

Dimension Reduction in Functional Regression with Applications

U. Amato^a A. Antoniadis^b I. De Feis^a

^a*Istituto per le Applicazioni del Calcolo 'M. Picone' CNR - Sezione di Napoli, Via Pietro Castellino 111, 80131 Napoli, Italy*

^b*Laboratoire de Modelisation et Calcul (LMC-IMAG), Universite Joseph Fourier, Tour IRMA, B.P. 53, 38041 Grenoble, CEDEX 9, France*

Abstract

Two dimensional reduction regression methods to predict a scalar response from a discretized sample path of a continuous time covariate process are presented. The methods take into account the functional nature of the predictor and are both based on appropriate wavelet decompositions. Using such decompositions, we derive prediction methods that are similar to minimum average variance estimation (MAVE) or functional sliced inverse regression (FSIR). We describe their practical implementation and we apply the method both in simulation and on real data analyzing three calibration examples of near infrared spectra.

Key words: dimension reduction; wavelets; MAVE; SIR

1 Introduction

In regression or classification problems one of the tasks is to study the structural relationship between a response variable Y and a vector of covariates $\mathbf{X} = (X_1, \dots, X_p)^T$ via $m(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ where $\mathbf{x} = (x_1, \dots, x_p)^T$ and $m(\mathbf{x}) = m(x_1, \dots, x_p)$. When the number of predictors is moderate and the number of samples that are observed are large, nonparametric statistical techniques are very useful for investigating $\mathbb{E}(Y|\mathbf{X})$, especially when an adequate parsimoniously parameterized model is not available. Quite often however, especially in longitudinal studies, for a number n of subjects one observes

Email addresses: u.amato@iac.cnr.it (U. Amato),
Anestis.Antoniadis@imag.fr (A. Antoniadis), i.defeis@iac.cnr.it (I. De Feis).

measurements of a scalar response variable Y and of an explanatory variable that is a continuous time process observed in a closed interval of \mathbb{R} . In such a case the vector \mathbf{x} represents discretized values of a function or of a stochastic process and this casts the regression problem into the class of functional data analysis.

In functional regression problems, one has a response variable Y to be predicted by a set of variables X_1, \dots, X_p that are discretizations of a same curve X at points t_1, \dots, t_p , that is $X_j = X(t_j)$, $j = 1, \dots, p$ where the discretization time points t_j lie in $[0, 1]$ without loss of generality. A typical set-up includes near infra-red spectroscopy (Y represents the proportion of a chemical constituent and \mathbf{x} is the spectrum of a sample, discretized at a sequence of wavelengths). Nonparametric techniques have been proved useful for analyzing such functional data. These techniques include growth curve modelling (Ramsay, 1982; Kneip and Gasser, 1992), principal component analysis (Besse and Ramsay, 1986); partial least squares (Goutis and Fearn, 1996); functional linear regression (Cardot et al., 1999; Ferraty and Vieu, 2002). They differ from classical longitudinal problems in that the covariance is unstructured and there are generally more measurements per subject than there are subjects. If classical techniques were used, these differences would generally produce singularity problems. Functional techniques overcome these problems. For an extensive review of such functional data approaches the reader is referred to the excellent monograph by Ramsay and Silvermann (1997).

Although surface smoothing techniques for estimating the regression function appear to be feasible, the increasing number of explanatory variables makes nonparametric estimation suffer from the curse of dimensionality. In practice, before applying any nonparametric regression technique to model real data, and in order to avoid the curse of dimensionality, a dimension reduction or model selection technique is crucial for appropriate smoothing. A first approach is to find a functional basis, decompose the covariate curve $X(t)$ accordingly and work with the coefficients in a spirit similar to that adopted by Martens and Naes (1989, Chapter 3), Alsberg (1993) and Denham and Brown (1993). The aim is to explain or predict the response through an expansion of the explanatory process in a relatively low dimensional basis in the space spanned by the measurements of X , thus revealing relationships that may not be otherwise apparent. Dimension reduction without loss of information is a dominant theme in such cases: one tries to reduce the dimension of X without losing information on $Y|X$, and without requiring a model for $Y|X$. Borrowing terminology from classical statistics, Cook (2000) calls this a sufficient dimension reduction which leads to the pursuit of sufficient summaries containing all of the information on $Y|X$ that is available from the sample.

In a first part of this paper we describe a novel regression approach for regression problems with high dimensional X variables. Our methodology relies

upon the above notion of sufficient dimension reduction and is based on developments in a recent paper by Xia et al. (2002) where an adaptive approach for effective dimension reduction called the (conditional) minimum average variance estimation (MAVE) method, is proposed within quite a general setting. In particular we show that the MAVE method is applicable to a wide range of functional data models, by expanding the covariate process into a wavelet basis and by estimating consistently the dimension of the most informative covariate space (EDR space). The idea is to transform each observed sample path of the covariate process X into a set of wavelet coefficients, the whole of which would suffice to reconstruct the sample paths, and select good predictors from amongst these. Let us point out that there are some good reasons to consider wavelet bases rather than other bases. Indeed, we will first show that wavelets can be used successfully for compression of a stochastic process, in the sense that the sample paths can be accurately reconstructed from a fraction of the full set of wavelet coefficients. Further, the wavelet decomposition of the sample paths will be a local one, so that if the information relevant to our prediction problem is contained in a particular part or parts of the sample path, as typically it is, this information will be carried by a very small number of wavelet coefficients. Moreover, the ability of wavelets to model the curve at different levels of resolution means that we have the option of selecting from the paths at a range of bandwidths. Also, for the application we have in mind, that is the analysis of infrared spectral data, detail coefficients typically contain important information concerning chemical composition, information which is usually contained in extreme features of higher derivatives of the spectra. Finally, note that while the above approach is inspired from a similar problem studied by Brown et al. (2001) and motivated by calibration problems in near infrared spectroscopy, it differs considerably from theirs since we are not using any bayesian variable selection method to find subsets of the wavelet coefficients with overall better predictive performance.

Another approach that is powerful for dimension reduction and for searching the EDR space is the sliced inverse regression (SIR) method proposed by Li (1991). In classical SIR the explanatory variable \mathbf{X} belongs to a finite dimensional space, say \mathbb{R}^p . Informally, SIR provides a basis $\{\beta_1, \dots, \beta_q\}$ of a subspace of \mathbb{R}^p and corresponding SIR predictors $\{b_1(\mathbf{X}), \dots, b_q(\mathbf{X})\}$ which are ordered according to their likely importance to the prediction or regression. Thus, the first SIR predictor $b_1(\mathbf{X})$ is likely more important than the second $b_2(\mathbf{X})$, and so on. Plots of Y versus various combinations of the SIR predictors can provide useful information on the regression. The method is essentially based on the eigenvalue decomposition of the matrix $(\text{var}(\mathbf{X}))^{-1}\text{var}(\mathbb{E}(\mathbf{X}|Y))$ and it has motivated a large amount of work by considering various estimates of $\text{var}(\mathbb{E}(\mathbf{X}|Y))$ (e.g. Hsing and Carroll, 1992; Zhu and Fang, 1996). SIR has been recently extended to a functional data setup by Dauxois et al. (2001), when the response Y is real and the covariate X is a random variable with values in a general Hilbert space, say H . In particular, these authors show that

under conditions similar to the ones required in classical SIR, the procedure to derive estimates of the EDR space still remains valid from a theoretical point of view. In Ferre and Yao (2000) a practical estimation method based on these results is derived, based on a truncated regularized estimate of the covariance operator of the X -process, similar to the one presented in Bosq (1991) for ARH(1) models and applied to the functional linear model by Cardot et al. (1999). Basically, it consists in projecting first the infinite dimensional data set onto a sequence of finite dimensional subspaces derived from the spectral decomposition of the operator $\text{var}(X)$. However, as noted in Ferre and Yao (2000), this procedure is often numerically instable unless the truncation is relatively severe.

In a second part of this paper, and in order to compare the MAVE method with the SIR approach we tackle the extension of the inverse regression method to the functional case differently using a spectral decomposition of a finite rank estimate of the operator $\text{var}(\mathbb{E}(X|Y))$. The estimate is based on a binned wavelet smoother of the regression $\mathbb{E}(X|Y)$ replacing the simple non-smooth nonparametric estimates of the inverse regression curves that classical SIR procedures are using, which may miss important relevant information as the continuous nature of the data is ignored.

However, one must note that the SIR method (classical or functional) imposes some strong probabilistic structure on X . Specifically, the method requires that basically X is within the class of elliptical symmetric distributions. Now, in time series analysis elliptical symmetry of X implies time-reversibility, a feature which is the exception rather than the rule in time series analysis.

The rest of this paper is organized as follows. Section 2 recalls some background on wavelets and wavelet decompositions that are going to be used in the sequel. In Section 3, we describe both the MAVE method and a procedure for determining the effective dimension reduction by means of cross-validation and demonstrate their applicability through a wavelet decomposition of the covariate process. Section 4 discusses the SIR extension that we have adopted and its comparison with other functional SIR existing methods, We also propose an extension of the sequential chi-square test procedure for estimating the dimension, originally proposed by Li (1991) for traditional SIR. To check and to compare our approaches, we have conducted many simulations, typical ones of which are reported in Section 5. The methodology is illustrated through an analysis of several spectrometric data sets in Section 6. Finally, theoretical proofs of our results are given in the appendix.

2 Wavelet-based orthogonal expansion of stochastic processes

The discrete wavelet transform (DWT) as formulated by Mallat (1989) and Daubechies (1992) is an increasingly popular tool for the statistical analysis of time series (e.g. Ogden, 1997; Carmona et al., 1998; Nason and von Sachs, 1999; Percival and Walden, 2000) and references therein). The DWT maps a time series into a set of wavelet coefficients. Each coefficient is associated with a particular scale, which is a measure of the amount of data that effectively determines the coefficient. Two distinct wavelet coefficients can be either ‘within-scale’ (i.e., both are associated with the same scale) or ‘between-scale’ (i.e., each has a distinct scale).

One reason for the popularity of the DWT in times series analysis is that measured data from most processes are inherently multiscale in nature owing to contributions from events occurring at different locations and with different localization in time and frequency. Consequently, data analysis and modelling methods that represent the measured variables at multiple scales are better suited for extracting information from measured data than methods that represent the variables at a single scale.

This section presents an overview of multiscale data analysis methods for stochastic processes based on wavelet decompositions. Since these methods exploit the ability of wavelets to extract events at different scales, compress deterministic features in a small number of relatively large coefficients, and approximately decorrelate a variety of stochastic processes, they will be used as our main tool in subsequent sections for decomposing the sample paths of the explanatory covariate process in our functional regression model.

Let (Ω, \mathcal{F}, P) denote a complete probability space and let $s(t)$ be a mean-square continuous process defined on $[0, 1]$, i.e.

$$s \in L^2(\Omega \times [0, 1]) = \left\{ X(t) : \Omega \rightarrow \mathbb{R}, t \in [0, 1] \mid \mathbb{E} \int_0^1 X^2(t) dt < \infty \right\}.$$

Recall that (e.g. Neveu, 1968) $L^2(\Omega \times [0, 1])$ is a separable Hilbert space with inner product defined by

$$\langle X, S \rangle = \mathbb{E} \int_0^1 X(t)S(t)dt.$$

To develop a wavelet decomposition analysis we mimick the construction of a multiresolution analysis of $L^2([0, 1])$ (Mallat, 1989). We will first consider a wavelet basis of $L^2([0, 1])$. We recall that $L^2([0, 1])$ is approximated by a multiresolution analysis, i.e. a ladder of closed subspaces

$$V_{j_0} \subset V_{j_0+1} \subset \cdots \subset L^2([0, 1])$$

with $j_0 \geq 0$, whose union is dense in L^2 and where V_j is spanned by 2^j orthonormal scaling functions $\phi_{j,k}$, $k = 0, \dots, 2^j - 1$, such that $\text{supp}(\phi_{j,k}) \subset [2^{-j}(k-c), 2^{-j}(k+c)]$ with c a constant not depending on j . At each resolution level j , the orthonormal complement W_j between V_j and V_{j+1} is generated by 2^j orthonormal wavelets $\psi_{j,k}$, $k = 0, \dots, 2^j - 1$, such that $\text{supp}(\psi_{j,k}) \subset [2^{-j}(k-c), 2^{-j}(k+c)]$. As a consequence, the family

$$\cup_{j \geq j_0} \{\psi_{j,k}\}_{k=0, \dots, 2^j - 1}$$

completed by $\{\phi_{j_0,k}\}_{k=0, \dots, 2^{j_0} - 1}$ constitutes an orthonormal basis of $L^2([0, 1])$. Similarly, we will define a sequence of approximating spaces of $L^2(\Omega \times [0, 1])$ by

$$V_j(\Omega \times [0, 1]) = \left\{ X \in L^2(\Omega \times [0, 1]) \mid X(t) = \sum_{\ell=0}^{2^j-1} \xi_{j,\ell} \phi_{j,\ell}(t), \sum_{\ell=0}^{2^j-1} \mathbb{E}(\xi_{j,\ell})^2 < \infty \right\},$$

where $\{\xi_{j,\ell}\}_{\ell=0, \dots, 2^j-1}$ is a sequence of random variables and $\phi_{j,\ell}$ is the scaling basis of V_j . Note that since $L^2(\Omega \times [0, 1])$ is isomorphic to the Hilbert tensor product $L^2(\Omega) \otimes L^2([0, 1])$, the stochastic approximating spaces $V_j(\Omega \times [0, 1])$ are closed subspaces of $L^2(\Omega \times [0, 1])$. Note also that for every $X \in V_j(\Omega \times [0, 1])$, one has $\mathbb{E}(X) \in L^2([0, 1])$ since

$$\int_0^1 [\mathbb{E}(X(t))]^2 dt = \int_0^1 \left(\mathbb{E} \sum_{k=0}^{2^j-1} \xi_{j,k} \phi_{j,k}(t) \right)^2 dt = \sum_{k=0}^{2^j-1} [\mathbb{E}(\xi_{j,k})]^2 \leq \sum_{k=0}^{2^j-1} \mathbb{E}(\xi_{j,k})^2,$$

by orthonormality of the scaling functions. Moreover, similar calculations show that, when $X \in V_j(\Omega \times [0, 1])$

$$\text{cov}(X(s), X(t)) = \sum_{k=0}^{2^j-1} \sum_{\ell=0}^{2^j-1} \text{cov}(\xi_{j,k}, \xi_{j,\ell}) \phi_{j,k}(s) \phi_{j,\ell}(t),$$

which belongs to $L^2([0, 1] \times [0, 1])$. In fact, the above calculations show that if $X \in L^2(\Omega \times [0, 1])$ and has a continuous covariance kernel R then the following two assertions are equivalent

$$X \in V_j(\Omega \times [0, 1]) \leftrightarrow R(\cdot, \cdot) \in V_j \otimes V_j$$

and

$$R(s, t) = \sum_{k,\ell} \gamma_{(j,k),(j,\ell)} (\phi_{j,k} \otimes \phi_{j,\ell})(s, t),$$

where γ is a positive definite symmetric matrix.

Following Cohen and D'Ales (1997) (see also Cheng and Tong, 1996), it is easy to see that $\{V_j(\Omega \times [0, 1]), j \in \mathbb{N}_0\}$ is a multiresolution analysis of $L^2(\Omega \times [0, 1])$. Moreover if $W_j(\Omega \times [0, 1])$ denotes the orthonormal complement of $V_j(\Omega \times [0, 1])$

in $V_{j+1}(\Omega \times [0, 1])$, then one naturally has the following stochastic wavelet decomposition:

$$X \in L^2(\Omega \times [0, 1]) \leftrightarrow X(t) = \sum_{k=0}^{2^{j_0}-1} \xi_{j_0,k} \phi_{j_0,k}(t) + \sum_{j \geq j_0} \sum_{\ell=0}^{2^j-1} \eta_{j,\ell} \psi_{j,\ell}(t),$$

where $\xi_{j,k} = \int_0^1 \phi_{j,k}(t) X(t) dt$ and $\eta_{j,k} = \int_0^1 \psi_{j,k}(t) X(t) dt$. The above remarks clearly show that the wavelet decomposition is a fundamental tool for viewing the stochastic process in both time and scale domains. Moreover, from an approximation perspective, if $R(t, t)$ is α -regular for some $\alpha > 0$, and if one uses regular enough wavelets, one may use Theorem 2.1. of Cohen and D'Ales (1997), to approximate in L^2 any sample path of the process X by its projection onto V_J at a rate of the order $\mathcal{O}(2^{-\alpha J})$, which is in fact a simple rephrasing, in the stochastic framework, of the deterministic results on the multiresolution approximation of functions in Sobolev spaces when the error is measured in the L^2 -norm. This result has the advantage that dimension reduction by basis truncation in the wavelet domain will be controlled more precisely.

3 A dimension-reduction functional regression model

In this section we describe our application of the MAVE method via a wavelet decomposition of the covariate process. We suppose that each realization $X(t)$ of the covariate process will be modelled as $X(t) = f(t) + s(t)$, $t \in [0, 1]$, where $f(t)$ represents the mean at time t and $s(t)$ is the observed residual variation, which will be regarded as a realization of a second order weakly stationary process. Since a large number of signal compression algorithms are based on second order statistical information we will concentrate on covariance modelling, and the mean function will be removed by filtering or simple averaging. Thus, we assume hereafter that the covariate process has been centered, so that $E(X(t)) = 0$ for all t .

Let us then consider the following model

$$Y = g(\langle \beta_1, X \rangle, \dots, \langle \beta_K, X \rangle) + \varepsilon \quad (1)$$

where ε is a scalar random variable independent of $X(t)$, $\{\beta_s(t), s = 1, \dots, K\}$ are K orthonormal functions belonging to $L^2([0, 1])$, and g is a smooth link function of \mathbb{R}^K into \mathbb{R} . For $D = K$, we obtain a standard regression model with all explanatory variables $\langle \beta_s, X \rangle$, $s = 1, \dots, K$ entering independently. Provided that $D < K$, the regression function depends on X only through D linear functionals of the explanatory process X . Hence, to explain the dependent variable Y , the space of K explanatory variables can be reduced to a

space with a smaller dimension D . The dimension reduction methods aim to find the dimension D of the reduction space and a basis defining this space.

Given a multiresolution analysis of $L^2([0, 1])$ and a primary level j_0 , as seen in Section 2, both $X(t)$ and $\beta_s(t)$ can be decomposed as

$$\begin{aligned} X(t) &= \sum_{k=0}^{2^{j_0}-1} \xi_{j_0,k} \phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \eta_{j,k} \psi_{j,k} \\ \beta_s(t) &= \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}^s \phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} d_{j,k}^s \psi_{j,k} \end{aligned}$$

with

$$\begin{aligned} \xi_{j_0,k} &= \langle X(t), \phi_{j_0,k} \rangle & \text{and} & & \eta_{j,k} &= \langle X(t), \psi_{j,k} \rangle \\ c_{j_0,k}^s &= \langle \beta_s(t), \phi_{j_0,k} \rangle & \text{and} & & d_{j,k}^s &= \langle \beta_s(t), \psi_{j,k} \rangle, \end{aligned}$$

$s = 1, \dots, K$ and $\{\xi_{j_0,k}, \eta_{j,k}\}_{j,k}$ sequences of random variables. By the isometry between $L^2([0, 1])$ and $\ell^2(\mathbb{R})$, model (1) can be also written as

$$Y = g(\langle \beta_1, \gamma \rangle, \dots, \langle \beta_K, \gamma \rangle) + \varepsilon. \quad (2)$$

We have indicated by β_s the ℓ^2 -sequence formed by the wavelet and scaling coefficients of $\beta_s(t)$, $s = 1, \dots, K$; and by γ the ℓ^2 -sequence formed by the wavelet and scaling coefficients of $X(t)$.

Usually, the sample paths of the process X are discretized. If we observe $p = 2^J$ values of $X(t)$, $(X_1, \dots, X_p) = (X(t_1), \dots, X(t_p))$, then, given the notation and results in Section 2, $X(t)$ can be approximated by its ‘empirical’ projection onto V_J defined as

$$X^J(t) = \sum_{k=0}^{2^{j_0}-1} \tilde{\xi}_{j_0,k} \phi_{j_0,k} + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \tilde{\eta}_{j,k} \psi_{j,k}$$

where $\tilde{\xi}_{j_0,k}$ and $\tilde{\eta}_{j,k}$ are the empirical scaling and wavelet coefficients. We will collect them into a vector $\tilde{\gamma}^J \in \mathbb{R}^p$. Let β_s^J be the projection of β_s onto V_J and let us denote by $\beta_s^J \in \mathbb{R}^p$ the vector collecting its scaling and wavelet coefficients, $s = 1, \dots, K$. The original model (2) is then replaced by its discrete counterpart

$$Y = g(\langle \beta_1^J, \tilde{\gamma}^J \rangle, \dots, \langle \beta_K^J, \tilde{\gamma}^J \rangle) + \varepsilon. \quad (3)$$

As much as K and J are large enough and thanks to the isometries between L_2 and ℓ_2 and the compression properties of the wavelet transform, the original functional regression model (1) may be replaced by the above model (3) which

will be our candidate for further dimension reduction by MAVE. A compact way to write down model (3) is

$$Y = g\left(B^T \tilde{\gamma}^J\right) + \varepsilon, \quad (4)$$

$B \in \mathbb{R}^{p \times K}$, $p = 2^J$. The method is then applied to Eq. (4). For the sake of completeness, we briefly describe hereafter the MAVE method, as it is applied on data obeying model (4).

Let d represent now the working dimension, $1 \leq d \leq K$. For an assumed number d of directions in model (3) and known directions B_0 , one would typically minimize

$$\min \mathbb{E}\{Y - E(Y|B_0^T \tilde{\gamma}^J)\}^2,$$

to obtain a nonparametric estimate of the regression function $\mathbb{E}(Y|B_0^T \tilde{\gamma}^J)$. The MAVE method is based on the local linear regression, which hinges at a point $\tilde{\gamma}_0^J$ on linear approximation

$$\mathbb{E}(Y|B_0^T \tilde{\gamma}^J) \approx a + b^T B_0^T (\tilde{\gamma}^J - \tilde{\gamma}_0^J). \quad (5)$$

Now, if directions B_0 are not known, we have to search their approximation B . Xia et al. (2002) propose to plug-in unknown directions B in the local linear approximation of the regression function and to optimize simultaneously with respect to B and local parameters a and b of local linear smoothing. Hence, given a sample $(\tilde{\gamma}_i^J, Y_i)_{i=1}^n$ from $(\tilde{\gamma}^J, Y)$, they perform local linear regression at every $\tilde{\gamma}_0^J = \tilde{\gamma}_i^J$, $i = 1, \dots, n$, and end up minimizing

$$\begin{aligned} \min_{B : B^T B = I_K} \quad & \sum_{l=1}^n \sum_{i=1}^n \left\{ Y_i - \left[a_l + b_l^T B^T (\tilde{\gamma}_i^J - \tilde{\gamma}_l^J) \right] \right\}^2 w_{il} \\ & a_l, b_l, l = 1, \dots, n \end{aligned}$$

where I_K represents the $K \times K$ identity matrix and w_{il} are weights describing the local character of the linear approximation (5) (i.e., w_{il} should depend on the distance of points $\tilde{\gamma}_i^J$ and $\tilde{\gamma}_l^J$).

Xia et al. (2002) call the resulting estimator of B , MAVE and show that the simultaneous minimization with respect to local linear approximation given by a_j, b_j and to directions B results in a convergence rate superior to any other dimension-reduction method. Initially, a natural choice of weights is given by a multidimensional kernel function K_h . At a given $\tilde{\gamma}_0^J$,

$$w_{i0} = K_h(\tilde{\gamma}_i^J - \tilde{\gamma}_0^J) / \sum_{i=1}^n K_h(\tilde{\gamma}_i^J - \tilde{\gamma}_0^J),$$

for $i = 1, \dots, n$ and a kernel function $K_h(\cdot)$, where h refers to a bandwidth.

Additionally, when we already have an initial estimate of the dimension reduction space given by \hat{B} , it is possible to iterate and search an improved estimate of the reduction space. Xia et al. (2002) do so by using the initial estimator \hat{B} to measure distances between points $\tilde{\gamma}_i^J$ and $\tilde{\gamma}_0^J$ in the reduced space. More precisely, they propose to choose in the iterative step weights

$$w_{i0} = K_h(\hat{B}^T(\tilde{\gamma}_i^J - \tilde{\gamma}_0^J)) / \sum_{i=1}^n K_h(\hat{B}^T(\tilde{\gamma}_i^J - \tilde{\gamma}_0^J)).$$

Repeating such iteration steps until convergence results in a refined MAVE (rMAVE) estimator. As one sees from the above equations, the initial estimate \hat{B} depends on local linear smoothing performed with weights computed via a multidimensional kernel on \mathbb{R}^p . Since, by Theorem 1 in Xia et al. (2002) the optimal kernel bandwidth h must be such that $h = \mathcal{O}(\log n/n^{1/p})$, in order to avoid the curse of dimensionality and to stabilize the computations it is therefore advisable in practice to reduce the initial resolution $\log_2 p$ to some resolution $J < \log_2 p$, and in such a way that the approximation of $\tilde{\gamma}$ by its projection $\tilde{\gamma}^J$ does not affect the asymptotics. When assuming that the process X is α -regular it is easy to see, by the results in Section 2, that such a condition holds if $2^\alpha J = \mathcal{O}(n^{1/p})$. From now on, whenever we talk or refer to MAVE, we mean its refined version rMAVE with such a choice of smoothing parameters.

In order to get the convergence rates of Xia et al. (2002), we need the same conditions as those required for the proof of their Theorem 1 and which may be found in the above mentioned paper. We would like however to comment their condition 1, which requires that (X_i, Y_i) is a stationary (with the same distribution as (X, Y)) and absolutely regular sequence, i.e.

$$\beta(k) = \sup_{i \geq 1} \mathbb{E} \left\{ \sup_{\mathcal{F}_{i+k}^\infty} |P(A|F_i^i) - P(A)| \right\} \rightarrow 0 \quad \text{as } k \rightarrow \infty$$

where \mathcal{F}_i^k denotes the σ -field generated by $\{(X_\ell, Y_\ell) : i \leq \ell \leq k\}$. Further $\beta(k)$ decreases at a geometric rate. It is easy to see that if compactly supported wavelets are used in the decomposition, such a stationarity and mixing condition holds also for the wavelet coefficients $\tilde{\gamma}_i$ of the original processes X_i , since the coefficients are integrals over bounded intervals of the sample paths.

The described methods are capable of estimating the dimension reduction space provided we can specify its dimension. To determine the dimension d , Xia et al. (2002) extend the cross-validation approach of Yao and Tong (1994). The cross-validation criterion is defined as

$$CV(d) = \sum_{j=1}^n \left[Y_j - \sum_{i=1, i \neq j}^n \frac{Y_i K_h(\hat{B}^T(\tilde{\gamma}_i^J - \tilde{\gamma}_j^J))}{\sum_{i=1, i \neq j}^n Y_i K_h(\hat{B}^T(\tilde{\gamma}_i^J - \tilde{\gamma}_j^J))} \right],$$

for $d > 0$ and for the special case of independent Y and X as

$$CV(0) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Consequently, the dimension is then determined as

$$\hat{d} = \operatorname{argmin}_{0 \leq d \leq K} CV(d),$$

where K represents the initial number of basis functions in model (1). This is the cross-validation procedure, that we have used in our simulation study and for the analysis of the real examples.

4 Functional Sliced Inverse Regression

As we previously noticed, the MAVE method is one among many that achieves a dimension reduction via a particular estimate of $\operatorname{var}(\mathbb{E}(X|Y))$. Another possible method, as discussed in Li (1991), is SIR which is an appealing method for explanatory analysis by providing a way to derive the direction in which X is best explained by Y . Standard SIR utilizes simple non-smooth nonparametric estimates of the inverse regression curves which may miss important relevant information as the continuous nature of the data is ignored. To address this limitation, in this section, we concentrate on a procedure obtained by smoothing the regression $\mathbb{E}(X|Y)$ via wavelet binning on the Y -values, which replaces slicing of Y by a more relevant smoothing procedure. However, classical SIR must be transposed to our functional data setting but this is supported by the fact that SIR features remain in the Hilbert space context as it has been shown recently by Dauxois et al. (2001).

Let us indicate by Γ the covariance operator of X and by Γ_e the covariance operator of $\mathbb{E}(X|Y)$. Both of them are symmetric, positive, compact and nuclear and therefore Hilbert-Schmidt.

We will consider the following assumption which is the Hilbert version of the ellipticity condition required in Li's Theorem 2.1.

Assumption 4.1 *For all \mathbf{b} in H , there exist c_0, \dots, c_K in R such that*

$$\mathbb{E}(\langle b, X \rangle \mid \langle \beta_1, X \rangle, \dots, \langle \beta_K, X \rangle) = c_0 + \sum_{j=1}^K c_j \langle \beta_j, X \rangle.$$

Under such an assumption, it is possible to prove (see Theorem 1 of Dauxois et al. (2001)), that $\mathbb{E}(X \mid \langle \beta_1, X \rangle, \dots, \langle \beta_K, X \rangle)$ belongs to $\overline{\Gamma(\operatorname{span}(\beta_1, \dots, \beta_k))}$

and to deduce that the operator Γ_e degenerates in any direction Γ -orthogonal to the EDR space.

The above show that the operator $\Gamma^{-1}\Gamma_e$ is a finite rank operator whose image is contained into the EDR space. This operator is a compact operator and it is self-adjoint with respect to the inner product induced by Γ . Using this remark, Ferre and Yao (2000), mimicking the multivariate case, approach the EDR space by the subspace spanned by the Γ -orthonormal eigenvectors $(\beta_1, \dots, \beta_K)$ associated with the K largest nonzero eigenvalues of the operator $\Gamma^{-1}\Gamma_e$. While their estimation procedure seems to be a straightforward replication of the multivariate SIR, this is not the case because while it exists, the inverse of Γ is not bounded. Using the fact that the operator Γ has finite trace so that the sequence of its eigenvalues has limit 0, similarly to Bosq (1991), they replace Γ by a sequence of finite rank operators, with bounded inverse and converging to Γ . We refer the reader to Ferre and Yao (2000) for more details on the procedure.

Remark 4.1 *The linearity condition in assumption 4.1 is similar to the one used in the finite dimensional case. As it is well-known, it is verified when the explanatory variable has symmetric elliptical distribution. Ferre and Yao (2000) have shown that this property holds also in the Hilbert case. As Li (1991) has pointed out, the linearity condition is not a severe restriction, since most low-dimensional projections of a high-dimensional data cloud are close to being normal (Diaconis and Freedman, 1984; Hall and Li, 1993). In addition, there often exist transformations of the predictors that make them comply with the linearity condition. However, note that if this linearity condition does not hold, the directions found by our functional SIR are not always characterizable but can still be useful in identifying some main features of the regression model. In such a case further research is needed to better understand the general relationship between the directions found by our methods and those of the EDR space.*

Using the above results, we now tackle the extension of the inverse regression method to the functional setting in a different way. We use the fact that $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$ is a finite rank operator whose eigenvectors are the ones of $\Gamma^{1/2}\Gamma_e\Gamma^{1/2}$, the reason being that a smooth estimate of Γ_e produces more stable estimates of the eigenvalue decomposition of Γ_e than that of the empirical estimate of Γ . Moreover our estimation method of Γ_e differs from that of Ferre and Yao (2000).

Let again (X_i, Y_i) , $i = 1, \dots, n$, a sample from (X, Y) . Then Γ can be estimated by

$$\hat{\Gamma}_n = \frac{1}{n} \sum_{i=1}^n X_i \otimes X_i. \quad (6)$$

Under the assumption that $\mathbb{E} \|X\|_{L^2([0,1])}^4 < \infty$, it is possible to show (see Bosq,

2000) that

$$\|\hat{\Gamma}_n - \Gamma\|_S = \mathcal{O}\left(\frac{1}{n}\right) \quad \text{in probability.}$$

Regarding Γ_e we may proceed in the following way. Let us assume that Y has a density f with respect to the Lebesgue measure on \mathbb{R} . Let $M_y(t) = E(X|Y = y)$ and let $\hat{M}_y(t)$ be the wavelet smoothing of $X(t)$ with design points the Y_i 's, $i = 1, \dots, n$ obtained through the BINWAV estimator (Antoniadis and Pham, 1998). Then we consider the following estimate for Γ_e :

$$\hat{\Gamma}_{e,n} = \frac{1}{n} \sum_{i=1}^n \hat{M}_{Y_i} \otimes \hat{M}_{Y_i}. \quad (7)$$

Denoting by $\|\cdot\|_S$ the Hilbert-Schmidt operator norm, it holds

Theorem 4.1 *Under the assumptions stated in the appendix*

$$\|\hat{\Gamma}_{e,n} - \Gamma_e\|_S = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \quad \text{in probability.}$$

as $n \rightarrow \infty$.

An outline of the proof of the above theorem is relegated to the Appendix.

Now, recall that a basis of the EDR space is given by the eigenvectors of $\Gamma^{1/2}\Gamma_e^+\Gamma^{1/2}$. By Theorem 4.1, and the above results, using a device similar to the one by Dauxois et al. (2001), one may easily show that the eigenvectors of $\hat{\Gamma}_n^{1/2}\hat{\Gamma}_{e,n}^+\hat{\Gamma}_n^{1/2}$ converge towards the eigenvectors of $\Gamma^{1/2}\Gamma_e^+\Gamma^{1/2}$ at a parametric rate. It follows then that once $\hat{\Gamma}_n$ and $\hat{\Gamma}_{e,n}$ have been estimated, in order to estimate the EDR space one must evaluate the spectral decomposition of $\hat{\Gamma}_n^{1/2}\hat{\Gamma}_{e,n}^+\hat{\Gamma}_n^{1/2}$. In order to improve the conditioning of $\hat{\Gamma}_{e,n}$, we apply a projection method before performing the spectral decomposition. Let us denote by $\hat{\pi}_{k_n}$ the orthogonal projector into the space spanned by the k_n eigenvectors of $\hat{\Gamma}_{e,n}$ corresponding to the k_n largest eigenvalues; we let $\hat{\Gamma}_{e,n}^{k_n} = \hat{\pi}_{k_n}\hat{\Gamma}_{e,n}\hat{\pi}_{k_n}$. Estimation of the EDR space is derived from the spectral decomposition of

$$\hat{\Gamma}_n^{1/2} \left(\hat{\Gamma}_{e,n}^{k_n}\right)^+ \hat{\Gamma}_n^{1/2}. \quad (8)$$

Let $(\alpha_i)_{i=1,\dots,K}$ be the smallest eigenvalues of (8) and η_i the corresponding eigenfunctions, then

$$\beta_i = \frac{1}{\alpha_i} \left(\hat{\Gamma}_{e,n}^{k_n}\right)^+ \hat{\Gamma}_n^{1/2} \eta_i. \quad (9)$$

Summarizing the procedure for computing the EDR directions $\beta_1(t), \dots, \beta_s(t)$ goes through the following steps:

Algorithm

- (1) calculate $\hat{M}_y(t)$, the wavelet smoothing of $X(t)$ with design points Y_1, \dots, Y_n , using the BINWAV estimator;
- (2) estimate $\hat{\Gamma}_n$ by (6) and $\hat{\Gamma}_{e,n}$ by (7);
- (3) evaluate the spectral decomposition of $\hat{\Gamma}_{e,n}$ and its projection $\hat{\Gamma}_{e,n}^{k_n}$;
- (4) evaluate the spectral decomposition of $\hat{\Gamma}_n^{1/2} \left(\hat{\Gamma}_{e,n}^{k_n} \right)^+ \hat{\Gamma}_n^{1/2}$ and estimate the EDR directions by Eq. (9).

Here again one of the important questions is to determine the number of effective dimensions. Most SIR procedures rely upon the asymptotic normality of $\sqrt{n}(\hat{\Gamma}_{e,n} - \Gamma_e)$, but we were not able to prove such a result in the infinite dimensional case and this question remains for future research. Whenever such a result is available, a way to estimate the dimension of the EDR space, when the sample paths of the process are discretized on p values (t_1, \dots, t_p) , is the chi-square sequential test proposed by Li (1991) to test the nullity of the $p - K$ smallest eigenvalues of $\hat{\Gamma}_n^{1/2} \left(\hat{\Gamma}_{e,n}^{k_n} \right)^+ \hat{\Gamma}_n^{1/2}$, with p number of covariates. Indeed, if the data were normally distributed, then (see Li, 1991) $n(p - K)\bar{\lambda}_{p-K}$ follows a χ^2 distribution with $(p - K)(H - K - 1)$ degrees of freedom asymptotically, where $\bar{\lambda}_{p-K}$ denotes the average of the smallest $p - K$ eigenvalues and H the length of the BINWAV estimator.

There is an alternative approach for more general predictor distributions, namely the trace criterion developed by Ferre (1998) which consists in examining, for $q = 1, \dots, K$ the values of

$$R(q) = \mathbb{E}(r(q)) = \mathbb{E} \left(\frac{1}{q} \text{tr}(\pi_q \hat{\pi}_q) \right),$$

where we have denoted by π_q the $\Gamma_{n,p}$ - orthogonal projector onto the space spanned by $(\beta_1, \dots, \beta_q)$ and by $\hat{\pi}_q$ the $\hat{\Gamma}_{n,p}$ - orthogonal projector onto the space spanned by $(\hat{\beta}_1, \dots, \hat{\beta}_q)$. In this case $\Gamma_{n,p}$ denotes the covariance matrix of the discretized process $X_{n,p}$ and $\hat{\Gamma}_{n,p}$ is its estimate by means of Eq. (6) in (t_1, \dots, t_p) . Since $R(q)$, for $q = 1, \dots, K$, will converge to 1 as n tends to infinity, then for a fixed n a reasonable way to assess that an EDR direction is available, is given by looking at how much $R(q)$ departs from 1. Consistent estimates of $R(q)$ are given in Ferre (1998) and their use for estimating the dimension of the EDR space are described in the above mentioned paper. However the computational cost in implementing such a procedure is high and we have not included this procedure in our comparisons.

5 Numerical experiments

To investigate and compare the performance of the two dimensional reduction regression methods proposed in this paper we have conducted a Monte Carlo simulation for a particular model. We have also applied the wavelet based MAVE and the functional sliced inverse regression to several spectrometric data sets. We have also compared our results with those obtained with the Bayesian wavelet variable selection method discussed in Brown et al. (2001). The algorithm to generate the wavelet MAVE and SIR approximations was implemented in MATLAB. The MATLAB scripts are available upon request for the interested readers.

5.1 Simulation

Let $H = L^2([0, 1])$, $\mathbf{X} = (X(t))_{t \in [0,1]}$ be a standard Brownian motion and ϵ a mean zero Gaussian distribution with variance σ^2 independent of \mathbf{X} . All curves $\beta_i(t)$ and $X(t)$ are discretized on the same grid generated from p equispaced points $t \in [0, 1]$. The observations Y were generated from i.i.d. observations of X and ϵ according to the following model:

Model 1

$$Y = \exp(\langle \beta_1(t), X(t) \rangle) + \exp(|\langle \beta_2(t), X(t) \rangle|) + \epsilon,$$

$$\beta_1(t) = \sin(3\pi t/2), \beta_2(t) = \sin(5\pi t/2), \sigma = 0.1.$$

The motivation for this example is that the functions β_1 and β_2 belong to the eigen-subspace of the covariance operator of X and therefore it represents the ideal situation where the EDR space is included in the central subspace.

The goal of our simulations is to compare prediction performance among the various methods. We therefore generated by Monte Carlo simulation two sets of samples, one for fitting (training set) and one for prediction (validation set). To get a good model selection during the training phase the number of samples for the training set is $n = 500$ for the training set; in order to have stable average estimates of RMSE and its distribution, n has been chosen equal to 2000 for the test set. Each continuous curve has been sampled at $p = 128$ equally spaced time points.

Among the methods that were compared by means of these simulations we have also included the original MAVE procedure by Xia et al. (2002), where each discretized X curve is modelled as a p -dimensional multivariate predictor. The following four methods have been considered for the analysis:

Table 1

Root Mean Squared Error and number of selected directions (variables in the case of BW) for WM, DM, WS, and BW predictors and for brownian motion data.

Brownian motion data				
	WM	DM	WS	BW
RMSE	3.4	3.5	3.4	3.8
Number of directions/variables	2	2	2	1

WM: wavelet-MAVE introduced in Section 3.;

DM: original MAVE developed in Xia et al. (2002);

WS: wavelet-SIR introduced in Section 4.;

BW: Bayesian wavelet variable selection introduced in Brown et al. (2001).

For all analyses, data were centered by subtraction of the training set means from the training and the validation data. To evaluate the predictive performance of the methods we have used the root mean squares prediction error defined by

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (Y_j - \hat{Y}_j)^2},$$

where \hat{Y}_j denote the predicted values and Y_j are the corresponding true values.

The root mean squared error of predictions on the $n = 2000$ validation samples for the Brownian motion example is reported in Table 1 together with the number of directions chosen (or variables selected in the case of BW).

The number of directions for the SIR method was selected by a simple analysis of the eigenvalues of the matrix in Eq. (8) thanks to their good behavior, since indeed there is a marked gap between the nonzero eigenvalues and the zero ones. This is consistent with the number of directions chosen by cross-validation MAVE (generally 2 to 3), which appears optimal since the functions β_1 and β_2 are respectively the second and third eigenvalues of the covariance operator of X . The Table 1 shows the good performance of wavelet based methods, followed closely by the original MAVE method, while the Bayesian wavelet variable selection method gives the worst result due probably to the fact that the directions β_1 and β_2 are of low frequency content.

Whatever the estimation of the directions β_i , $i = 1, 2$ are, the quality of prediction is related to how close the estimated projections $\langle \hat{\beta}_i, X \rangle$ are to the true projections $\langle \beta_i, X \rangle$. In order to check this, Figure 1 displays the indexes $\langle \hat{\beta}_i, X \rangle$ versus $\langle \beta_i, X \rangle$ for each of the wavelet based dimension reduction methods and for each validation sample, showing a quite satisfactory estimation, even if the WM method produces estimates that seem more variable.

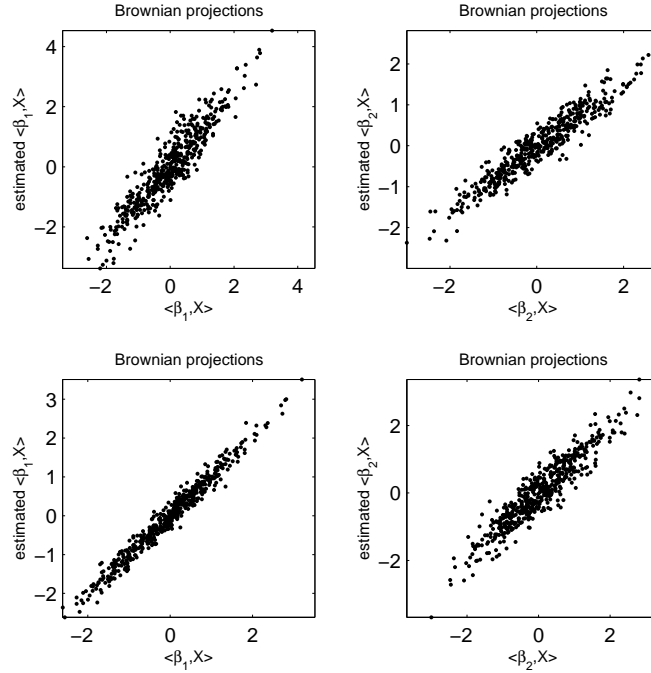


Fig. 1. Estimated projections versus true ones for the simulated example: top for wavelet MAVE method (WM), bottom for wavelet SIR method (WS).

5.2 Real data examples

We now focus on the analysis of several real data sets from spectrometric experiments that have already been used by various authors to illustrate functional data regression procedures. These real data examples are respectively: the `tecator` data set, the `cookie` data set and the `temperature` data set. All of them are chemometric data derived from Near InfraRed (NIR) spectroscopy experiments where the aim is to predict a component concentration of interest starting from measurements of the spectrum of a mixture of several compounds.

Real example 1 *The `tecator` data are recorded by a Tecator near-infrared spectrometer (the Tecator Infratec Food and Feed Analyzer) which measures the spectrum of light transmitted through a sample of minced pork meat in the region 850 - 1050 nanometers (nm) (Thodberg, 1996). Each sample contains finely chopped pure meat with different moisture, fat and protein contents. For each meat sample the data consists of a 100 channel spectrum of absorbances and the contents of moisture (water), fat and protein. The absorbance is $-\log_{10}$ of the transmittance measured by the spectrometer. The three contents, measured in percent, are determined by analytic chemistry. The total number of samples is 240; the first 125 spectra have been used to train the methods and the last 115 to test them. The aim is to predict the percentage of each content Y given the corresponding spectrometric curve X . The spectral range used for the*

wavelet MAVE methods is 902 - 1028 nm, for a total of 64 channels, because it contains the most informative part of the spectrum for the three contents.

Real example 2 The `cookie` data consist of two NIR spectra data sets aimed at measuring the composition of biscuit dough pieces (formed but unbaked biscuits). The variables under investigation are fat, sucrose, dry flour, and water (for a full description of the experiment, see Osborne et al. (1984)). The first data set used for training contains 39 spectra, and the one used for testing contains 31 spectra. The original spectral data are composed of 700 channels in the spectral range 1100 - 2498 nm in steps of 2 nm, but following Brown et al. (2001), we have reduced the channels to 256, considering just the spectral range 1380 - 2400 nm, over which we have taken every other point, thus increasing the gap to 4 nm. The first 140 and last 49 wavelengths have been removed because it is supposed they contain little useful spectral information.

Real example 3 The `temperature` data refer to the experiment presented by Wulfert et al. (1998) which involves mixtures of ethanol, water and isopropanol, and are available at <http://www-its.chem.uva.nl>. The data set consists of 19 NIR spectra measured in the spectral range 580 - 1091 nm in 1 nm intervals for a total of 512 channels under several temperature conditions: 30, 40, 50, 60 and 70 °C ($\pm 0.2^\circ\text{C}$). For absorption and noise reasons, all data analyses were performed using just 128 channels, i.e., the sub-band 896 - 1023 nm. The problem is to predict the percentage of ethanol, water and isopropanol using as a training set only spectra at 30 °C or at 50 °C, and then see how stable prediction is using spectra at the other temperatures. See Marx and Eilers (2002) for a panoramic of three high dimensional modelling strategies applied to this problem.

As for the simulations, we have again compared the four dimension reduction procedures WM, DM, WS and BW. However, since the response variables in these examples are in fact percentages, following Brown et al. (2001), we have also transformed the original Y 's into log ratios of the form $Z = \log(Y/1 - Y)$ before applying our wavelet based procedures resulting into 3 extra procedures with acronyms LWM, LDM and LWS, that are as follows:

LWM: wavelet-MAVE with logit transformation of the response;

LDM: original MAVE method with logit transformation of the response as in LWM;

LWS: wavelet-SIR with data subject to the same logit transform as in LWM.

For each model, the best dimensionality reduction model was selected by fitting the training set. For each best model, we finally computed the corresponding predictions on the validation set. To get an idea of the variability in the prediction errors we also display boxplots of the errors $|Y_j - \hat{Y}_j|$, $j = 1, \dots, n$, where n denotes the number of observations in the validation set for each

Table 2

Root Mean Squared Error and number of directions/variables selected for WM, DM, LWM, LDM, WS, LWS and BW predictors and for tecator data.

Tecator data							
	WM	DM	LWM	LDM	WS	LWS	BW
water	0.042 (2)	0.048 (2)	0.041 (2)	0.051 (2)	0.038 (8)	0.036 (8)	0.037 (4)
fat	0.054 (2)	0.063 (2)	0.11 (2)	0.11 (2)	0.047 (8)	0.083 (7)	0.050 (6)
protein	0.014 (2)	0.027 (2)	0.01 (2)	0.027 (2)	0.013 (8)	0.013 (8)	0.017 (3)

example.

Table 2 summarizes the results for Example 1. The number of EDR directions for wavelet-SIR was selected by the χ^2 test (Li, 1991) as discussed in Section 4. Boxplots of the error are displayed in Figure 2. Both the table and the boxplots show that all the methods have good and equivalent performances.

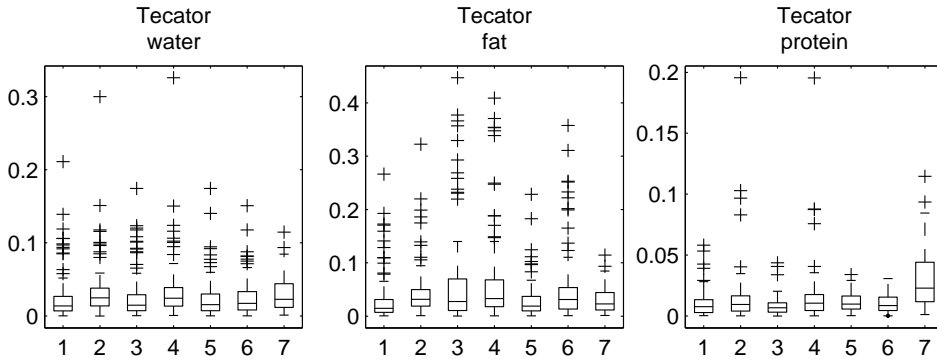


Fig. 2. Boxplot of prediction error for tecator data and for each of the predictors (1: WM, 2: DM, 3: LWM, 4: LDM, 5: WS, 6: LWS, 7: BW).

The results for Example 2 are summarized in Table 3 and in Figure 3. The number of EDR directions for wavelet-SIR was selected empirically by looking directly at the eigenvalues of the matrix in Eq. (8), since the smallest K eigenvalues are significantly different from zero. The boxplot for the BW predictor is not shown in Figure 3 because of its bad performance respect to the other methods that would not permit a clear visualization of the results.

Finally, for the last Example 3, Table 4 summarizes the results together with Figure 4 which displays the boxplots of the prediction error. The number of EDR directions for wavelet SIR was chosen through the analysis of the χ^2 test. Both the wavelet MAVE method and the original MAVE work better than the other methods when the testing model refers to temperature conditions close to the ones of the training model, as for example when the training is made on the temperature at 30 °C and validation for the sample at 40-50 °C. Similar results, not displayed here, hold when training is made at 50 °C and testing

Table 3

Root Mean Squared Error and number of directions/variables selected for WM, DM, LWM, LDM, WS, LWS and BW predictors and for cookie data.

Cookie data							
	WM	DM	LWM	LDM	WS	LWS	BW
fat	0.0065 (2)	0.0056 (2)	0.0071 (2)	0.0051 (2)	0.0038 (7)	0.0036 (7)	0.24 (10)
sucrose	0.029 (2)	0.021 (2)	0.031 (2)	0.022 (2)	0.0079 (7)	0.011 (7)	0.68 (10)
flour	0.028 (2)	0.021 (2)	0.028 (2)	0.021 (2)	0.0062 (7)	0.0062 (7)	0.59 (10)
water	0.0068 (2)	0.0055 (2)	0.0062 (2)	0.0050 (2)	0.0034 (7)	0.0034 (7)	0.22 (10)

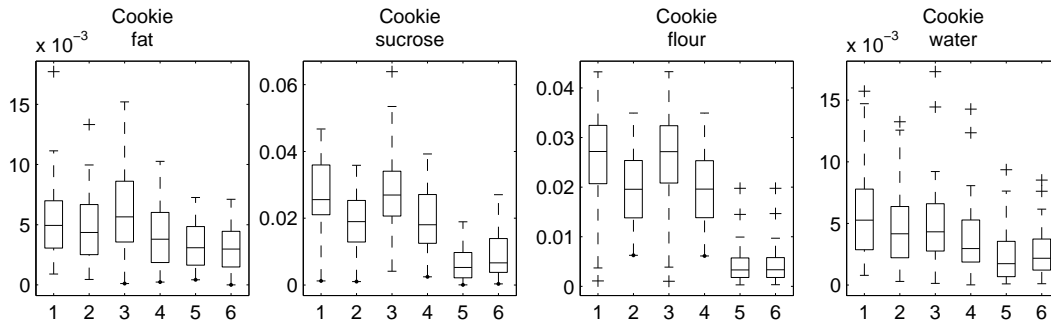


Fig. 3. Boxplot of prediction errors for the cookie data and the six predictors (1: WM, 2: DM, 3: LWM, 4: LDM, 5: WS, 6: LWS).

at 40-60 °C.

Finally we observe that the computational time required to select directions/variables strongly depends on the method: it ranges from the order of hours on a Pentium IV PC for the Bayesian wavelet variable selection (for each trial on the prior) to a few minutes for the MAVE procedures and less than one minute for wavelet SIR method on the same platform.

From these examples, we can see that our methods are in particular well adapted to spectrometric calibration problems and compare equally and sometimes favorably in terms of prediction quality with other methods.

6 Conclusion

In this paper we provide a wavelet based variant of two popular dimension reduction methods that are traditionally used in practice, namely MAVE and SIR. The proposed procedures are computationally fast and seem to be well adapted to spectrometric calibration problems. Although it is in general non-trivial to extend multivariate dimension reduction methods for

Table 4

Root Mean Squared Error and number of directions/variables selected for WM, DM, WS, and BW predictors and for temperature data, training 30 °C, testing 40, 50, 60 and 70 °C.

Temperature data				
	40 °C			
	WM	DM	WS	BW
ethanol	0.066 (2)	0.052 (2)	0.19 (2)	0.19 (1)
water	0.022 (2)	0.019 (2)	0.071 (2)	0.067 (1)
isopropanol	0.058 (2)	0.049 (2)	0.19 (2)	0.15 (1)
	50 °C			
	WM	DM	WS	BW
ethanol	0.12 (2)	0.10 (2)	0.25 (2)	0.30 (1)
water	0.053 (2)	0.047 (2)	0.085 (2)	0.10 (1)
isopropanol	0.11 (2)	0.10 (2)	0.33 (2)	0.24 (1)
	60 °C			
	WM	DM	WS	BW
ethanol	0.18 (2)	0.16 (2)	0.16 (2)	0.23 (1)
water	0.10 (2)	0.075 (2)	0.14 (2)	0.14 (1)
isopropanol	0.16 (2)	0.17 (2)	0.45 (2)	0.26 (1)
	70 °C			
	WM	DM	WS	BW
ethanol	0.23 (2)	0.22 (2)	0.20 (2)	0.16(1)
water	0.16 (2)	0.12 (2)	0.16 (2)	0.11 (1)
isopropanol	0.21 (2)	0.24 (2)	0.55 (2)	0.11(1)

functional data, we show that both these methods can be generalized in a relatively straightforward way to such a setting. Theoretically the wavelet MAVE procedure might be the most appealing given its simplicity and the ease of its implementation.

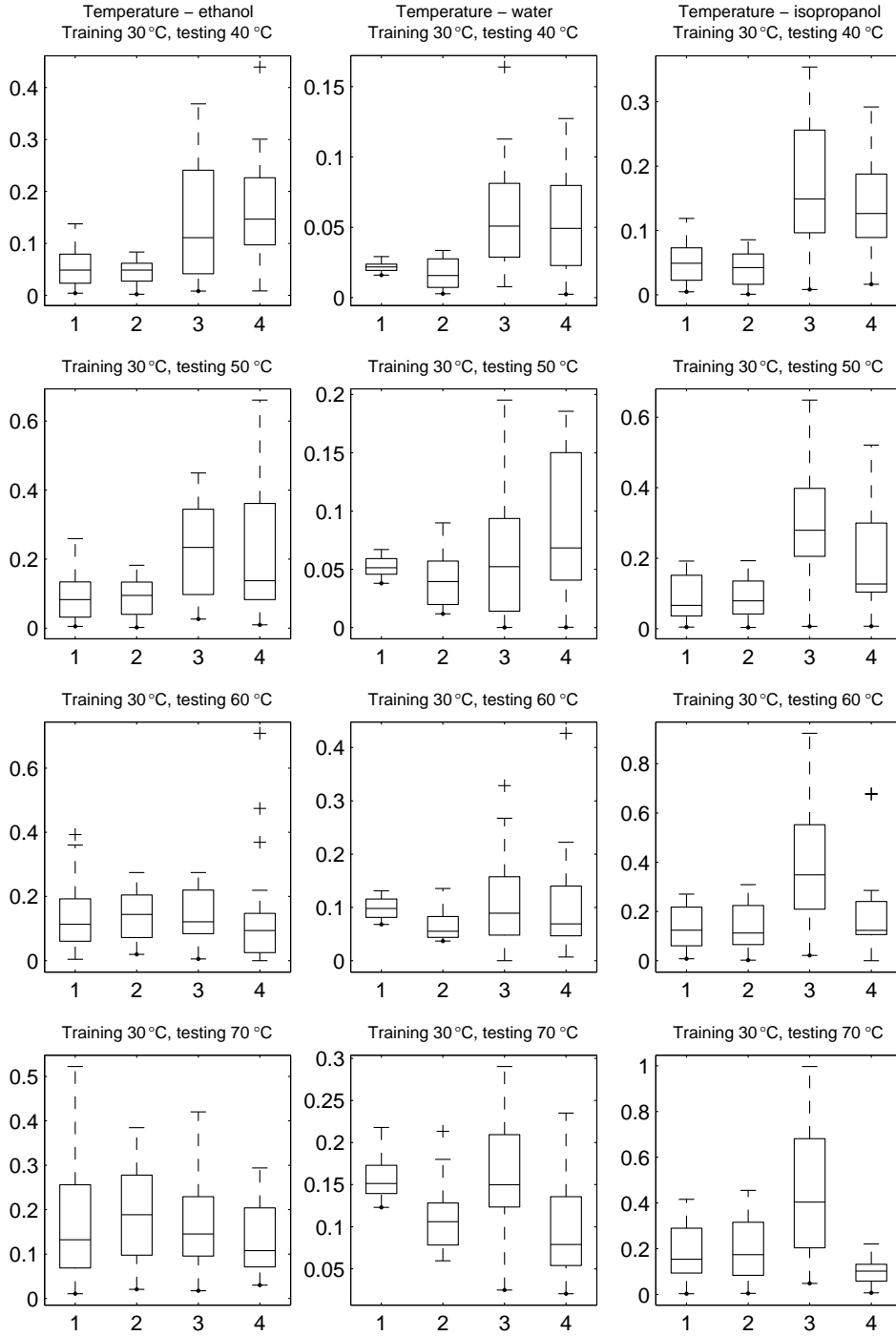


Fig. 4. Boxplots of prediction errors for the temperature data (training set: 30 °C; testing set: 40, 50, 60 and 70 °C), and for the four predictors (1: WM, 2: DM, 3: WS, 4: BW).

Appendix

We state here the necessary assumptions for Theorem 4.1 to hold. For the functional SIR method, the following conditions are supposed to hold:

- [H-0] The marginal density f of Y is compactly supported and bounded below by a positive constant.
- [H-1] The marginal density f of Y and, for each $t \in [0, 1]$, $g_t(y) = \mathbb{E}(X(t)|Y = y)$ are α -times continuously differentiable, where α is an integer larger than 1.
- [H-2] There exist $\delta > 0$ such that $\mathbb{E}(\|X\|^{4+\delta}) < +\infty$.
- [H-3] The wavelet system used for estimation is compactly supported and of regularity larger than α .

The rest of this section is devoted to the proof of Theorem 4.1.

Proof (Theorem 4.1). Let

$$\tilde{\Gamma}_{e,n} = \frac{1}{n} \sum_{i=1}^n M_{Y_i} \otimes M_{Y_i}.$$

Then

$$\Gamma_e - \hat{\Gamma}_{e,n} = \Gamma_e - \tilde{\Gamma}_{e,n} + \tilde{\Gamma}_{e,n} - \hat{\Gamma}_{e,n} = O_p\left(\frac{1}{n}\right) + \tilde{\Gamma}_{e,n} - \hat{\Gamma}_{e,n}$$

by the weak law of large numbers. Now let us concentrate on $\tilde{\Gamma}_{e,n} - \hat{\Gamma}_{e,n}$. It holds

$$\begin{aligned} \hat{\Gamma}_{e,n} - \tilde{\Gamma}_{e,n} &= \frac{1}{n} \sum_{i=1}^n (M_{Y_i} \otimes M_{Y_i} - \hat{M}_{Y_i} \otimes \hat{M}_{Y_i}) = \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{M}_{Y_i} - M_{Y_i}) \otimes (\hat{M}_{Y_i} - M_{Y_i}) + \frac{1}{n} \sum_{i=1}^n M_{Y_i} \otimes (\hat{M}_{Y_i} - M_{Y_i}) + \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{M}_{Y_i} - M_{Y_i}) \otimes M_{Y_i} = A_1 + A_2 + A_3. \end{aligned}$$

We will show that

$$A_1 = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right), A_2 = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right), A_3 = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right).$$

We recall first here that the estimator \hat{M}_Y of M_Y is obtained by wavelet smoothing of $X(t)$ with design points the Y_i 's, $i = 1, \dots, n$ through the BINWAV estimator (Antoniadis and Pham, 1998). Note that, similarly to a Naradaya-Watson estimator, this estimator is defined as the ratio of two estimators, the numerator \hat{g} being an estimate of $g = M_Y f$ and the denominator $\hat{f}(y)$ being also a wavelet binned estimator of the marginal density of Y . However, since the latter may not be a bona fide estimator, we will use instead an estimator of the form $\hat{f}_n(y) = [\hat{f}(y) - \mu_n]_+$, where μ_n is a sequence of positive numbers converging to 0 as $n \rightarrow +\infty$. By the results of Efromovich

(1989) the corresponding estimator $\hat{M}_{n,Y}$ retains the same asymptotic properties as those of \hat{M}_Y . To not complicate notations we will still denote $\hat{M}_{n,Y}$ by \hat{M}_Y in the sequel.

Let us consider first A_1 . We have

$$\hat{M}_{Y_i} - M_{Y_i} = \frac{M_{Y_i}}{\hat{f}_n(Y_i)} \left(f(Y_i) - \hat{f}_n(Y_i) \right) + \frac{1}{\hat{f}_n(Y_i)} \left(\hat{g}(Y_i) - g(Y_i) \right).$$

Since \hat{f}_n is bounded below by μ_n , it follows that

$$\begin{aligned} \|\hat{M}_{Y_i} - M_{Y_i}\| &\leq \frac{\|M_{Y_i}\|}{\mu_n} \left| f(Y_i) - \hat{f}_n(Y_i) \right| + \frac{1}{\mu_n} \|\hat{g}(Y_i) - g(Y_i)\| \\ &\leq \frac{\|M_{Y_i}\|}{\mu_n} \sup_Y \left| f(Y) - \hat{f}_n(Y) \right| + \frac{1}{\mu_n} \sup_Y \|\hat{g}(Y) - g(Y)\| \end{aligned}$$

and consequently

$$\|\hat{M}_{Y_i} - M_{Y_i}\|^2 \leq 2 \frac{\|M_{Y_i}\|^2}{\mu_n^2} \sup_Y \left| f(Y) - \hat{f}_n(Y) \right|^2 + \frac{2}{\mu_n^2} \sup_Y \|\hat{g}(Y) - g(Y)\|^2.$$

From this it follows that

$$\begin{aligned} \sqrt{n} \|A_1\|_S &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^n \|\hat{M}_{Y_i} - M_{Y_i}\|^2 \\ &\leq \sqrt{n} \frac{2}{\mu_n^2} \left(\frac{1}{n} \sum_{i=1}^n \|M_{Y_i}\|^2 \sup_Y \left| f(Y) - \hat{f}_n(Y) \right|^2 \right) + \sqrt{n} \frac{2}{\mu_n^2} \sup_Y \|\hat{g}(Y) - g(Y)\|^2. \end{aligned}$$

By assumption [H-2] and the weak law of large numbers we have

$$\frac{1}{n} \sum_{i=1}^n \|M_{Y_i}\|^2 = \mathbb{E} \left(\|M_y(t)\|^2 \right) + \mathcal{O}_p \left(\frac{1}{n} \right) \quad (10)$$

Now by assumptions [H-1] and [H-3] and using the results from Antoniadis and Pham (1998), we also have

$$\sup_Y \left| f(Y) - \hat{f}_n(Y) \right|^2 = \mathcal{O}_p \left(h^\alpha + \frac{\sqrt{\log n}}{\sqrt{n}\sqrt{h}} \right)^2 \quad (11)$$

$$\sup_Y \|\hat{g}(Y) - g(Y)\|^2 = \mathcal{O}_p \left(h^\alpha + \frac{\sqrt{\log n}}{\sqrt{n}\sqrt{h}} \right)^2, \quad (12)$$

where we have indicated by h the binwidth.

By equalities (10), (12) and (12) it follows that

$$\sqrt{n}A_1 = \sqrt{n}\mathcal{O}_p\left(\frac{1}{\mu_n}\left(h^\alpha + \frac{\sqrt{\log n}}{\sqrt{n}\sqrt{h}}\right)\right)^2. \quad (13)$$

Let now ϵ_1 and ϵ_2 two positive constants such that

$$\epsilon_2 < \frac{2\alpha - 1}{4(2\alpha + 1)}, \quad \frac{\epsilon_2}{\alpha} + \frac{1}{4\alpha} \leq \epsilon_1 < \frac{1}{2} - 2\epsilon_2.$$

A simple calculation, shows that taking $h = n^{-\epsilon_1}$ and $\mu_n = n^{-\epsilon_2}$ leads to the appropriate rate of convergence.

Expression A_3 is the symmetric of A_2 so it is only necessary to consider A_2 . It holds

$$\begin{aligned} A_2 &= \frac{1}{n} \sum_{i=1}^n M_{Y_i} \otimes (\hat{M}_{Y_i} - M_{Y_i}) \\ &= \frac{1}{n} \sum_{i=1}^n M_{Y_i} \otimes \frac{\hat{g}(Y_i) - g(Y_i)}{f(Y_i)} + \frac{1}{n} \sum_{i=1}^n \frac{f(Y_i) - \hat{f}_n(Y_i)}{f(Y_i)\hat{f}_n(Y_i)} M_{Y_i} \otimes (\hat{g}(Y_i) - g(Y_i)) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \frac{f(Y_i) - \hat{f}_n(Y_i)}{\hat{f}_n(Y_i)} M_{Y_i} \otimes M_{Y_i}. \end{aligned}$$

It follows that

$$\begin{aligned} \|A_2\|_S &\leq \frac{1}{\mu_n} \left(\frac{1}{n} \sum_{i=1}^n \|M_{Y_i}\| \sup_Y \|\hat{g}(Y) - g(Y)\| \right) \\ &\quad + \frac{1}{\mu_n} \left(\frac{1}{n} \sum_{i=1}^n \|M_{Y_i}\| \sup_Y \|\hat{g}(Y) - g(Y)\| \sup_Y |f(Y) - \hat{f}_n(Y)| \right) \\ &\quad + \frac{1}{\mu_n} \left(\frac{1}{n} \sum_{i=1}^n \|M_{Y_i}\|^2 \sup_Y |f(Y) - \hat{f}_n(Y)| \right), \end{aligned}$$

By the law of large numbers and assumptions [H-2], both terms $\frac{1}{n} \sum_{i=1}^n \|M_{Y_i}\|$ and $\frac{1}{n} \sum_{i=1}^n \|M_{Y_i}\|^2$ are bounded in probability. Using now assumptions [H-0], [H-1] and [H-3] and arguments similar to those used for A_1 we have

$$\sqrt{n}A_2 = \mathcal{O}_p\left(\frac{\sqrt{n}}{\mu_n}\left(h^\alpha + \frac{\sqrt{\log n}}{\sqrt{n}\sqrt{h}}\right)\right) \quad (14)$$

Regarding A_3 we proceed as for the A_2 term, obtaining the same rate of convergence:

$$\sqrt{n}A_3 = \mathcal{O}_p\left(\frac{\sqrt{n}}{\mu_n}\left(h^\alpha + \frac{\sqrt{\log n}}{\sqrt{n}\sqrt{h}}\right)\right). \quad (15)$$

Using the same choice for h and μ_n as before and inequalities (13), (14) and (15) the theorem follows.

Acknowledgements

Umberto Amato and Italia De Feis work was supported by ASI and ‘Project CNR-CNRS’. Antoniadis Antoniadis was supported by ‘Project IDOPT, INRIA-CNRS-IMAG’, ‘Project CNR-CNRS’ and ‘Project AMOA, IMAG’. Anestis Antoniadis would like to thank Umberto Amato for excellent hospitality while visiting Naples to carry out this work. The authors would like to thank Prof. Marina Vannucci for providing her Matlab code and the cookie data set. They would also like to thank Ferre and Yao for providing a preprint of their paper prior to publication and Yingcun Xia, Howell Tong, W. K. Li and Li-Xing Zhu for providing a set of Matlab functions implementing their MAVE procedures for additive noise models.

References

- [1] Alsberg, B. K. (1993). Representation of spectra by continuous functions. *Journal of Chemometrics*, 7, 177–193.
- [2] Antoniadis, A. and D.-T. Pham (1998). Wavelet regression for random or irregular design. *Computational Statistics & data analysis*, 28, 353–369.
- [3] Besse, P. and J. Ramsay (1986). Principal components analysis of sampled functions. *Psychometrika*, 51, 285–311.
- [4] Bosq, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. In *Nonparametric Functional Estimation and Related Topics, Proc. NATO Advanced Inst. Spetsai*. Ed. G. Roussas, pp. 509–529. Kluwer.
- [5] Bosq, D. (2000). *Linear processes in function space, Theory and Application*. Lectures Notes in Statistics n. 149, New York: Springer.
- [6] Brown, P. J., T. Fearn, and M. Vannucci (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454), 398–408.
- [7] Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & Probability Letters*, 45, 11–22.
- [8] Carmona, R., W. L. Hwang, and B. Torresani (1998). Practical time-frequency analysis: Gabor and wavelet transforms with an implementation in s. In *Wavelet Analysis and its Applications Series, Volume 9*. San Diego: Academic Press.
- [9] Cheng, B. and H. Tong (1996). A theory of wavelet representation and decomposition for a general stochastic process. *Lect. Notes Statist.*, 115, 115–129.

- [10] Cohen, A. and J. P. D'Ales (1997). Nonlinear approximation of random functions. *SIAM J. Appl. Math.*, 57, 518–540.
- [11] Cook, D. (2000). Save: A method for dimension reduction and graphics in regression. *Communications in Statistics: Theory and Methods*, 29, 161–175.
- [12] Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.
- [13] Dauxois, J., L. Ferre, and A. F. Yao (2001). Un modèle semi-paramétrique pour variables aléatoires hilbertiennes. *Comptes Rendus de l'Académie des Sciences*, 333, 947–952.
- [14] Denham, M. C. and P. J. Brown (1993). Calibration with many variables. *Applied Statistics*, 42(3), 515–528.
- [15] Diaconis, P. and D. Freedman (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12, 793–815.
- [16] Efromovich, S. (1989). On sequential nonparametric estimation of a density. *Theory Probab. Applications*, 34, 228–239.
- [17] Ferraty, F. and P. Vieu (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17, 545–564.
- [18] Ferre, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93(441), 132–140.
- [19] Ferre, L. and A. F. Yao (2000). Functional sliced inverse regression analysis. Technical Report LSP-2000-14, University of Toulouse.
- [20] Goutis, C. and T. Fearn (1996). Partial least squares regression on smooth factors. *Journal of American Statistical Association*, 91(434), 627–632.
- [21] Hall, P. and K. C. Li (1993). On almost linearity of low dimensional projection from high dimensional data. *Annals of Statistics*, 21(2), 867–889.
- [22] Hsing, T. and R. J. Carroll (1992). An asymptotic theory for sliced inverse regression. *Annals of Statistics*, 20(2), 1040–1061.
- [23] Kneip, A. and T. Gasser (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics*, 20(2), 1260–1305.
- [24] Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of American Statistical Association*, 86(414), 316–342.
- [25] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11, 674–693.
- [26] Martens, H. and T. Naes (1989). *Multivariate Calibration*. New York: John Wiley & Sons, Inc.
- [27] Marx, B. D. and P. H. C. Eilers (2002). Multivariate calibration stability: a comparison of methods. *Journal of Chemometrics*, 16, 129–140.
- [28] Nason, G. P. and R. von Sachs (1999). Wavelets in time series analysis. *Philosophical Transactions of the Royal Society of London A*, 357(1760), 2511–2526.
- [29] Neveu, J. (1968). Processus aleatoires gaussiens. Technical report, Presses de l'Université de Montréal, Montréal, Quebec.
- [30] Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data*

- Analysis*. Boston: Birkhäuser.
- [31] Osborne, B. G., T. Fearn, A. R. Miller, and S. Douglas (1984). Application of near-infrared reflectance spectroscopy to compositional analysis of biscuits and biscuits doughs. *Journal of the Science of Food and Agriculture*, 35, 99–105.
 - [32] Percival, D. B. and A. T. Walden (2000). *Wavelet Methods for Time Series Analysis*. Cambridge: Cambridge University Press.
 - [33] Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling. *Journal of Royal Statistical Society, Series A*, 145, 285–312.
 - [34] Ramsay, J. O. and B. W. Silvermann (1997). *Functional Data Analysis*. New York: Springer.
 - [35] Thodberg, H. H. (1996). A review of bayesian neural networks with an application to near infrared spectroscopy. *IEEE Trans. on Neural Networks*, 7(1), 56–72.
 - [36] Wülfert, F., W. Kok, and A. Smilde (1998). Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Anal. Chem.*, 70, 1761–1767.
 - [37] Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society, Series B*, 64(3), 363–410.
 - [38] Yao, Q. and H. Tong (1994). On subset selection in nonparametric stochastic regression. *Stat. Sin.*, 4, 51–70.
 - [39] Zhu, L. X. and K. T. Fang (1996). Asymptotics for kernel estimate of sliced inverse regression. *Annals of Statistics*, 24(3), 1053–1068.