{\sum \alpha_{i'i}} \left(K\left(\frac{d(\chi, \chi_i)}{h} \right) \right) = \mathbb{E}^{k} \left(K\left(\frac{d(\chi, \chi_i)}{h} \right) \right) \geq C_{1,1}^{k} \Phi_H^{k}.$$

Then, using successively, the definition of $K_H(\chi,\chi_i)$ in $\delta_{i\chi}$, the fact that r is bounded, the fact that W is bounded, the Newton binome, condition (3.25), and condition (3.16), we have

$$
\begin{aligned}
&\left|\mathbb{E}\Big(\Big[\prod_{i,i'=1}^{m}\delta_{i'i}^{\alpha_{i'i}}\Big]\Big[\prod_{i=1}^{m}W(\chi_i)^{\beta_i}\Big]\Big)\right| \\
\leq\ & \frac{1}{C_{1,1}^k\Phi_H^k}\left|\mathbb{E}\left(\left[\prod_{i,i'=1}^{m}\Big((Y_{i'}-r(\chi_i))K\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)\Big)^{\alpha_{i'i}}\right]\Big[\prod_{i=1}^{m}W(\chi_i)^{\beta_i}\Big]\right)\right| \\
\leq\ & \frac{1}{C_{1,1}^k\Phi_H^k}\left|\mathbb{E}\left(\left[\prod_{i,i'=1}^{m}\Big((|Y_{i'}|+C_r)K\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)\Big)^{\alpha_{i'i}}\right]\Big[\prod_{i=1}^{m}W(\chi_i)^{\beta_i}\Big]\right)\right| \\
\leq\ & \frac{1}{C_{1,1}^k\Phi_H^k}\left|\mathbb{E}\left(\left(\mathbb{E}\left[\prod_{i,i'=1}^{m}\Big((|Y_{i'}|+C_r)K\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)\Big)^{\alpha_{i'i}}\right]\Big[\prod_{i=1}^{m}W(\chi_i)^{\beta_i}\Big]\Big|\chi_1\dots\chi_m\right)\right)\right| \\
\leq\ & \frac{1}{C_{1,1}^k\Phi_H^k}\left|\mathbb{E}\left(C_W\prod_{i,i'=1}^{m}\mathbb{E}\Big((|Y_{i'}|+C_r)^{\alpha_{i'i}}\Big|\chi_1\dots\chi_m\Big)K^{\alpha_{i'i}}\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)\right)\right| \qquad \text{(A.20)} \\
\leq\ & \frac{1}{C_{1,1}^k\Phi_H^k}\left|\mathbb{E}\left(C_W\prod_{i,i'=1}^{m}2^{\alpha_{i'i}}\Big(\mathbb{E}\Big(|Y_{i'}|^{\alpha_{i'i}}\Big|\chi_1\dots\chi_m\Big)+C_r{}^{\alpha_{i'i}}\Big)K^{\alpha_{i'i}}\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)\right)\right| \qquad \text{(A.21)} \\
\leq\ & \frac{1}{C_{1,1}^k\Phi_H^k}\left|\mathbb{E}\left(C_W\prod_{i,i'=1}^{m}\sup_{s\in 1,2,\dots k}\Big(2^s(C_k^s+C_r^s)\Big)K^{\alpha_{i'i}}\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)\right)\right| \qquad \text{(A.22)} \\
\leq\ & \frac{1}{C_{1,1}^k\Phi_H^k}\left|\mathbb{E}\left(C_W\prod_{i,i'=1}^{m}C_k'K^{\alpha_{i'i}}\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)\right)\right| \qquad \text{(A.23)} \\
=\ & \frac{C_k''}{\Phi_H^k}\left|\mathbb{E}\left(\prod_{i,i'=1}^{m}K\Big(\frac{d(\chi_i,\chi_{i'})}{h}\Big)^{\alpha_{i'i}}\right)\right| \leq C_k''\Phi_H^{-k+\frac{m}{2}}. \qquad \text{(A.24)}
\end{aligned}
$$

The last inequality comes from condition (3.16) and from the fact that it is always possible to find floor($\frac{m+1}{2}$) pairs (i_l, j_l) such that for all $l = 1\dots\text{floor}(\frac{m+1}{2})$,$i_l \neq j_l$ and i_l or j_l is unique among the set of pairs $(i_l, j_l)_{l=1\dots\text{floor}(\frac{m+1}{2})}$. There are thus at least $\frac{m}{2}$ pairs i, i' such that $\alpha_{i'i}$ is non zero. As a consequence, (A.24) is valid and it proves statement (A.11) . The proof of statement (A.14) follows nearly the same steps, and we have

$$
\left|\mathbb{E}\left(\left[\prod_{i,i'=1}^{m}\widetilde{\delta}_{i'i}^{\alpha_{i'i}}\right]\right)\right| \leq \frac{1}{C_{1,2}^{2k}\Phi_H^{2k}}\left|\mathbb{E}\left(\prod_{i,i'=1}^{m}\left(\int K\Big(\frac{d(\chi_i,\chi)}{h}\Big)K\Big(\frac{d(\chi_{i'},\chi)}{h}\Big)dP_\chi(\chi)\right)^{\alpha_{i'i}}\right)\right|.
$$

Again, we know that there exists at least $\frac{m}{2}$ pairs i, i' such that $\alpha_{i'i}$ is non zero and i or i' is unique. For those pairs, we use Holder's inequality and condition (3.16) to see that

$$
\mathbb{E}\left(\Big(\int K\Big(\frac{d(\chi_i,\chi)}{h}\Big)K\Big(\frac{d(\chi_{i'},\chi)}{h}\Big)dP_\chi(\chi)\Big)^{\alpha_{i'i}}\Big|\chi_i,\chi_{i'},\ i\neq i'\right) \leq C\Phi_H^{\alpha_{i'i}}.
$$

As the kernel is bounded, this integral is bounded for all other pairs. Thus, we have

$$
\left|\mathbb{E}\left(\prod_{i,i'=1}^{m}\widetilde{\delta}_{i'i}^{\alpha_{i'i}}\right)\right| \leq \frac{1}{C_{1,2}^{2k}\Phi_H^{2k}}C{\Phi_H}^{k+m/2} = (\frac{1}{\Phi_H})^{(k-\frac{m}{2})},
$$

which proves statement (A.14). We now consider statement (A.17). Using the Theorem of Fubini to permute the

integrals, we have

$$\Big|\int\int\int \delta_{j\chi}\delta_{i\chi}W(\chi)dP_\chi(\chi)dP_\chi(\chi_i)dP_\chi(\chi_j)\Big| = \Big|\int\int \delta_{j\chi}dP_\chi(\chi_j)\int \delta_{i\chi}dP_\chi(\chi_i)W(\chi)dP_\chi(\chi)\Big|$$
$$= \mathbb{E}(\mathbb{E}(\delta_{j\chi})\mathbb{E}(\delta_{i\chi})W(\chi)).$$

Then, using condition (3.25) and the fact that r is bounded, we have for $j = i, i'$,

$$\begin{aligned}
\mathbb{E}(|\delta_{j\chi}|) &\leq \mathbb{E}\left(\left|(Y_j - r(\chi))\frac{K\left(\frac{d(\chi,\chi_j)}{h}\right)}{\mathbb{E}\left(K\left(\frac{d(\chi,\chi_j)}{h}\right)\right)}\right|\right) \\
&\leq \mathbb{E}\left(\mathbb{E}(|Y_j|\chi_j)\frac{K\left(\frac{d(\chi,\chi_j)}{h}\right)}{\mathbb{E}\left(K\left(\frac{d(\chi,\chi_j)}{h}\right)\right)}\right) + \mathbb{E}\left(C_r\frac{K\left(\frac{d(\chi,\chi_j)}{h}\right)}{\mathbb{E}\left(K\left(\frac{d(\chi,\chi_j)}{h}\right)\right)}\right) \\
&\leq (C_1 + C_r)\mathbb{E}\left(\frac{K\left(\frac{d(\chi,\chi_j)}{h}\right)}{\mathbb{E}\left(K\left(\frac{d(\chi,\chi_j)}{h}\right)\right)}\right) = C_1 + C_r.
\end{aligned}$$

Combining this result and the fact that W is bounded, we have

$$\mathbb{E}\Big(\mathbb{E}(|\delta_{j\chi}|)\mathbb{E}(|\delta_{i\chi}|)W(\chi)\Big) \leq C,$$

which proves statement (A.17). In a view to prove statement (A.18), we note that

$$(Y_i - r(\chi))^2 = ((Y_i - r(\chi_i) + r(\chi_i) - r(\chi))^2 = (\epsilon_i + r(\chi_i) - r(\chi))^2 = \epsilon_i^2 + 2\epsilon_i(r(\chi_i) - r(\chi)) + (r(\chi_i) - r(\chi))^2,$$

and

$$\mathbb{E}((Y_i - r(\chi))^2|\chi_i) \geq \mathbb{E}(\epsilon_i^2|\chi_i)$$

because of condition (3.21). Then, using conditions (3.22) and (3.16) successively, we have

$$\begin{aligned}
\int \widetilde{\delta}_{ii}dP_\chi(\chi_i) &= \int\int \delta_{i\chi}\delta_{i\chi}W(\chi)dP_\chi(\chi)dP_\chi(i) \\
&\geq C\mathbb{E}\left((Y_i - r(\chi))^2\frac{K^2\left(\frac{d(\chi,\chi_j)}{h}\right)}{\mathbb{E}^2\left(K\left(\frac{d(\chi,\chi_j)}{h}\right)\right)}\right) \\
&\geq C\mathbb{E}\left(\mathbb{E}(\epsilon_i^2|\chi_i)\frac{K^2\left(\frac{d(\chi,\chi_j)}{h}\right)}{\mathbb{E}^2\left(K\left(\frac{d(\chi,\chi_j)}{h}\right)\right)}\right) \\
&\geq C\sigma_0^2\mathbb{E}\left(\frac{K^2\left(\frac{d(\chi,\chi_j)}{h}\right)}{\mathbb{E}^2\left(K(\frac{d(\chi,\chi_j)}{h})\right)}\right) \geq \frac{C}{\Phi_H}.
\end{aligned}$$

Finally, statement (A.19) is shown true by taking $C_k = C^{2k-2}$. □

Lemma 8. *Under conditions of Theorem 2, we have*

$$\sup_{H\in\mathcal{H}_n}\left|\frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)}\right| \longrightarrow 0 \quad a.s,$$

with

$$ISE^*(H) = \frac{1}{n^2}\int\sum_{j,k=1}^{n}\delta_{j\chi}\delta_{k\chi}W(\chi)dP_\chi(\chi). \tag{A.25}$$

Proof. The proof relies on the following decomposition:

$$ISE^*(H) = R_1(H) + R_2(H) + R_3(H) + 2(1-\frac{1}{n})S(H) + b_H(1-\frac{1}{n}),$$

with $R_3(H)$ defined by (A.4) and

$$\begin{aligned} R_1(H) &= \frac{1}{n^2}\int \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \Big(\delta_{ix} - \mathbb{E}(\delta_{ix})\Big)\Big(\delta_{jx} - \mathbb{E}(\delta_{jx}))\Big)W(\chi)dP_\chi(\chi), \\ R_2(H) &= \frac{1}{n^2}\int \sum_{i=1}^{n} \Big(\delta_{ix}^2 - \mathbb{E}(\delta_{ix}^2)\Big)W(\chi)dP_\chi(\chi), \qquad \text{(A.26)} \\ S(H) &= \frac{1}{n}\int \mathbb{E}\Big(\delta_{jx}\Big)\sum_{i=1}^{n}(\delta_{ix} - \mathbb{E}(\delta_{ix})\Big)W(\chi)dP_\chi(\chi) \qquad \text{(A.27)} \end{aligned}$$

Those definitions meet the one of Marron and Hardle [21]. Then, because of this decomposition, and because of decomposition (A.3), we have

$$\begin{aligned} \sup_{H\in\mathcal{H}_n}\left|\frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)}\right| &= \sup_{H\in\mathcal{H}_n}\left|\frac{R_1(H) + R_2(H) + 2(1-\frac{1}{n})S(H)}{MISE^*(H)}\right| \\ &\leq \sup_{H\in\mathcal{H}_n}\left|\frac{R_1(H)}{MISE^*(H)}\right| + \sup_{H\in\mathcal{H}_n}\left|\frac{R_2(H)}{MISE^*H)}\right| \\ &\quad +2\frac{n-1}{n}\sup_{H\in\mathcal{H}_n}\left|\frac{S(H)}{MISE^*(H)}\right|. \end{aligned}$$

We then use exactly the same steps as Marron and Hardle [21, p.105], with conditions (3.1) and (3.3) in that paper replaced by conditions (A.10), (3.24), and (3.17) and (3.18) respectively, and where the conditions equivalent to (3.4) to (3.7) in that paper are shown valid by Lemma 7. This allows to show that for all $k = 1, 2, \ldots$, we have

$$\mathbb{E}\left(\Big(\frac{R_1(H)}{MISE^*(H)}\Big)^{2k}\right) \leq C_k n^{-\gamma k}, \quad \mathbb{E}\left(\Big(\frac{R_2(H)}{MISE^*(H)}\Big)^{2k}\right) \leq C_k n^{-\gamma k}, \text{ and } \mathbb{E}\left(\Big(\frac{S(H)}{MISE^*(H)}\Big)^{2k}\right) \leq C_k n^{-\gamma k}. \qquad \text{(A.28)}$$

Then, given this, given condition (3.24), and using Markov inequality, we have

$$\begin{aligned} P(\sup_{H\in\mathcal{H}_n}\left|\frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)}\right| > \epsilon) &\leq \#\mathcal{H}_n \sup_{H\in\mathcal{H}_n} P(\left|\frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)}\right| > \epsilon) \\ &\leq C_n n^\alpha \frac{1}{\epsilon^{2k}}\sup_{H\in\mathcal{H}_n}\mathbb{E}((\frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)})^{2k}). \\ &\leq C_n n^\alpha \frac{1}{\epsilon^{2k}} C_k n^{-\gamma k} = Cn^{-\gamma k+\alpha} \end{aligned}$$

As $\alpha, \gamma > 0$, the validity of equation (A.28) for all $k = 1, 2, \ldots$ tells us that, for all $\epsilon > 0$, one can always chose a k such that $\gamma k - \alpha > 1$ and a N large enough so that for all $n > N$, the above probability is smaller than ϵ. This indicates the almost sure convergence of $\sup_{H\in\mathcal{H}_n}\left|\frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)}\right|$, which proves the Lemma. □

Lemma 9. *Under conditions of Theorem 2, we have*

$$\sup_{H\in\mathcal{H}_n}\left|\frac{MISE^*(H) - ASE^*(H)}{MISE^*(H)}\right| \longrightarrow 0 \quad a.s,$$

with

$$ASE^*(H) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_{2H}(\chi_i) - r(\chi_i)\hat{r}_{1H}(\chi_i)\right)^2 W(\chi_i) = \frac{1}{n^3}\sum_{i,j,k=1}^{n}\delta_{ji}\delta_{ki}W(\chi_i). \tag{A.29}$$

Proof. The proof of this Lemma follows the same steps as Theorem 4 in Marron and Hardle [21]. First, long but simple calculations lead to the following decomposition of $ASE^*(H)$:

$$\begin{aligned} ASE^*(H) &= \frac{n-2}{n}ISE^*(H) \\ &+ T_1(H) + T_2(H) + \frac{n-1}{n}T_3(H) + 2T_4(H) + 2T_5(H) + T_6(H) + T_7(H) \\ &+ 2\frac{n-2}{n}U_1(H) + 2\frac{n-1}{n}(U_2(H) + U_3(H)) + \frac{(n-2)(n-1)}{n^2}V(H) \\ &+ \frac{1}{n}R_2(H) + \frac{1}{n}R_3(H), \end{aligned} \tag{A.30}$$

with $ISE^*(H)$, $R_2(H)$ and $R_3(H)$ defined by equations (A.25), (A.26) and (A.4), and with the following definitions:

$$\begin{aligned} T_1(H) &= \frac{1}{n^3}\sum_{\substack{i,j,k=1\\ i\neq j\neq k\neq i}}^{n}(\delta_{ij}\delta_{ik}W(\chi_i) + \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)) - \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_j,\chi_k) - \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_i,\chi_k) \\ &\quad -\mathbb{E}(\delta_{ij}\delta_{ik}W(\chi)|\chi_i,\chi_j) + \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_k) + \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_j) - \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_i)), \\ T_2(H) &= \frac{1}{n^3}\sum_{\substack{i,j=1\\ i\neq j}}^{n}(\delta_{ij}^2W(\chi_i) - \mathbb{E}(\delta_{ij}^2W(\chi_i)|\chi_i) - \mathbb{E}(\delta_{ij}^2W(\chi_i)|\chi_j) + \mathbb{E}(\delta_{ij}^2W(\chi_i))) \\ T_3(H) &= \frac{1}{n^2}\sum_{i=1}^{n}(\mathbb{E}(\delta_{ij}^2W(\chi_i)|\chi_i) - \mathbb{E}(\delta_{ij}^2W(\chi_i))), \\ T_4(H) &= \frac{1}{n^3}\sum_{\substack{i,j=1\\ i\neq j}}^{n}(\delta_{ii}\delta_{ij}W(\chi_i) - \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i)|\chi_i) - \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i)|\chi_j) + \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i))), \\ T_5(H) &= \frac{1}{n^3}\sum_{\substack{i,j=1\\ i\neq j}}^{n}(\mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i)|\chi_j) - \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i))), \\ T_6(H) &= \frac{1}{n^3}\sum_{i=1}^{n}(\delta_{ii}^2W(\chi_i) - \mathbb{E}(\delta_{ii}^2W(\chi_i))), \\ T_7(H) &= \frac{1}{n^2}\mathbb{E}(\delta_{ii}^2W(\chi_i)), \\ U_1(H) &= \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}(\delta_{ij}W(\chi_i)B(\chi_i) - \mathbb{E}(B(\chi_i)W(\chi_i)\delta_{ij}|\chi_j)), \\ U_2(H) + U_3(H) &= \frac{1}{n^2}\sum_{i=1}^{n}\int\delta_{ii}\delta_{ix}W(\chi_i)dP_\chi(\chi), \\ V(H) &= \frac{1}{n}\sum_{i=1}^{n}(W(\chi_i)B^2(\chi_i) - \mathbb{E}(B^2(\chi_i)W(\chi_i))). \end{aligned}$$

Those terms are the same as the ones identified by the same name in Marron and Hardle [21]. As a consequence of

this decomposition, we have

$$\begin{aligned}
&\frac{AS\,E^*(H) - MIS\,E^*(H)}{MIS\,E^*(H)} \qquad (A.31)\\
=&\left(\frac{\frac{n-2}{n} IS\,E^*(H) + \frac{1}{n} R_3(H) - MIS\,E^*(H)}{MIS\,E^*(H)} \right) + \left(\frac{T_1(H)}{MIS\,E^*(H)} \right) + \left(\frac{T_2(H)}{MIS\,E^*(H)} \right)\\
+ &\frac{n-1}{n}\left(\frac{T_3(H)}{MIS\,E^*(H)} \right) + 2\left(\frac{T_4(H)}{MIS\,E^*(H)} \right) + 2\left(\frac{T_5(H)}{MIS\,E^*(H)} \right) + \left(\frac{T_6(H)}{MIS\,E^*(H)} \right) + \left(\frac{T_7(H)}{MIS\,E^*(H)} \right)\\
+ &\frac{2(n-2)}{n}\left(\frac{U_1(H)}{MIS\,E^*(H)} \right) + 2\frac{n-1}{n}\left(\frac{U_2(H) + U_3(H)}{MIS\,E^*(H)} \right) + \frac{(n-2)(n-1)}{n^2}\left(\frac{V(H)}{MIS\,E^*(H)} \right)\\
+ &\frac{1}{n}\left(\frac{R_2(H)}{MIS\,E^*(H)} \right)
\end{aligned}$$

In this expression, the eleven last terms on the right hand side are shown to converge to 0 almost surely in exactly the same manner as in Marron and Hardle [21], with conditions (3.1) and (3.3) in that paper replaced by conditions (A.10), (3.24), and (3.17) and (3.18) respectively, and where the conditions equivalent to (3.4) to (3.7) in that paper are shown valid by Lemma 7. Now, going back to equation (A.31), it remains to prove that

$$\frac{\frac{n-2}{n} IS\,E^*(H) + \frac{1}{n} R_3(H) - MIS\,E^*(H)}{MIS\,E^*(H)} \longrightarrow 0 \quad a.s. \qquad (A.32)$$

Therefore, we consider the following decomposition:

$$\begin{aligned}
\frac{\frac{n-2}{n} IS\,E^*(H) + \frac{1}{n} R_3(H) - MIS\,E^*(H)}{MIS\,E^*(H)} &= \frac{n-2}{n}\frac{IS\,E^*(H) - MIS\,E^*(H)}{MIS\,E^*(H)}\\
&+ \frac{1}{n}\frac{R_3(H) - MIS\,E^*(H)}{MIS\,E^*(H)} + \frac{1}{n}.
\end{aligned}$$

Using (A.3) and Lemma 5, we observe that

$$\left| \frac{1}{n}\frac{R_3(H) - MIS\,E^*(H)}{MIS\,E^*(H)} \right| = \left| \frac{n-1}{n^2}\frac{b_H}{MIS\,E^*(H)} \right| \leq \frac{1}{n}$$

which converges to 0 as n goes to ∞. We then make use of Lemma 8, and equation (A.32) is proved, which proves the Lemma. □

Lemma 10. *Under conditions of Theorem 2, we have*

$$\sup_{H \in \mathcal{H}_n} \left| \frac{MIS\,E^*(H) - AS\,E(H)}{MIS\,E^*(H)} \right| \longrightarrow 0 \quad a.s.$$

Proof. We first state that

$$\sup_{H \in \mathcal{H}_n} \left| \frac{AS\,E^*(H) - AS\,E(H)}{AS\,E^*(H)} \right| \longrightarrow 0 \quad a.s., \qquad (A.33)$$

with $AS\,E^*(H)$ defined by (A.29). We have, for all $H \in \mathcal{H}_n$,

$$\begin{aligned}
AS\,E(H) &= \frac{1}{n}\sum_{i=1}^{n} (\hat{r}_H(\chi_i) - r(\chi_i))^2\, W(\chi_i)\\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\hat{r}_{2H}(\chi_i) - r(\chi_i)\hat{r}_{1H}(\chi_i)}{\hat{r}_{1H}(\chi_i)} \right)^2 W(\chi_i)\\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\hat{r}_{2H}(\chi_i) - r(\chi_i)\hat{r}_{1H}(\chi_i) \right)^2 W(\chi_i) \left(1 + \frac{1 - \hat{r}_{1H}(\chi_i)}{\hat{r}_{1H}(\chi_i)} \right)^2.
\end{aligned}$$

Expression (A.33) follows from this decomposition, because of Lemma 6. Given this, we can write

$$\begin{aligned}
&\sup_{H\in\mathcal{H}_n}\left|\frac{MIS\,E^*(H)-AS\,E(H)}{MIS\,E^*(H)}\right| \\
&\leq \sup_{H\in\mathcal{H}_n}\left|\frac{MIS\,E^*(H)-AS\,E^*(H)}{MIS\,E^*(H)}\right| + \sup_{H\in\mathcal{H}_n}\left|\frac{AS\,E^*(H)-AS\,E(H)}{MIS\,E^*(H)}\right| \\
&= \sup_{H\in\mathcal{H}_n}\left|\frac{MIS\,E^*(H)-AS\,E^*(H)}{MIS\,E^*(H)}\right| + \sup_{H\in\mathcal{H}_n}\left(\left|\frac{AS\,E^*(H)-AS\,E(H)}{AS\,E^*(H)}\right|\left|\frac{AS\,E^*(H)-MIS\,E^*(H)}{MIS\,E^*(H)}+1\right|\right)
\end{aligned}$$

The proof is then completed because of (A.33) and of Lemma 9. □

Lemma 11. *Under conditions of Theorem 2, we have*

$$\sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{AS\,E}(H)-AS\,E(H)}{MIS\,E^*(H)}\right| \longrightarrow 0 \quad a.s.$$

Proof. For all $H\in\mathcal{H}_n$, we consider the result (A.33) for $AS\,E^*(H)$, and the following decomposition of $\widetilde{AS\,E}(H)$:

$$\begin{aligned}
\widetilde{AS\,E}(H) &= \frac{1}{n}\sum_{j=1}^{n}\left(\hat{r}_H^{-j}(\chi_j)-r(\chi_j)\right)^2 W(\chi_j), \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\hat{r}_{2H}^{-j}(\chi_j)-r(\chi_j)\hat{r}_{1H}^{-j}(\chi_j)}{\hat{r}_{1H}^{-j}(\chi_j)}\right)^2 W(\chi_j) \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\hat{r}_{2H}^{-j}(\chi_j)-r(\chi_j)\hat{r}_{1H}^{-j}(\chi_j)\right)^2 W(\chi_j)\left(1+\frac{1-\hat{r}_{1H}^{-j}(\chi_j)}{\hat{r}_{1H}^{-j}(\chi_j)}\right)^2.
\end{aligned}$$

By the same arguments as in (A.33), it results in

$$\sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{AS\,E}(H)-\widetilde{AS\,E}^*(H)}{\widetilde{AS\,E}^*(H)}\right| \longrightarrow 0 \quad a.s \tag{A.34}$$

with

$$\widetilde{AS\,E}^*(H) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{r}_{2H}^{-j}(\chi_j)-r(\chi_j)\hat{r}_{1H}^{-j}(\chi_j)\right)^2 W(\chi_j).$$

Consequently, we have

$$
\begin{aligned}
&\sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{ASE}(H)-ASE(H)}{MISE^*(H)}\right| \\
&\leq \sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{ASE}^*(H)-\widetilde{ASE}(H)}{MISE^*(H)}\right| + \sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{ASE}^*(H)-ASE(H)}{MISE^*(H)}\right| \\
&\leq \sup_{H\in\mathcal{H}_n}\left(\left|\frac{\widetilde{ASE}^*(H)-\widetilde{ASE}(H)}{\widetilde{ASE}^*(H)}\right|\left(\left|\frac{\widetilde{ASE}^*(H)-MISE^*(H)}{MISE^*(H)}\right|+1\right)\right) \\
&+ \sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{ASE}^*(H)-ASE^*(H)}{MISE^*(H)}\right| + \sup_{H\in\mathcal{H}_n}\left(\left|\frac{ASE^*(H)-ASE(H)}{ASE^*(H)}\right|\left(\left|\frac{ASE^*(H)-MISE^*(H)}{MISE^*(H)}\right|+1\right)\right) \\
&\leq \sup_{H\in\mathcal{H}_n}\left(\left|\frac{\widetilde{ASE}^*(H)-\widetilde{ASE}(H)}{\widetilde{ASE}^*(H)}\right|\left(\left|\frac{\widetilde{ASE}^*(H)-ASE^*(H)}{MISE^*(H)}\right|+\left|\frac{ASE^*(H)-ASE(H)}{MISE^*(H)}\right|+1\right)\right) \\
&+ \sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{ASE}^*(H)-ASE^*(H)}{MISE^*(H)}\right| + \sup_{H\in\mathcal{H}_n}\left(\left|\frac{ASE^*(H)-ASE(H)}{ASE^*(H)}\right|\left(\left|\frac{ASE^*(H)-MISE^*(H)}{MISE^*(H)}\right|+1\right)\right) \\
&\leq \sup_{H\in\mathcal{H}_n}\left(\left|\frac{\widetilde{ASE}^*(H)-\widetilde{ASE}(H)}{\widetilde{ASE}^*(H)}\right|\left(\left|\frac{\widetilde{ASE}^*(H)-ASE^*(H)}{MISE^*(H)}\right|+1\right)\right) \\
&+ \sup_{H\in\mathcal{H}_n}\left(\left|\frac{\widetilde{ASE}^*(H)-\widetilde{ASE}(H)}{\widetilde{ASE}^*(H)}\right|\left(\left|\frac{ASE^*(H)-MISE^*(H)}{MISE^*(H)}\right|\right)\right) \\
&+ \sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{ASE}^*(H)-ASE^*(H)}{MISE^*(H)}\right| + \sup_{H\in\mathcal{H}_n}\left(\left|\frac{ASE^*(H)-ASE(H)}{ASE^*(H)}\right|\left(\left|\frac{ASE^*(H)-MISE^*(H)}{MISE^*(H)}\right|+1\right)\right)
\end{aligned}
$$

Thus, the Lemma will be proved by using (A.33), (A.34) and Lemma 9, plus the fact that

$$\sup_{H\in\mathcal{H}_n}\left|\frac{\widetilde{ASE}^*(H)-ASE^*(H)}{MISE^*(H)}\right| \to 0 \quad a.s.$$

This last result is shown as follows. Because of decomposition (A.30) and of the following decomposition

$$
\begin{aligned}
\widetilde{ASE}^*(H) &= \frac{n-2}{n}ISE^*(H) + T_1(H) + T_2(H) + \frac{n-1}{n}T_3(H) \\
&\quad + \frac{1}{n}(R_2(H)+R_3(H)) + 2\frac{n-2}{n}U_1(H) + \frac{(n-2)(n-1)}{n^2}V(H),
\end{aligned}
$$

we have, for all $H \in \mathcal{H}_n$,

$$
\begin{aligned}
\left|\frac{\widetilde{ASE}^*(H)-ASE^*(H)}{MISE^*(H)}\right| &\leq 2\left|\frac{T_4(H)}{MISE^*(H)}\right| + 2\left|\frac{T_5(H)}{MISE^*(H)}\right| + \left|\frac{T_6(H)}{MISE^*(H)}\right| \\
&\quad + \left|\frac{T_7(H)}{MISE^*(H)}\right| + 2\frac{n-1}{n}\left|\frac{U_2(H)+U_3(H)}{MISE^*(H)}\right|.
\end{aligned}
$$

As discussed for equation (A.31) that contains the same terms, all the terms on the right hand side converge to zero almost surely. This is enough to proof Lemma 11. □

Lemma 12. *Under conditions of Theorem 2, we have*

$$\sup_{H\in\mathcal{H}_n}\left|\frac{CT(H)}{MISE^*(H)}\right| \longrightarrow 0 \quad a.s.$$

Proof. Using the definitions of $CT(H)$, $\hat{r}_{1H}(\chi)$ and $\hat{r}_{2H}(\chi)$, one gets:

$$\begin{aligned}
&\sup_{H\in\mathcal{H}_n} \frac{|CT(H)|}{MISE^*(H)} \\
\leq\ & \sup_{H\in\mathcal{H}_n} \left|(MISE^*(H))^{-1}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j(\hat{r}^{-j}(\chi_j)-r(\chi_j))W(\chi_j)\right| \\
\leq\ & \sup_{H\in\mathcal{H}_n} \left|(MISE^*(H))^{-1}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j\frac{(\hat{r}_{2H}^{-j}(\chi_j)-r(\chi_j)r_{1H}^{-j}(\chi_j))}{r_{1H}^{-j}(\chi_j)}W(\chi_j)\right| \\
\leq\ & \sup_{H\in\mathcal{H}_n} \left|(MISE^*(H))^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\epsilon_j\epsilon_i\Delta_i(\chi_j)\frac{W(\chi_j)}{(n-1)\mathbb{E}\left(\Delta_i(\chi_j)\right)r_{1H}^{-j}(\chi_j)}\right| \\
&+ \sup_{H\in\mathcal{H}_n} \left|(MISE^*(H))^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\epsilon_j(r(\chi_i)-r(\chi_j))\Delta_i(\chi_j)\frac{W(\chi_j)}{(n-1)\mathbb{E}\left(\Delta_i(\chi_j)\right)r_{1H}^{-j}(\chi_j)}\right|
\end{aligned}$$

We note hereafter

$$U_{i,j} = (MISE^*(H))^{-1}\epsilon_j\epsilon_i\Delta_i(\chi_j)\frac{W(\chi_j)}{(n-1)\mathbb{E}\left(\Delta_i(\chi_j)\right)r_{1H}^{-j}(\chi_j)}$$

and

$$V_{i,j} = MISE^*(H))^{-1}\epsilon_j(r(\chi_i)-r(\chi_j))\Delta_i(\chi_j)\frac{W(\chi_j)}{(n-1)\mathbb{E}\left(\Delta(\chi_j)\right)r_{1H}^{-j}(\chi_j)}$$

Now the aim is to show

$$\sup_{H\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}U_{i,j}\right| = o_{a.s.}(1) \tag{A.35}$$

and

$$\sup_{H\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}V_{i,j}\right| = o_{a.s.}(1) \tag{A.36}$$

To state (A.35) and (A.36) we are going to show that for any $\epsilon > 0$ the series

$$\sum_{n=1}^{\infty}P\Big(\sup_{H\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}U_{i,j}\right| > \epsilon\Big) \quad \text{and} \quad \sum_{n=1}^{\infty}P\Big(\sup_{H\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}V_{i,j}\right| > \epsilon\Big)$$

are convergent. We consider the set

$$A_n = \{\inf_{H\in\mathcal{H}_n,\chi\in\mathcal{W},1\leq j\leq n} r_{1H}^{-j}(\chi) < \frac{1}{2}\}.$$

From results obtained in the proof of Lemma 6 it is fairly easy to get for any $s > 0$:

$$\begin{aligned}
P(A_n) &\leq n\#\mathcal{H}_n \sup_{H\in\mathcal{H}_n,\,1\leq j\leq n} P\left(\inf_{\chi\in\mathcal{W}} r_{1H}^{-j}(\chi) < \frac{1}{2}\right) \\
&\leq Cn^{\alpha+1} \sup_{H\in\mathcal{H}_n,\,1\leq j\leq n} P\left(\sup_{\chi\in\mathcal{W}}|r_{1H}^{-j}(\chi)-1| > \frac{1}{2}\right) \\
&\leq C_s n^{-1-s} \text{ for n large enough (i.e. } n\geq n_s).
\end{aligned} \tag{A.37}$$

On the other hand, for any $\epsilon > 0$ and any positive integer k, it comes from Markov inequality and condition (3.24),

$$\begin{aligned} P(\sup_{H\in\mathcal{H}_n} |\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} U_{i,j}| > \epsilon \cap \overline{A}_n) &\leq \#\mathcal{H}_n \sup_{H\in\mathcal{H}_n} P\Big(|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} U_{i,j}\mathbf{1}_{\overline{A}_n}| > \epsilon\Big) \\ &\leq \epsilon^{-2k}\#\mathcal{H}_n \sup_{H\in\mathcal{H}_n} \mathbb{E}\Big(|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} U_{i,j}\mathbf{1}_{\overline{A}_n}|^{2k}\Big) \qquad (A.38) \\ &\leq Cn^{\alpha} \sup_{H\in\mathcal{H}_n} \mathbb{E}\Big(|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} U_{i,j}\mathbf{1}_{\overline{A}_n}|^{2k}\Big) \end{aligned}$$

Hence, it is enough to show that for some positive integer k

$$\sum_{n=1}^{\infty} n^{\alpha} \sup_{H\in\mathcal{H}_n} \mathbb{E}\Big(|\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} U_{i,j}\mathbf{1}_{\overline{A}_n}|^{2k}\Big) < +\infty \qquad (A.39)$$

Now, we define

$$I_{2k} = \{(i_1, \ldots, i_{2k}, j_1, \ldots, j_{2k}) \in \{1, \ldots, n\}^{2k} | i_l \neq j_l \forall 1 \leq l \leq 2k\}.$$

It comes

$$\begin{aligned} &\mathbb{E}\Big(\frac{1}{n^{2k}}\Big(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} U_{i,j}\Big)^{2k}\mathbf{1}_{\overline{A}_n}\Big) \\ = \; & n^{-2k} \sum_{(i_l,j_l)_{1\leq l\leq 2k}\in I_{2k}} \Big|\mathbb{E}\Big(\Pi_{l=1}^{2k} U_{i_l,j_l}\mathbf{1}_{\overline{A}_n}\Big)\Big| \\ \leq \; & n^{-2k} \sum_{m=2}^{4k} \sum_{(i_l,j_l)_{1\leq l\leq 2k}\in J_m} \Big|\mathbb{E}\Big(\Pi_{l=1}^{2k}(MIS\,E^*(H))^{-1}\epsilon_{j_l}\epsilon_{i_l}\Delta_{i_l}(\chi_{j_l})\frac{W(\chi_{j_l})}{(n-1)\mathbb{E}\Big(\Delta_{i_l}(\chi_{j_l})\Big)r_{1H}^{-j_l}(\chi_{j_l})}\mathbf{1}_{\overline{A}_n}\Big)\Big|, \\ \leq \; & n^{-2k} \sum_{m=2}^{2k} \sum_{(i_l,j_l)_{1\leq l\leq 2k}\in J_m} \Big|\mathbb{E}\Big(\Pi_{l=1}^{2k}(MIS\,E^*(H))^{-1}\frac{\epsilon_{j_l}\epsilon_{i_l}\Delta_{i_l}(\chi_{j_l})W(\chi_{j_l})}{(n-1)\mathbb{E}\Big(\Delta_{i_l}(\chi_{j_l})\Big)r_{1H}^{-j_l}(\chi_{j_l})}\mathbf{1}_{\overline{A}_n}\Big)\Big|, \qquad (A.40) \end{aligned}$$

where J_m contains elements of I_{2k} that involve exactly m different indices. The last line comes from the fact that when $m > 2k$ at least one of the ϵ's appears with exponent 1 and hence the mean equals 0 (conditioning w.r.t. $\chi_1, \ldots, \chi_n$). Using the fact that W is bounded, using conditions (3.16) and (3.25), it comes directly from (A.40)

$$\begin{aligned} &\mathbb{E}\Big(\frac{1}{n^{2k}}\Big(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} U_{i,j}\Big)^{2k}\mathbf{1}_{\overline{A}_n}\Big) \\ \leq \; & C'_k n^{-2k}(MIS\,E^*(H)n\Phi_H)^{-2k} \sum_{m=2}^{2k} \sum_{(i_l,j_l)_{1\leq l\leq 2k}\in J_m} \mathbb{E}\Big(\Pi_{l=1}^{2k}\Delta_{i_l}(\chi_{j_l})W(\chi_{j_l})\Big) \\ \leq \; & C'_k n^{-2k} \sum_{m=2}^{2k} \#J_m \Phi_H^{\frac{m}{2}} \\ \leq \; & C'_k \Phi_H^k \sum_{m=2}^{2k} (n\Phi_H)^{m-2k}\Phi_H^{k-\frac{m}{2}} \\ \leq \; & C'_k \Phi_H^k \qquad (A.41) \end{aligned}$$

where the second inequality comes from the fact that one can extract at least $\frac{m}{2}$ pairs (i_{l_p}, j_{l_p}) in which for each p either i_{l_p} or j_{l_p} appears uniquely once. Now, (A.41) and (3.17) are enough to state (A.39) for k large enough. The proof of (A.36) is very similar.

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n^{2k}}\Big(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} V_{i,j}\Big)^{2k}\mathbf{1}_{\overline{A}_n}\Big)\\
\leq\ & n^{-2k}\sum_{m=2}^{3k}\sum_{(i_l,j_l)_{1\leq l\leq 2k}\in J_m}\left|\mathbb{E}\Big(\Pi_{l=1}^{2k}\frac{(r(\chi_{i_l})-r(\chi_{j_l}))\epsilon_{i_l}\Delta_{i_l}(\chi_{j_l})w(\chi_{j_l})}{(n-1)\,MISE^*(H)\,\mathbb{E}\big(\Delta(\chi_{j_l})\big)\,r_{1H}^{-j_l}(\chi_{j_l})}\mathbf{1}_{\overline{A}_n}\Big)\right|, \qquad (A.42)
\end{aligned}$$

The last line comes from the fact that when $m > 3k$ at least one of the ϵ's appears with exponent 1 and hence the mean equals 0 (conditioning w.r.t. $\chi_1, \ldots, \chi_n$). Now conditions (3.23), (3.17) and(3.18), as well as Lemma (5) are used to get

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n^{2k}}\Big(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} V_{i,j}\Big)^{2k}1_{\overline{A}_n}\Big)\\
\leq\ & C'_k\sum_{m=2}^{3k} n^{m-2k}\Big(MISE^*(H)(n\Phi_H)^2\Big)^{-k}\Phi_H^{\frac{m}{2}}\\
\leq\ & C'_k\Phi_H^{\frac{k}{2}}\sum_{m=2}^{3k}(n\Phi_H)^{m-3k}\Phi_H^{\frac{3k-m}{2}}\\
\leq\ & C'_k\Phi_H^{\frac{k}{2}},
\end{aligned}$$

what is enough to get (A.36). □

Lemma 13. *Under conditions of Theorem 2, we have*

$$\sup_{H\in\mathcal{H}_n}\left|\frac{MISE(H)-MISE^*(H)}{MISE^*(H)}\right|\longrightarrow 0 \quad a.s$$

Proof. We consider the following decomposition of $MISE(H)$:

$$\begin{aligned}
MISE(H) &= \int\mathbb{E}\left(\frac{(\hat{r}_{2H}(\chi)-r(\chi)\hat{r}_{1H}(\chi))^2}{\hat{r}_{1H}(\chi)^2}\right)W(\chi)dP_\chi(\chi)\\
&= \int\mathbb{E}\left((\hat{r}_{2H}(\chi)-r(\chi)\hat{r}_{1H}(\chi))^2(1+(1-\hat{r}_{1H}(\chi))+\frac{(1-\hat{r}_{1H}(\chi))^2}{\hat{r}_{1H}(\chi)})^2\right)W(\chi)dP_\chi(\chi)\\
&= MISE^*(H)+A_1(H)+A_2(H)+2A_3(H)+2A_4(H)+2A_5(H),
\end{aligned}$$

with

$$\begin{aligned}
A_1(H) &= \int\mathbb{E}\Big((\hat{r}_{2H}(\chi)-r(\chi)\hat{r}_{1H}(\chi))^2(1-\hat{r}_{1H}(\chi))^2)\Big)W(\chi)dP_\chi(\chi)\\
A_2(H) &= \int\mathbb{E}\Big((\hat{r}(\chi)-r(\chi))^2(1-\hat{r}_{1H}(\chi))^4)\Big)W(\chi)dP_\chi(\chi)\\
A_3(H) &= \int\mathbb{E}\Big((\hat{r}_{2H}(\chi)-r(\chi)\hat{r}_{1H}(\chi))^2(1-\hat{r}_{1H}(\chi)))\Big)W(\chi)dP_\chi(\chi)\\
A_4(H) &= \int\mathbb{E}\Big((\hat{r}_{2H}(\chi)-r(\chi)\hat{r}_{1H}(\chi))(\hat{r}(\chi)-r(\chi))(1-\hat{r}_{1H}(\chi))^2\Big)W(\chi)dP_\chi(\chi)\\
A_5(H) &= \int\mathbb{E}\Big((\hat{r}_{2H}(\chi)-r(\chi)\hat{r}_{1H}(\chi))(\hat{r}(\chi)-r(\chi))(1-\hat{r}_{1H}(\chi))^3\Big)W(\chi)dP_\chi(\chi).
\end{aligned}$$

We consider the term $A_1(H)$. Long but simple calculations leads to the following decomposition:

$$A_1(H) = A_{11} + A_{12} + A_{13} + 2A_{14} + 2A_{15} + 2A_{16},$$

which holds because of Lemma 6, with

$$\begin{aligned}
A_{11} &= \int \mathbb{E}\Big((\hat{r}_{2H}(\chi) - E(\hat{r}_{2H}(\chi)))^2(1-\hat{r}_{1H}(\chi))^2\Big) W(\chi)dP_\chi(\chi) \\
A_{12} &= \int B^2(\chi)\mathbb{E}\Big((1-\hat{r}_{1H}(\chi))^2\Big) W(\chi)dP_\chi(\chi) \\
A_{13} &= \int r^2(\chi)\mathbb{E}\Big((1-\hat{r}_{1H}(\chi))^4\Big) W(\chi)dP_\chi(\chi) \\
A_{14} &= \int B(\chi)\mathbb{E}\Big((\hat{r}_{2H}(\chi) - \mathbb{E}(\hat{r}_{2H}(\chi)))(1-\hat{r}_{1H}(\chi))^2\Big) W(\chi)dP_\chi(\chi) \\
A_{15} &= \int r(\chi)\mathbb{E}\Big((\hat{r}_{2H}(\chi) - \mathbb{E}(\hat{r}_{2H}(\chi)))(1-\hat{r}_{1H}(\chi))^3\Big) W(\chi)dP_\chi(\chi) \\
A_{16} &= \int B(\chi)r(\chi)\mathbb{E}\Big((1-\hat{r}_{1H}(\chi))^3\Big) W(\chi)dP_\chi(\chi).
\end{aligned}$$

For dealing with term A_{11}, we first observe that

$$\begin{aligned}
&\mathbb{E}\Big(\big(\hat{r}_{2H}(\chi) - \mathbb{E}(\hat{r}_{2H}(\chi))\big)^2\big(1-\hat{r}_{1H}(\chi)\big)^2\Big) \\
=&\frac{1}{n^4\mathbb{E}^4\big(\Delta_1(\chi)\big)}\Bigg[\sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{E}\Big(\big(Y_i\Delta_i(\chi) - \mathbb{E}\big(Y_i\Delta_i(\chi)\big)\big)^2\Big)\mathbb{E}\Big(\big(\Delta_j(\chi) - \mathbb{E}\big(\Delta_j(\chi)\big)\big)^2\Big) \qquad \text{(A.43)} \\
&\quad + 2\sum_{\substack{i,,j=1\\ i\neq j=1}}^{n}\mathbb{E}\Big(\big(Y_i\Delta_i(\chi) - \mathbb{E}\big(Y_i\Delta_i(\chi)\big)\big)\big(\Delta_i(\chi) - \mathbb{E}\big(\Delta_i(\chi)\big)\big)\big(Y_j\Delta_j(\chi) - \mathbb{E}\big(Y_j\Delta_j(\chi)\big)\big)\big(\Delta_j(\chi) - \mathbb{E}\big(\Delta_j(\chi)\big)\big)\Big) \\
&\quad + \sum_{i=1}^{n}\mathbb{E}\Big(\big(Y_i\Delta_i(\chi) - \mathbb{E}\big(Y_i\Delta_i(\chi)\big)\big)^2\big(\Delta_i(\chi) - \mathbb{E}\big(\Delta_i(\chi)\big)\big)^2\Big)\Bigg] \\
=&\frac{1}{n^4\mathbb{E}^4\big(\Delta_1(\chi)\big)}\Big[n(n-1)Var\big(Y_1\Delta_1(\chi)\big)Var\big(\Delta_1(\chi)\big) \\
&\quad + 2n(n-1)\mathbb{E}^2\Big(\big(Y_1\Delta_1(\chi) - \mathbb{E}\big(Y_1\Delta_1(\chi)\big)\big)\big(\Delta_1(\chi) - \mathbb{E}\big(\Delta_1(\chi)\big)\big)\Big) \\
&\quad + n\mathbb{E}\Big(\big(Y_1\Delta_1(\chi) - \mathbb{E}\big(Y_1\Delta_1(\chi)\big)\big)^2\big(\Delta_1(\chi) - \mathbb{E}\big(\Delta_1(\chi)\big)\big)^2\Big)\Big],
\end{aligned}$$

which relies on condition (3.15). Because of condition (3.16), we have

$$Var\big(\Delta_1(\chi)\big) = \mathbb{E}\big(\Delta_1^2(\chi)\big) - \mathbb{E}^2\big(\Delta_1(\chi)\big) \leq \mathbb{E}\big(\Delta_1^2(\chi)\big) \leq C_{2,1}\Phi_H. \qquad \text{(A.44)}$$

Similarly, because of condition (3.26), we have

$$\begin{aligned}
Var\big(Y_1\Delta_1(\chi)\big) &= \mathbb{E}\big(Y_1^2\Delta_1^2(\chi)\big) - \mathbb{E}^2\big(Y_1\Delta_1(\chi)\big) \qquad \text{(A.45)} \\
&\leq \mathbb{E}\Big(Y_1^2\Delta_1^2(\chi)\Big) \leq C\Phi_H.
\end{aligned}$$

Because of conditions (3.16) and (3.26), and because $\Phi_H \leq 1$, we have

$$\begin{aligned}
&\mathbb{E}^2\Big(\big(Y_1\Delta_1(\chi) - \mathbb{E}\big(Y_1\Delta_1(\chi)\big)\big)\big(\Delta_1(\chi) - \mathbb{E}\big(\Delta_1(\chi)\big)\big)\Big) \\
= &\Big(2\mathbb{E}\big(Y_1\Delta_1(\chi)\big)\mathbb{E}\big(\Delta_1(\chi)\big) - 2\mathbb{E}\big(Y_1\Delta_1^2(\chi)\big)\Big)^2 \\
= &\ 4\Big(\mathbb{E}^2\big(Y_1\Delta_1(\chi)\big)\mathbb{E}^2\big(\Delta_1(\chi)\big) + \mathbb{E}^2\big(Y_1\Delta_1^2(\chi)\big) - 2\mathbb{E}\big(Y_1\Delta_1^2(\chi)\big)\mathbb{E}\big(Y_1\Delta_1(\chi)\big)\mathbb{E}\big(\Delta_1(\chi)\big)\Big) \\
\leq &\ C\big(\Phi_H^4 + \Phi_H^2 + \Phi_H^3\big) \leq C'\Phi_H^2.
\end{aligned} \tag{A.46}$$

Using the same conditions, we also have

$$\begin{aligned}
&\mathbb{E}\Big(\big(Y_1\Delta_1(\chi) - \mathbb{E}\big(Y_1\Delta_1(\chi)\big)\big)^2\big(\Delta_1(\chi) - \mathbb{E}\big(\Delta_1(\chi)\big)\big)^2\Big)\Big] \\
=&\mathbb{E}\big(Y_1^2\Delta_1^2(\chi)\big)\mathbb{E}^2\big(\Delta_1(\chi)\big) + \mathbb{E}\big(Y_1^2\Delta_1^4(\chi)\big) - 2\mathbb{E}\big(Y_1^2\Delta_1^3(\chi)\big)\mathbb{E}\big(\Delta_1(\chi)\big) + \mathbb{E}^4\big(Y_1\Delta_1(\chi)\big) + \mathbb{E}\big(\Delta_1^2(\chi)\big)\mathbb{E}^2\big(Y_1\Delta_1(\chi)\big) \\
&- 4\mathbb{E}^2\big(\Delta_1(\chi)\big)\mathbb{E}^2\big(Y_1\Delta_1(\chi)\big) - 2\mathbb{E}\big(Y_1\Delta_1^3(\chi)\big)\mathbb{E}\big(Y_1\Delta_1(\chi)\big) + 4\mathbb{E}\big(Y_1\Delta_1^2(\chi)\big)\mathbb{E}\big(\Delta_1(\chi)\big)\mathbb{E}\big(Y_1\Delta_1(\chi)\big) \\
\leq &C(\Phi_H + \Phi_H^2 + \Phi_H^3 + \Phi_H^4) \leq C'\Phi_H.
\end{aligned} \tag{A.47}$$

Now, given equations (A.43), (A.44), (A.45),(A.46) and (A.47), and using condition (3.16) and the fact that W is bounded, we obtain

$$\begin{aligned}
A_{11} &= \int \mathbb{E}\big((\hat{r}_{2H}(\chi) - E(\hat{r}_{2H}(\chi)))^2(1 - \hat{r}_{1H}(\chi))^2\big)\, W(\chi) dP_\chi(\chi) \\
&\leq \frac{1}{n^4\Phi_H^4}(n(n-1)C\Phi_H^2 + 2n(n-1)C'\Phi_H^2 + nC''\Phi_H) = O\Big(\frac{1}{n^2\Phi_H^2}\Big).
\end{aligned}$$

As a result, because of Lemma 5 and using condition (3.18), we have

$$\sup_{H\in\mathcal{H}_n} \frac{|A_{11}|}{MISE^*(H)} \leq \sup_{H\in\mathcal{H}_n} C\frac{n\Phi_H}{n^2\Phi_H^2} = \sup_{H\in\mathcal{H}_n} C\frac{1}{n\Phi_H} = O\Big(\frac{1}{\inf\limits_{H\in\mathcal{H}_n} n\Phi_H}\Big) = o(1).$$

We then consider the term A_{12}. Using condition (3.15) and equation (A.44), we have

$$\mathbb{E}\Big((1 - \hat{r}_{1H}(\chi))^2\Big) = \frac{1}{n^2\mathbb{E}^2\big(\Delta_1(\chi)\big)}\sum_{i=1}^{n}\mathbb{E}\Big(\mathbb{E}\big(\Delta_i(\chi)\big) - \Delta_i(\chi)\big)^2\Big) = \frac{nVar\big(\Delta_1(\chi)\big)}{n^2\mathbb{E}^2\big(\Delta_1(\chi)\big)} \leq \frac{C}{n\Phi_H}$$

Consequently, and because W is bounded,

$$A_{12} \leq \frac{C}{n\Phi_H}\int B^2(\chi)W(\chi)dP_\chi(\chi) \leq \frac{C'}{n\Phi_H}b_H,$$

so that

$$\sup_{H\in\mathcal{H}_n} \frac{|A_{12}|}{MISE^*(H)} \leq C\frac{b_H}{n\Phi_H}\frac{n}{(n-1)\,b_H} = O\Big(\frac{1}{\inf\limits_{H\in\mathcal{H}_n} n\Phi_H}\Big) = o(1),$$

because of condition (3.18), and because of Lemma 5. In view of dealing with term A_{13}, and because of condition

(3.15), we note that

$$\begin{aligned}
&\mathbb{E}\Big(1-\hat{r}_{1H}(\chi))^4\Big) \qquad (A.48)\\
=&\frac{1}{n^4\mathbb{E}^4\Big(\Delta_1(\chi)\Big)}\Bigg(3\sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{E}\Bigg(\Big(\mathbb{E}\Big(\Delta_i(\chi)\Big)-\Delta_i(\chi)\Big)^2\Big(\mathbb{E}\Big(\Delta_j(\chi)\Big)-\Delta_j(\chi)\Big)^2\Bigg)+\sum_{i=1}^{n}\mathbb{E}\Bigg(\Big(\mathbb{E}\Big(\Delta_i(\chi)\Big)-\Delta_i(\chi)\Big)\Bigg)\\
=&\frac{3n(n-1)Var^2\Big(\Delta_1(\chi)\Big)}{n^4\mathbb{E}^4\Big(\Delta_1(\chi)\Big)}+\frac{1}{n^3\mathbb{E}^4\Big(\Delta_1(\chi)\Big)}\mathbb{E}\Bigg(\Big(\mathbb{E}\Big(\Delta_1(\chi)\Big)-\Delta_1(\chi)\Big)^4\Bigg).
\end{aligned}$$

Because of condition (3.16), and because $\Phi_H < 1$, we have

$$\begin{aligned}
&\mathbb{E}\Bigg(\Big(\mathbb{E}\Big(\Delta_1(\chi)\Big)-\Delta_1(\chi)\Big)\Big)^4\Bigg)\\
&=\mathbb{E}\Big(\Delta_1^4(\chi)\Big)-4\mathbb{E}\Big(\Delta_1(\chi)\Big)\mathbb{E}\Big(\Delta_1^3(\chi)\Big)+6\mathbb{E}^2\Big(\Delta_1(\chi)\Big)\mathbb{E}\Big(\Delta_1^2(\chi)\Big)-4\mathbb{E}^4\Big(\Delta_1(\chi)\Big)+\mathbb{E}^4\Big(\Delta_1(\chi)\Big)\\
&\leq C(\Big(\Phi+\Phi_H^2+\Phi_H^3+\Phi_H^4\Big)\leq C'\Phi_H.
\end{aligned}$$

Moreover, due to equation (A.44), condition (3.16), and because $\Phi_H < 1$, we have

$$\mathbb{E}\Big(1-\hat{r}_{1H}(\chi))^4\Big)\leq C\frac{1}{n^4\Phi_H^4}n^2\Phi_H^2+C'\frac{1}{n^3\Phi_H^4}\Phi_H\leq C\frac{1}{n^2\Phi_H^2}$$

Thus, as r and W are bounded, we obtain

$$A_{13}\leq\frac{C}{n^2\Phi_H^2}\int r^2(\chi)W(\chi)dP_\chi(\chi)\leq\frac{C'}{n^2\Phi_H^2}$$

Consequently, using (3.18) and Lemma 5, we have

$$\sup_{H\in\mathcal{H}_n}\frac{|A_{13}|}{MISE^*(H)}\quad\leq\sup_{H\in\mathcal{H}_n}\quad C\Bigg(\frac{1}{n^2\Phi_H^2}+\frac{1}{n^3\Phi_H^3}\Bigg)Cn\phi_H\leq O\Big(\frac{1}{\inf\limits_{H\in\mathcal{H}_n}n\Phi_H}\Big)=o(1).$$

Now, using Cauchy-Schwarz inequality, we show that

$$|A_{14}|\leq\sqrt{|A_{11}|}\sqrt{|A_{12}|},\quad |A_{15}|\leq\sqrt{|A_{11}|}\sqrt{|A_{13}|}\quad\text{and } |A_{16}|\leq\sqrt{|A_{12}|}\sqrt{|A_{13}|}.$$

It results from this that

$$\sup_{H\in\mathcal{H}_n}\frac{|A_{14}|}{MISE^*(H)}=o(1),\quad\sup_{H\in\mathcal{H}_n}\frac{|A_{15}|}{MISE^*(H)}=o(1)\quad\text{and}\quad\sup_{H\in\mathcal{H}_n}\frac{|A_{16}|}{MISE^*(H)}=o(1).$$

Thus, $\sup\limits_{H\in\mathcal{H}_n}\frac{|A_1|}{MISE^*(H)}=o(1)$. In a view to deal with term A_2, we observe that

$$\mathbb{E}\Big((\hat{r}(\chi)-r(\chi))^2|\chi_1\ldots\chi_n\Big)\leq 2\Big(\mathbb{E}(\hat{r}^2(\chi)|\chi_1\ldots\chi_n)+r^2(\chi)\Big)\leq C,$$

the last inequality deriving from conditions (3.25) and (3.16), and from the fact that r is bounded. Consequently, we have

$$A_2\leq C\int\mathbb{E}\Big((1-\hat{r}_{1H}(\chi))^4\Big)W(\chi)dP_\chi(\chi),$$

where the integral is dealt with in the same way as for term A_{13}, so that we obtain $\sup\limits_{H\in\mathcal{H}_n}\frac{|A_2|}{MISE^*(H)}=o(1)$. Finally, using Cauchy-Schwarz inequality, we observe that

$$\frac{|A_3|}{2}\leq\sqrt{MISE^*(H)}\sqrt{|A_1|},\quad\frac{|A_4|}{2}\leq\sqrt{MISE^*(H)}\sqrt{|A_2|}\quad\text{and }\frac{|A_5|}{2}\leq\sqrt{|A_1|}\sqrt{|A_2|}.$$

It results from this that

$$\sup_{H\in\mathcal{H}_n}\frac{|A_3|}{MISE^*(H)}=o(1),\quad \sup_{H\in\mathcal{H}_n}\frac{|A_4|}{MISE^*(H)}=o(1)\quad\text{and}\quad \sup_{H\in\mathcal{H}_n}\frac{|A_5|}{MISE^*(H)}=o(1)..$$

The proof of our Lemma is then completed. □

Lemma 14. *Under conditions of Theorem 2, we have*

$$\frac{MISE^*(H^{CV})}{MISE^*(H^*)}\longrightarrow 1\quad a.s.$$

Proof. The proof of this Lemma is completed as soon as we can prove that

$$\left|\frac{MISE^*(H^{CV})-MISE^*(H^*)}{MISE^*(H^*)}\right|\longrightarrow 0\quad a.s.$$

In order to show this convergence, we make use of the following upper bound:

$$\begin{aligned}&\left|MISE^*(H^{CV})-MISE^*(H^*)\right| \\ &\leq \left|MISE^*(H^{CV})-ASE(H^{CV})\right|+\left|ASE(H^{CV})-ASE(H^n)\right| \\ &+\left|ASE(H^n)-ASE(H^*)\right|+\left|ASE(H^*)-MISE^*(H^*)\right|,\end{aligned}\tag{A.49}$$

with $H^n=\arg\min_{H\in\mathcal{H}_n}ASE(H)$. Given (A.49), we can further major $|MISE^*(H^{CV})-MISE^*(H^*)|$ using the inequality

$$\begin{aligned}&\left|ASE(H^{CV})-ASE(H^n)\right| \\ &\leq \left|ASE(H^{CV})-ASE(H^n)-CV(H^{CV})+CV(H^n)\right| \\ &\leq \left|\widetilde{ASE}(H^{CV})-\widetilde{ASE}(H^n)-CV(H^{CV})+CV(H^n)\right| \\ &+\left|\widetilde{ASE}(H^{CV})-ASE(H^{CV})\right|+\left|\widetilde{ASE}(H^n)-ASE(H^n)\right| \\ &\leq 2\left|CT(H^{CV})\right|+2\left|CT(H^n)\right|+\left|\widetilde{ASE}(H^{CV})-ASE(H^{CV})\right|+\left|\widetilde{ASE}(H^n)-ASE(H^n)\right|\end{aligned}\tag{A.50}$$

which is valid because $CV(H^n)\geq CV(H^{CV})$ and $ASE(H^n)\leq ASE(H^{CV})$, by construction, and because of the decomposition

$$CV(H)=\widetilde{ASE}(H)-2CT(H)+R,$$

with $R=\frac{1}{n}\sum_{j=1}^{n}(Y_j-r(\chi_j))^2W(\chi_j)$. Similarly, because $MISE^*(H^*)\leq MISE^*(H^n)$, one gets:

$$\begin{aligned}\left|ASE(H^n)-ASE(H^*)\right|&\leq\left|ASE(H^n)-ASE(H^*)-MISE^*(H^n)+MISE^*(H^*)\right| \\ &\leq\left|ASE(H^n)-MISE^*(H^n)\right|+\left|ASE(H^*)-MISE^*(H^*)\right|\end{aligned}\tag{A.51}$$

Then, inequalities (A.49), (A.50) and (A.51) lead to

$$\begin{aligned}
&\left|\frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)}\right| \\
\leq& 2\left|\frac{ASE(H^*) - MISE^*(H^*)}{MISE^*(H^*)}\right| \\
&+ \frac{MISE^*(H^{CV})}{MISE^*(H^*)}\left(\left|\frac{MISE^*(H^{CV}) - ASE(H^{CV})}{MISE^*(H^{CV})}\right| + \left|\frac{\widetilde{ASE}(H^{CV}) - ASE(H^{CV})}{MISE^*(H^{CV})}\right| + 2\left|\frac{CT(H^{CV})}{MISE^*(H^{CV})}\right|\right) \\
&+ \frac{MISE^*(H^n)}{MISE^*(H^*)}\left(\left|\frac{MISE^*(H^n) - ASE(H^n)}{MISE^*(H^n)}\right| + \left|\frac{\widetilde{ASE}(H^n) - ASE(H^n)}{MISE^*(H^n)}\right| + 2\left|\frac{CT(H^n)}{MISE^*(H^n)}\right|\right).
\end{aligned}$$

We also have

$$\begin{aligned}
&\frac{MISE^*(H^{CV})}{MISE^*(H^*)} = \frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)} + 1, \\
&\frac{MISE^*(H^n)}{MISE^*(H^*)} \leq \frac{MISE^*(H^n)}{ASE(H^n)}\frac{ASE(H^*)}{MISE^*(H^*)} = \frac{1}{1 + \frac{|ASE(H^n) - MISE^*(H^n)|}{MISE^*(H^n)}}\left(\frac{|ASE(H^*) - MISE^*(H^*)|}{MISE^*(H^*)} + 1\right),
\end{aligned}$$

where the second inequality holds because $ASE(H^n) \leq ASE(H^{CV})$. Combining all those results, and using notations (3.29), we finally get

$$\left|\frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)}\right|\Big(1 - T_\alpha - T_\beta - T_\gamma\Big) \leq 2T_\alpha + \Big(T_\alpha + T_\beta + T_\gamma\Big)\frac{2}{1 - T_\alpha}.$$

Consequently, the proof of Lemma 14 is completed because of Lemmas 10, 11 and 12 which ensure the convergence to zero of T_α, T_β, and T_γ respectively. . □

Given those Lemmas, the proof of Theorem 2 is as follows. We consider the following decomposition, which holds because of Lemma 13:

$$\begin{aligned}
&\left|\frac{MISE(H^{CV})}{MISE(H^*)} - 1\right| = \left|\frac{MISE(H^{CV}) - MISE(H^*)}{MISE^*(H^*)}\right|\frac{MISE^*(H^*)}{MISE(H^*)} \\
&\leq \frac{MISE^*(H^*)}{MISE(H^*)}\left(\frac{|MISE(H^{CV}) - MISE^*(H^{CV})|}{MISE^*(H^*)} + \frac{|MISE^*(H^{CV}) - MISE^*(H^*)|}{MISE^*(H^*)}\right. \\
&\qquad\qquad\left. + \frac{|MISE(H^*) - MISE^*(H^*)|}{MISE^*(H^*)}\right) \\
&\leq \left(\frac{1}{1 - \sup_{H\in\mathcal{H}_n}\frac{|MISE(H) - MISE^*(H)|}{MISE^*(H)}}\right)\left(\frac{MISE^*(H^{CV})}{MISE^*(H^*)}\sup_{H\in\mathcal{H}_n}\frac{|MISE(H^{CV}) - MISE^*(H^{CV})|}{MISE^*(H^{CV})}\right. \\
&\qquad\qquad\left. + \frac{|MISE^*(H^{CV}) - MISE^*(H^*)|}{MISE^*(H^*)} + \sup_{H\in\mathcal{H}_n}\frac{|MISE(H^*) - MISE^*(H^*)|}{MISE^*(H^*)}\right)
\end{aligned}$$

This tends to zero almost surely because of Lemmas 13 and 14. The proof of Theorem 2 is then completed.

References

[1] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer Series in Statistics, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] C. Timmermans, R. von Sachs, Bagidis, a new method for statistical analysis of differences between curves with sharp local patterns, under revision (2010).

[3] J. Ramsay, B. Silverman, Applied functional data analysis : methods and case studies, Springer Series in Statistics, Springer-Verlag, New York, 2002.

[4] J. Ramsay, B. Silverman, Functional Data Analysis, Springer Series in Statistics, Springer, 2nd edition, 2005.
[5] D. Bosq, Linear Processes in Function Spaces, Springer, Berlin, 2000.
[6] F. Ferraty, Y. Romain, The Oxford Handbook of Functional Data Analysis, Oxford University Press, USA, City, 2010.
[7] C. J. Stone, Optimal global rates of convergence for nonparametric regression, The Annals of Statistics 10 (1982) 1040–1053.
[8] C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, Annals of Statistics (2008).
[9] J. Ramsay, C. Dalzell, Some tools for functional data analysis (with discussion), Journal of the Royal Statistical Society, Series B 53 (1991) 539–572.
[10] T. T. Cai, P. Hall, Prediction in functional linear regression, Annals of Statistics 34 (2006) 2159.
[11] A. Sood, G. M. James, G. J. Tellis, Functional regression: A new model for predicting market penetration of new products, Marketing Science 28 (2009) 36–51.
[12] A. Ait-Saidi, F. Ferraty, R. Kassa, P. Vieu, Cross-validated estimations in the single-functional index model, Statistics 42 (2008) 475–494.
[13] G. Aneiros-Pérez, P. Vieu, Nonparametric time series prediction: A semi-functional partial linear modeling, Journal of Multivariate Analysis 99 (2008) 834–857.
[14] F. Ferraty, A. Rabhi, P. Vieu, Conditional Quantiles for Dependent Functional Data with Application to the Climatic El Nino Phenomenon , Sankhya : The Indian Journal of Statistics 67 (2005) 378–398. Special Issue on Quantile Regression and Related Methods.
[15] G. Aneiros-Pérez, H. Cardot, G. Estévez-Pérez, P. Vieu, Maximum ozone concentration forecasting by functional non-parametric approaches, Environmetrics 15 (2004) 675–685.
[16] M. Girardi, W. Sweldens, A new class of unbalanced Haar wavelets that form an unconditional basis for Lp on general measure spaces, Journal of Fourier Analysis and Applications 3 (1997) 457–474.
[17] P. Fryzlewicz, Unbalanced Haar technique for non parametric function estimation, Journal of the American Statistical Association 102 (2007) 1318–1327.
[18] F. Ferraty, A. Laksaci, A. Tadj, P. Vieu, Rate of uniform consistency for nonparametric estimates with functional variables, Journal of Statistical Planning and Inference 140 (2010) 335–352.
[19] F. Ferraty, A. Laksaci, A. Tadj, P. Vieu, Kernel regression with functional response, Electronic Journal of Statistics 5 (2011) 159–171.
[20] M. Rachdi, P. Vieu, Nonparametric regression for functional data: Automatic smoothing parameter selection, Journal of Statistical Planning and Inference 137 (2007) 2784–2801.
[21] J. Marron, W. Hardle, Random approximations to some measures of accuracy in nonparametric curve estimation, Journal of Multivariate Analysis 20 (1986) 91–113.
[22] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
[23] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[24] F. Burba, F. Ferraty, P. Vieu, k-Nearest Neighbour method in functional nonparametric regression, Journal of Nonparametric Statistics 21 (2009) 453–469.
[25] A. Van der Vaart, J. Wellner, Weak convergence and empirical processes, Springer Series in Statistics, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.