Statistica Sinica **15**(2005), 665-683

SMOOTHED FUNCTIONAL INVERSE REGRESSION

Louis Ferré and Anne-Françoise Yao

Université Toulouse Le Mirail and Centre d'Océanologie de Marseille

Abstract: A generalization of Sliced Inverse Regression to functional regressors was introduced by Ferré and Yao (2003). Here we first address the issue of the identifiability of the Effective Dimension Reduction (EDR) space. Next, we estimate the covariance operator of the conditional expectation by means of kernel estimates. Consistency is proved and this extends the results of Zhu and Fang (1996) in the multivariate context to the functional case. We also suggest a new way for estimating the EDR Space for functional data which avoids inverting the covariance operator of the regressor. We apply our method to a prediction problem where the regressors are spectrometric curves.

Key words and phrases: Dimension reduction, functional data analysis, inverse regression, prediction.

1. Introduction

We consider the Tecator data set where the problem is one of predicting the fat content of pieces of meat from a near infrared absorbance spectrum. Fat content is evaluated by analytic chemistry with high cost, while infrared analysis is substantially cheaper. Achieving a good prediction of the fat content from the infrared analysis is an economic challenge for the Tecator company. The point is now that while the response variable Y, the fat content, is real, the regressor variable X, the spectrum, is a curve, see Figure 1. The abscissa correspond to the wavelength (100 channels) and the ordinates to the absorbance, that is $-\log_{10}$ of the transmittance measured by the spectrometer. The solution to this problem is clearly a functional regression with a real response. Functional data analysis is now an important research field with many applications. A substantial body of work has been developed in this area and an extensive review is given in Ramsay and Silverman (1997). In particular, functional regression has been investigated from the linear point of view (see e.g., Cardot, Ferraty and Sarda (1999)), but also by kernel nonparametric regression in Ferraty and Vieu (2002).

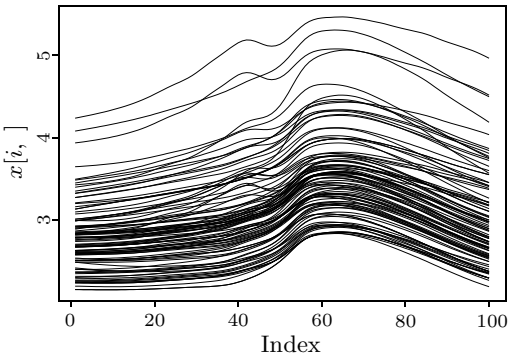

Figure 1. The regressor curves.

At the same time, Dauxois, Ferré and Yao (2001) have proposed a semi-parametric model for Hilbertian variables which corresponds, when the response is real (Ferré and Yao (2003)), to the functional version of Li's Sliced Inverse Regression (SIR) (Li (1991)). This model is:

$$Y = g\left(\int_a^b \theta_1(t)X(t), \ldots, \int_a^b \theta_D(t)X(t), \varepsilon\right), \tag{1}$$

where $\theta_1, \ldots, \theta_D$ are D functions in $L^2([a,b])$, the space of square integrable functions from $[a,b]$ into $\Re$, linearly independent and spanning a subspace E_D, ε is a real random variable independent of X and g is a function from $\Re^{D+1}$ to $\Re$. The space E_D is usually called the Effective Dimension Reduction (EDR) space.

In the multivariate context, SIR has led to numerous works either for determining the dimensionality, such as Li (1991), Schott (1994), Ferré (1998) and Velilla (1998), or for improving the method. In this latter case, different estimates of the covariance of the conditional mean have been proposed in Hsing and Carroll (1992), Zhu and Ng (1995) and Zhu and Fang (1996). Moreover, other methods have been proposed for estimating E_D: pHd (Li (1992)), SAVE (Cook (1991)), PIR (Bura and Cook (2001)), MAVE (Xia, Tong, Li and Zhu (2002)), estimation of the Central Mean Subspace (Cook and Li (2002)) and use of the conditional kth moment (Yin and Cook (2002)). In functional analysis, the reduction of dimensionality is sizeable since this sufficient subspace (sufficient since the relationship between Y and X only involves the projection of X onto E_D) is assumed finite.

Unlike the multivariate SIR, functional inverse regression has to face up to some technical difficulties. Let Γ (respectively Γ_e) be the covariance operator of X (resp. $E(X|Y)$). A previous problem arises since, under the assumptions made on X, the operator Γ is a Hilbert-Schmidt operator and not invertible as an operator from $L^2([a,b])$ to $L^2([a,b])$. By considering some restrictions, a

definition of an "inverse" operator, Γ^{-1}, is given and, by using the results of Dauxois et al. (2001), it can be proved that the EDR subspace contains the Γ-orthonormed eigenvectors of $\Gamma^{-1}\Gamma_e$ associated with the D positive eigenvalues. This is the generalization of Li (1991) on SIR to the infinite-dimensional case. The operator Γ^{-1} is unbounded so that the identifiability of the EDR space is not insured. Similarly to He, Muller and Wang (2003) for functional canonical analysis, we give a sufficient condition on X and Y that guarantees the existence of a basis of the EDR space. These issues are investigated in Section 2.

In Ferré and Yao (2003), the properties of Functional SIR are studied when Γ_e is estimated by slicing the range of Y. Here, we propose in Section 3 a new version of Functional SIR obtained by replacing the slicing by kernel smoother regressions: the Functional Inverse Regression (FIR). For multivariate regressors, this approach has been tackled by Zhu and Fang (1996), who show the consistency of the obtained estimate. Our purpose here is to show the consistency of the estimate in the functional context.

The fact that Γ^{-1} is unbounded leads to estimating it by an ill-conditioned matrix. While in Ferré and Yao (2003) a method of filtering is used to overcome this problem, we suggest a different way to provide a consistent estimate of E_D in Section 4.

Section 5 is devoted to applications. We first give simulations that show the efficiency in practice of FIR. Then, the spectrometric data set is treated. Our model is particularly well adapted to the problem mentioned above since predictions can be obtained after E_D and g have been estimated. Our approach is compared to several competitors, most of them relying on dimension reduction reached by means of PCA. The conclusion is that FIR leads to better results than PCA and that, combined with simple prediction methods, it can compete with more sophisticated approaches.

2. The Model

Let (X, Y) be a random variable that takes values in $H \times \Re$, where H is a functional space, for instance $H = L^2([a, b])$. Actually, our purpose extends to the more general context where H is a general Hilbert space with inner product $< \cdot, \cdot >$ and associated norm denoted by $||.||$. While all the proofs are given in this general context, the remainder of the paper may be read with H considered either as a Hilbert space or simply as a functional space.

We assume that X is centered, without loss of generality, and that

(A-1) $E\left(\|X\|^4\right) < +\infty$.

Under Assumption (A-1), the covariance operators of X and E(X|Y) exist and are denoted Γ and Γ_e.

(A-2) Γ is positive definite.

Under Assumption (A-1), both Γ and Γ_e are Hilbert-Schmidt operators. The operator Γ is therefore compact and is not invertible as defined from H to H. We denote by $(\delta_i)_{i=1,\ldots,\infty}$ the sequence of eigenvalues of Γ, $(u_i)_{i=1,\ldots,\infty}$ the sequence of associated eigenvectors and $(\pi)_{i=1,\ldots,\infty}$ the sequence of associated eigenprojectors. For convenience, for $(x,y) \in H \times H$, let $x \otimes y$ denote the tensor product operator from H to H defined as the operator which associates to any z in H, $(x \otimes y)(z) =< x, z > y$. Note that we have $\Gamma = E(X \otimes X)$. We let R_Γ be the range of Γ and $\mathrm{R}_\Gamma^{-1} = \{h \in H : h = \sum_{i=1}^{\infty} 1/\delta_i (u_i \otimes u_i)(f), f \in \mathrm{R}_\Gamma\}$. We have that (restricted to R_Γ^{-1}) Γ is a one-to-one mapping from R_Γ^{-1} onto R_Γ whose inverse, Γ^{-1}, is defined by $\Gamma^{-1} = \sum_{i=1}^{\infty} (1/\delta_i)\pi_i$.

We denote by $(\xi_i)_{i=1,\ldots,\infty}$ the coordinates of X on the Hilbertian basis $(u_i)_{i=1,\ldots,\infty}$. If $H = L^2([a,b])$, $X = \sum_{i=1}^{\infty} \xi_i u_i$ is the Karunen-Loeve decomposition of the stochastic process X.

Condition 1. *For any b in H, there exists a vector C in $\Re^D$ satisfying $E(< b, X > |B) = C'B$ with $B' = (< \theta_1, X >, \ldots, < \theta_D, X >)$.*

Using these notations, Model (1) becomes

$$Y = g(< \theta_1, X >, \ldots, < \theta_D, X >, \varepsilon),$$

and, under Condition 1, it can be shown that the EDR subspace contains the Γ-orthonormed eigenvectors of $\Gamma^{-1}\Gamma_e$ associated with the D positive eigenvalues. A basis of the EDR space is thus given by the eigenvectors of $\Gamma^{-1}\Gamma_e$, but the operator Γ^{-1} is not bounded and there is no guarantee that these eigenvectors exist in H.

(A-3) $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (1/\delta_i^2 \delta_j) E(E(\xi_i|Y) E(\xi_j|Y))^2 < \infty$.

Theorem 2.1. *Under* (A-1), (A-2) *and* (A-3), *the eigensubspace associated with the D positive eigenvalues of $\Gamma^{-1}\Gamma_e$ is well defined in H.*

The proof is given in Appendix 6.1. This theorem is the analogue of the results given in He et al. (2003) for functional canonical analysis.

3. Estimation of the Conditional Covariance Operator

Let (X_i, Y_i), be an i.i.d. sample, $i = 1, \ldots, n$. The estimation of the EDR space is obtained by replacing the covariance operators by suitable estimates just as in Li (1991), although here we do not use the same form for the conditional covariance operator. The operator Γ is estimated by $\Gamma_n = (1/n) \sum_{i=1}^{n} (X_i - \overline{X}) \otimes (X_i - \overline{X})$ where $\overline{X}$ is the empirical mean of X, and Γ_e remains to be estimated. We consider a kernel based estimate.

Assume that Y has a probability density f and let $r(\cdot) = E(X \mid Y = \cdot)$. The Nadaraya-Watson estimate of f is $\widehat{f}$ and, for any y in $\Re$, a kernel estimate of

$r(y)$ is given by

$$\widehat{r}(y) = \frac{\sum_{i=1}^{n} X_i K(\frac{Y_i - y}{h})}{\sum_{i=1}^{n} K(\frac{Y_i - y}{h})} = \frac{\hat{m}}{\hat{f}},$$

where h is the bandwidth and $K(.)$ an order k kernel. To avoid the effect of the small values of f, we consider $f_{e_n} = \max(f, e_n)$ and $\widehat{f}_{e_n} = \max(\widehat{f}, e_n)$ instead of f and $\widehat{f}$, $(e_n)_{n \in \aleph^*}$ being a sequence of numbers which tends to zero. Thus take $\widehat{r}_{e_n} = \widehat{m}/\widehat{f}_{e_n}$. We estimate Γ_e by the operator

$$\widehat{\Gamma}_e = \frac{1}{n} \sum_{i=1}^{n} \widehat{r}_{e_n}(Y_i) \otimes \widehat{r}_{e_n}(Y_i) - \overline{X} \otimes \overline{X},$$

which is the functional version of the Zhu and Fang (1996) estimate for a multivariate regressor. Now we make the following assumptions

(A-4) f and r belong to C^k;

(A-5) $K(\cdot)$ is an order k kernel with compact support;

(A-6) there exists d_1 and d_2 satisfying $\sup_y |f^{(k)}(y)| < d_1$ and $\sup_y ||r^{(k)}(y)|| < d_2$;

(A-7) $h \simeq n^{-c_1}$ and $e_n \simeq n^{-c_2}$, where c_1 and c_2 are positive scalars such that $(c_2/k) + (1/4k) < c_1 < 1/4 - c_2$;

(A-8) the function $\varphi(.) = E(||X||^2 | Y = .)$ is continuous;

(A-9) $\sqrt{n}\ E\left(\|r(Y)\|^2 I_{\{f(Y) < e_n\}}\right)$ tends to 0.

The consistency of $\hat{\Gamma}_e$ is ensured by the following theorem whose proof is given in Appendix 6.2 ($||.||_{hs}$ here denotes the Hilbert-Schmidt norm for operators).

Theorem 3.1. *Under* (1), (A-1) *and* (A-4) *to* (A-9), *we have* $||\widehat{\Gamma}_e - \Gamma_e||_{hs} = O_p(1/\sqrt{n})$.

A central limit theorem is also obtained and its proof is presented in Appendix 6.3:

Theorem 3.2. *Under* (1), (A-1) *and* (A-4) *to* (A-9), $\sqrt{n}(\widehat{\Gamma}_e - \Gamma_e)$ *converges in distribution, in the space of Hilbert-Schmidt operators, to a centered Hilbertian Gaussian variable with covariance operator* $\mathrm{Var}\,(r(Y) \otimes r(Y))$.

These results extend those obtained by Zhu and Fang (1996) in a multivariate context. In addition, the consistency of Γ_n is derived from the Law of Large Numbers for Hilbertian variables, see, e.g., Fortet (1995).

Theorem 3.1 offers an interesting perspective for functional inverse regression because determining the optimal bandwidths by cross-validation is an advantage of kernel covariance estimates over sliced covariance estimates. Indeed, even if the method is not sensitive to the number of slices, H, one can artificially reduce the dimension by taking too few slices. While this can be overcome in the multivariate

context by choosing H larger than or at least close to the number of regressors, no such guideline is available in the infinite-dimensional case and, unlike the kernel approach for the bandwidth, no automatic and optimal procedure for selecting H has been available.

4. Estimation of the EDR Space

In a multivariate context, the estimation of the EDR space does not pose any problem since Γ^{-1} is accurately estimated by the inverse of the empirical covariance matrix of X. Unfortunately, this is no longer true for functional inverse regression if we assume that Γ is a Hilbert-Schmidt operator: this inverse might be ill-conditioned. To overcome this difficulty Ferré and Yao (2003) suggest to replacing Γ_n by a generalized inverse of $\hat{\pi}_{k_n}\Gamma_n\hat{\pi}_{k_n}$, where $\hat{\pi}_{k_n}$ is the eigenprojector associated with the first k_n eigenvalues of Γ_n. The authors use an estimate of Γ_e based on slicing and they call the method Functional Sliced Inverse Regression (FSIR). Computation of FSIR relies on the choice of two parameters: the number of slices H and the value of k_n.

An alternative is derived from the following observation. Under our model, $\Gamma^{-1}\Gamma_e$ has finite rank. Then, it has the same eigen subspace associated with positive eigenvalues as $\Gamma_e^+\Gamma$, where Γ_e^+ is a generalized inverse of Γ_e. Now, we avoid the inversion of Γ by estimating the EDR space from the spectral decomposition of $\widehat{\Gamma}_e^+\Gamma_n$, where $\widehat{\Gamma}_e^+$ is a generalized inverse of $\widehat{\Gamma}_e$. The price to pay is that computing $\widehat{\Gamma}_e^+$ requires the knowledge of D.

Let $(\alpha_1, \ldots, \alpha_D)$ be the decreasing sequence of eigenvalues of $\Gamma_e^+\Gamma$, assumed to be distinct, and let $\beta_1, \ldots, \beta_D$ (respectively $\hat{\beta}_1, \ldots, \hat{\beta}_D$) be the eigenvectors, Γ -orthonormed, of $\Gamma_e^+\Gamma$ (resp., Γ_n -orthonormed of $\widehat{\Gamma}_e^+\Gamma_n$). From Theorem 3.1, and by using the consistency of Γ_n, it is easy to derive the consistency of $\widehat{\Gamma}_e^+\Gamma_n$. Then we deduce that $(1/\sqrt{n})||\widehat{\Gamma}_e^+\Gamma_n - \Gamma_e^+\Gamma||_{hs}$ is bounded in probability by a constant M. The consistency of the estimated EDR space is thus ensured by the following theorem.

Theorem 4.1. *Under* (1) *and* (A-1) *to* (A-9), if $(1/\sqrt{n}) \leq [(\min_{i=1,\ldots,D} \mid \alpha_i - \alpha_{i-1}|)/(2M)]$, then, for any $i = 1, \ldots, D$, $||\widehat{\beta}_i - \beta_i|| = O_p(1/\sqrt{n})$.

This result is a straightforward application of perturbation theory for linear operators, see, e.g., Kato (1966).

Determining D can be tackled in different ways and it depends on the goal of the analysis. If the purpose is prediction, the dimensionality can be treated as a parameter of the whole model and adjusted according to the performance of the prediction. This has been successfully experimented with in applications. But FIR can also be used in a descriptive way, for instance to provide relevant scatter plots; in that case it is necessary to estimate D *a priori*. We suggest

evaluating D by applying to Γ_e a criterion measuring the quality of estimation of its eigensubspaces by those of $\widehat{\Gamma}_e$. This criterion has been proposed to determine the dimensionality in multivariate SIR by Ferré (1998).

Let Π_q (respectively $\widehat{\Pi}_q$) be the eigenprojector associated with the q largest eigenvalues of Γ_e (resp. $\widehat{\Gamma}_e$), we consider the loss function

$$R(q) = 1 - \frac{1}{q} E(\mathrm{tr}(\Pi_q \widehat{\Pi}_q)).$$

Let $U_n = \sqrt{n}(\widehat{\Gamma}_e - \Gamma_e)$ and $(\lambda_i)_{i=1,\ldots,D+1}$ (respectively $(\widehat{\lambda}_i)_{i\in\aleph^*}$) be the increasing sequence of eigenvalues of Γ_e (respectively $\widehat{\Gamma}_e$) with $\lambda_{D+1} = 0$. We assume, to simplify, that the λ_i, $i = 1, \ldots, D$, are distinct and, for $i = 1, \ldots, D$, let b_i be the eigenvector associated with λ_i. An estimate of $R(q)$ is given by the following theorem.

Theorem 4.2. *If* $1/\sqrt{n} < \min_{i<j}(\lambda_i - \lambda_j)/2C$ *where* $||U_n||_\infty$ *is bounded in probability by* C, *if* X *satisfies Condition* (1), *if* (A-1) *and* (A-4) *to* (A-7) *are satisfied and if, for any* (i, j), $\widehat{\mathrm{Var}}(< r(Y), b_j >< r(Y), b_i >)$ *is the empirical variance of* $(< \hat{r}(Y), b_j >< \hat{r}(Y), b_i >)$, *then for* $q = 1, \ldots, D-1$,

$$\hat{R}(q) = \frac{1}{nq} \sum_{i=1}^{q} \sum_{j=q+1}^{D} \frac{\widehat{\mathrm{Var}}(< r(Y), b_j >< r(Y), b_i >)}{(\lambda_j - \lambda_i)^2}$$

is an estimate of $R(q)$ *which satisfies* $E(\hat{R}(q)) = R(q) + O(n^{-3/2})$.

This result is again a straightforward application of perturbation theory associated with elementary calculus in Hilbert spaces.

Actually, $\widehat{R}(q)$ depends on the unknown D. We suggest using the same strategy as in Ferré (1998): compute $\widehat{R}(q, D)$ for $D = 1, 2, \ldots$ and $q = 1, \ldots, d$, and retain the dimension D as the common value of q for which $\widehat{R}(q+1, d)$ clearly departs from zero.

5. Applications

5.1. Algorithm

We apply the following algorithm.

(1) Center X;
(2) compute Γ_n;
(3) compute the kernel estimate of $r(y)$, the bandwidth h being selected by cross-validation;
(4) compute $\widehat{\Gamma}_e$, the empirical variance of $r(Y)$;
(5) perform the eigenvalue decomposition of $\widehat{\Gamma}_e$;
(6) determine D by computing the $\widehat{R}(q, D)$'s;

(7) perform the eigenvalue decomposition of $(\Gamma_n^{1/2}(\widehat{\Pi}_D\widehat{\Gamma}_e\widehat{\Pi}_D)^+\Gamma_n^{1/2})^+$ where $\widehat{\Pi}_D$ is the eigen projector associated with the D largest eigenvalues of $\widehat{\Gamma}_e$;
(8) for each $i = 1, \ldots, D$, compute $\widehat{\beta}_i = \widehat{\alpha}_i(\widehat{\Pi}_D\widehat{\Gamma}_e\widehat{\Pi}_D)^+\Gamma_n^{1/2}\widehat{\eta}_i$, where $(\widehat{\alpha}_i)_{i=1,\ldots,D}$ are the D largest eigenvalues of $(\Gamma_n^{1/2}(\widehat{\Pi}_D\widehat{\Gamma}_e\widehat{\Pi}_D)^+\Gamma_n^{1/2})^+$ and $(\widehat{\eta}_i)_{i=1},\ldots,_D$ are the associated eigenvectors.

Note that in a prediction context, steps (5) and (6) can be omitted and step (7) can be replicated for several values of D.

5.2. Simulations

In this section, we report on some simulations that provide an insight in the behavior of our approach in practice. Our goal is to evaluate the ability of FIR to estimate the EDR space, but we also want to test how it behaves as a preliminary step for predictions.

Data were generated according to the following model:

$$Y = \sin(\frac{\pi}{2} < \theta_1, X >) + < \theta_2, X > +\varepsilon,$$

where X is a standard Brownian motion, ε is a $N(0,1)$ variable, and the EDR space was generated by the functions satisfying $\theta_1(x) = (2x-1)^3+1$ and $\theta_2(x) = \cos(\pi(2x-1))+1$.

The sample size was $n = 500$ and each curve X was evaluated at $p = 100$ points. Since prediction is one of our goals, the sample was divided into three parts: a training sample to estimate the EDR space (size 300), S_1; a monitoring sample (size 100) to select the best "parameters" of the prediction method, S_2; and a test sample to compute the prediction error, S_3. We performed 100 replications of the simulations in order to compute the average and standard error of the criteria used to evaluate the performance of the method.

To compute $\widehat{\Gamma}_e$ on S_1, a bandwidth selection was required and cross-validation returned values around 1.8 for each curve. This value was stable over the replications. We performed the analysis with the true value $D = 2$ and the erroneous $D = 3$. In Table 1, we give the corresponding eigenvalues for a single simulation, those values being very stable over the simulations. The two analyses are not nested and the first two eigenvalues are of course different. Moreover, the third eigenvalue in the second analysis is almost 0, indicating that the correct dimension of the model is indeed 2. This fact was also confirmed by the criterion $R(q)$: we found $\hat{R}(q, D) \leq 0.1$ for $D = 1, 2, 3$ and $q = 1, 2$ and $\hat{R}(3, 3) > 0.45$.

Table 1. Eigenvalues for the simulations.

k	1	2	3
$D = 2$	0.481	0.002	
$D = 3$	0.01	0.02	$6e^{-18}$

Next, to study the ability of our method to estimate the EDR space, we used $R^2(\hat{B})$, the squared trace correlation between the EDR space and its estimates, i.e., the average of the squared coefficients between $< \hat{\theta}_1, X >$, $< \hat{\theta}_2, X >$ and $< \theta_1, X >$, $< \theta_2, X >$. Table 2 reports its averaged values over the 100 simulations and it shows that our method yielded rather good results. Moreover, FIR worked better than FSIR (this method used the optimal $k_n = 2$), even if the variance of $R^2(\hat{B})$ was larger for FIR. To confirm this, we observed that in 84% of the simulations FIR outperformed FSIR.

Table 2. Average over the 100 replications of $R^2(\hat{B})$ and SEP (standard error in parenthesis.

	FIR	FSIR
$R^2(\hat{B})$	0.925	0.891
	(0.060)	(0.016)
SEP	0.382	0.464
	(0.070)	(0.042)

Finally, we compared the predictive performance of FIR and FSIR. Here, the dimension of the EDR space was known to be two and the amount of data being rather large, we selected a bivariate Nadaraya-Watson kernel smoother to estimate g. Computed on S_2, the optimal bandwidth was around 1 for every replication. Finally, we computed the Standard Error of Prediction, SEP, on S_3, for the bivariate kernel smoother applied to the indices $(< \hat{\theta}_1, X >, < \hat{\theta}_2, X >)$ of FIR and FSIR. The averages of the SEP over the hundred replications are also reported in Table 2. On the average, FIR provided better results than FSIR and we also observe that FIR outperformed FSIR for 85% of the replications. Finally, we tested the effects of the parameters D and k_n for FIR and FSIR by also computing the SEP for the erroneous values $D = 3$ and $k_n = 3$. We got, respectively, 0.46 and 0.53 for the averaged SEP which points out the crucial role of those parameters.

5.3 Application to spectrometric data

We return to our initial problem. We have $n = 215$ i.i.d. observations (X_i, Y_i) of the couple (X, Y), where we recall that X is the spectrum of absorbance discretized at one hundred points and Y is the lipid rate. This data set has been treated in several ways and, in order to compare efficiently our results with those analyses, we have adopted the same scheme as Thodberg (1995): the sample has been divided into a training sample, S_{11} (size 129) used to estimate the EDR space, a monitoring sample S_{12} (size 43) used to estimate the function g, and a

testing sample, S_2 (size 43) used to compute the Standard Error of Prediction (SEP).

The computation of $\hat{\Gamma}_e$ was performed from S_{11} and for each inverse regression the bandwidth selected by cross-validation was around 2. The estimation of $\widehat{R}(q)$ led to a ten-dimensional solution since all $\hat{R}(q, D)$ were close to 0 for any $q \leq 10$ and they presented a gap for $q > 10$ when $D > 10$. The eigenvalues of FIR are given in Figure 2. We also give in this Figure the eigenvalues obtained by FSIR for $H = 20$ and $k_n = 10$. This latter value was selected after several trials as the one which provided the best SEP. Note also that the ten eigenvalues of FIR are very close to the first ten eigenvalues of FSIR. We give in Figure 3 the ten curves that span the estimated EDR space.

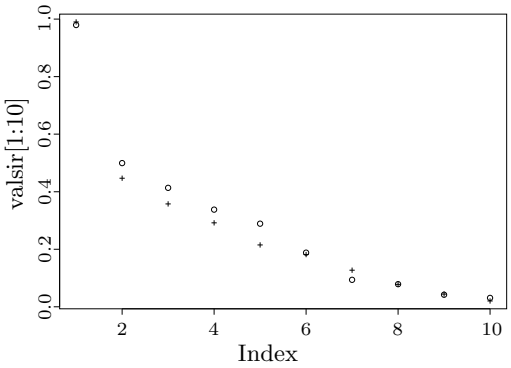

Figure 2. Eigenvalues of FIR (cross)and FSIR (circle).

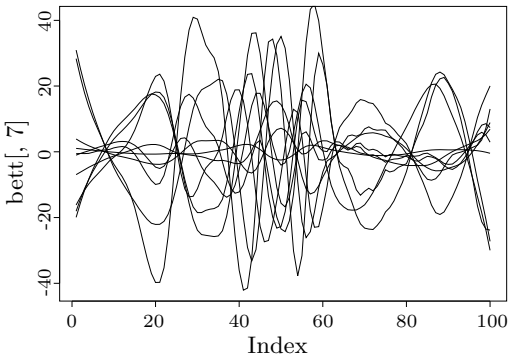

Figure 3. The ten curves spanning the EDR space.

Unfortunately, any nonparametric method like kernel regression, spline regression or wavelet regression, is excluded from estimating g (recall that we deal

with a ten-dimensional regressor but only 43 observations in S_{12} for prediction). A method insensitive to the curse of dimensionality is required. Multi-layer perceptrons were used since, due to their low Vapnick-Chernonenkis dimension (Vapnick (1998)), they are not affected by this curse.

Then, a Multi-layer Neural Network with the projections of X onto the space E_D as input and Y as output was computed. Actually, Ferré and Villa(2005) show that any functional inverse regression (it can be FSIR, FIR, or any other method performing functional inverse regression) can be used as a relevant first stage to compute networks with functional inputs: the functional data are first projected onto a convenient basis (spanning the EDR space) before a (multivariate) Neural Network is applied. They also show that the estimated weights converge to the optimal weights of the networks. Previously, other projections have been suggested. For instance, in Conan-Guez and Rossi (2002), spline basis, orthonormal polynomial, Fourier basis, are considered; but the main drawback is that one does not know which elements of the basis are relevant. In order to overcome this problem, the basis can be extracted from the data themselves by means of Principal Component Analysis (PCA) and this is the choice of Borggard and Thodberg (1992) and Thodberg (1995). Unfortunately, this only depends on X while "conditional" information is required here, and this is exactly what FIR does.

The training sample S_{11} was used by Borggard and Thodberg (1992) and Thodberg (1995) to compute the first ten principal components without justification of the choice of this dimension. Those ten multivariate regressors were used to perform linear regression (lm) and several Neural Networks: a classical six hidden unit multi-layer perceptron (NN1) and two Bayesian neural networks (NN2) and (NN3), using previous knowledge on the data set (see Thodberg (1995) for details). When FIR is used to estimate the EDR space, the optimal architecture is found to have five hidden units. Note that we have performed the analysis with several values of D and several architectures for the network, but the lower SEP has been reached for a five hidden unit network and $D = 10$ (this confirms that ten is indeed the correct dimension). On the other hand, we have built different networks that reached the best SEP for NN1 when the number of PC's is larger than ten.

We give in Table 3 the values of the SEP for the methods mentioned above. Let us first examine the SEP computed for (lm), the linear regression on the ten PCs, and the linear regression (flm) on the projection of X onto the EDR space. The results are very close. This is explained by the fact that, for linear regression, the prediction based on X is correlated with the prediction based on the projection of X onto the EDR space (simple calculations prove it). The slight difference observed here results from the 'approximation' of X by the ten

PCs. Anyway, linear regression is not a suitable solution here. We also include in the comparison the functional nonparametric regression (fnp) of Ferraty and Vieu (2002), based on a Nadaraya-Watson estimate computed from a semi-norm in the regressor space. Even if it outperforms the linear regressions, (fnp) is still far from obtaining the best results.

Table 3. Standard Error of Prediction (SEP) for the compared methods.

Method	Linear Reg.	Non-parametric	**FIR +Linear**	**FSIR +N.N.**
SEP	2.79	2.13	2.75	0.56
Method	N.N. early stopping	N.N. with prunning	N.N. with committee	**FIR +N.N.**
SEP	0.65	0.55	0.52	0.55

The best performances are reached by Neural Networks; note that all these methods provide rather close results. Particularly, using FIR or FSIR only makes a slight difference in favor of FIR. Neither of them outperforms the Bayesian approach based on committee, but the combination of FIR and Neural Networks provides results similar to the Bayesian network with pruning based on a single evidence. Thus, it can compete with sophisticated methods while being simple and requiring much less computational time. Our feeling is that the results could have been improved by combining FIR with a Bayesian network, but this is beyond the scope of the paper. The main point here is that FIR combined with neural networks works better than PCA combined with neural networks: since the same type of neural networks is involved, the improvement only results from the performance of FIR.

6. Appendix-Technical Results

6.1. Proof of Theorem 2.1

To get the Γ-orthonormed eigenvectors of $\Gamma^{-1}\Gamma_e$, $\beta_1, \ldots, \beta_D$, it is convenient to determine $\eta_1, \ldots, \eta_D$, the orthonormed eigenvectors of $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$, and to use, for any $d = 1, \ldots, D$, $\beta_d = \Gamma^{-1/2}\eta_d$.

The proof the theorem is close to the one used by He et al. (2003) to define the canonical functions in functional canonical analysis, we use the same definitions. We denote by F_{XX}, the range of $\Gamma^{1/2} = \sum_{i=1}^{\infty} \delta_i^{1/2} u_i \otimes u_i$. This is $F_{XX} = \{u \in H, \sum_{i=1}^{\infty}(1/\delta_i) < u, u_i >^2 < \infty\}$. As an operator from $F_{XX}^{-1} = \{h \in H, h = \sum_{i=1}^{\infty}(1/\delta_i)(u_i \otimes u_i)u, u \in F_{XX}\}$ onto F_{XX}, $\Gamma^{1/2}$ is a one to one mapping whose inverse is denoted by $\Gamma^{-1/2}$. We first prove the existence of the vectors $\eta_d's$.

For that, we show that, for any vector $u \in \mathrm{F}_{XX}$, $\Gamma_e\Gamma^{-1/2}u \in \mathrm{F}_{XX}$. From the definition of $\Gamma^{1/2}$, we have

$$\Gamma^{-1/2} = \sum_{i=1}^{\infty} \frac{1}{\delta_i^{1/2}} u_i \otimes u_i$$

and, from the decomposition of X on the basis $(u_i)_{i=1,\ldots,\infty}$,

$$\Gamma_e = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} E(E(\xi_i|Y)E(\xi_j|Y))(u_i \otimes u_j).$$

Thus, we have

$$\Gamma_e\Gamma^{-1/2}u = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \frac{E(E(\xi_i|Y)E(\xi_j|Y))}{\delta_i^{1/2}}(u_i \otimes u_j)(u),$$

$$\sum_{i=1}^{\infty} \frac{1}{\delta_i} < \Gamma_e\Gamma^{-1/2}u, u_i >^2 \leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \frac{E(E(\xi_i|Y)E(\xi_j|Y))^2}{\delta_i\delta_j} \parallel u \parallel^2,$$

by using the Cauchy-Schwarz inequality.

But (A-3) implies $\sum_{i=1}^{\infty}\sum_{i=1}^{\infty}(1/\delta_i\delta_j)E(E(\xi_i|Y)E(\xi_j|Y))^2 < \infty$. Then Γ_e $\Gamma^{-1/2}u \in \mathrm{F}_{XX}$, for any u, and the operator $\Gamma^{-1/2}\Gamma\Gamma^{-1/2}$ is well defined.

It remains to prove that $\eta_d \in \mathrm{F}_{XX}$ for any d. Since η_d is an eigenvector of $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$, we just have to show that $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}\eta_d$ belongs to F_{XX}. We have, for any d,

$$\sum_{i=1}^{\infty} \frac{1}{\delta_i} < \Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}\eta_d, u_i >^2$$

$= \sum_{i=1}^{\infty} \frac{1}{\delta_i}(\sum_{j=1}^{\infty} \frac{E(E(\xi_j)E(\xi_i))}{\delta_j^{1/2}\delta_i^{1/2}} < u_j, \eta_d >)^2 \leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \frac{E(E(\xi_i)E(\xi_j))^2}{\delta_i^2\delta_j} < \infty$ by the Cauchy-Schwarz inequality and assumption (A-3), which completes the proof.

6.2. Proof of Theorem 3.1

Recall that $< ., . >_{hs}$ is the inner product in the space of Hilbert-Schmidt operators and, for any bounded operator A, take $||A||_\infty = \sup_{||x||=1} ||Ax||$.

In the following, $m * f$ denotes the convolution product of m and f. To simplify, we set $K(\frac{\cdot}{h}) = K_h(.)$.

Lemma 6.1. (Rao (1983)) *Under assumptions* (A-4), (A-5) *and* (A-6), *we have* $\sup_y |\widehat{f}(y) - f(y)| = O_p(h^k + (\sqrt{\log n}/\sqrt{n}h))$ *and* $\sup_y |(f * K_h(y)/h) - f(y)| = O(h^k)$.

Lemma 6.2. (Yao (2001) *Under assumptions* (A-4) *to* (A-6), $\sup_y \|\widehat{m}(y) - m(y)\| = O_p(h^k + \sqrt{\log n}/\sqrt{n}h)$ and $\sup_y ||(m * K_h(y)/h) - m(y)|| = O(h^k)$.

To prove Theorem 3.1, we set $r_{e_n}(y) = m(y)/f_{e_n}(y)$, $\overline{\Gamma}_e = \frac{1}{n}\sum_{i=1}^n r_{e_n}(Y_i) \otimes r_{e_n}(Y_i)$ and $\widetilde{\Gamma}_e = \frac{1}{n}\sum_{i=1}^n r(Y_i) \otimes r(Y_i)$. Then

$$\widehat{\Gamma}_e - \Gamma_e = (\widehat{\Gamma}_e - \overline{\Gamma}_e) + (\overline{\Gamma}_e - \widetilde{\Gamma}_e) + (\widetilde{\Gamma}_e - \Gamma_e). \tag{2}$$

First we show that

$$\overline{\Gamma}_e - \widetilde{\Gamma}_e = o_p(\frac{1}{\sqrt{n}}). \tag{3}$$

Some calculus, using the relation $(1/f^2_{e_n}(Y)) = (1/f^2(Y))I_{\{f(y)\geq e_n\}} - (1/e_n^2)$ $I_{\{f(y)<e_n\}}$, leads to

$$E\left(\left\|\sqrt{n}\left[\overline{\Gamma}_e - \widetilde{\Gamma}_e - E\left(\overline{\Gamma}_e - \widetilde{\Gamma}_e\right)\right]\right\|^2_{hs}\right) \leq E(\|r(y)\otimes r(y)\|^2_{hs}\, I_{\{f(y)<e_n\}}).$$

Since $E(\|r(y)\otimes r(y)\|^2_{hs} I_{\{f(y)<e_n\}})$ tends to 0 by dominated convergence, Tchebychev's inequality gives $\overline{\Gamma}_e - \widetilde{\Gamma}_e = E(\overline{\Gamma}_e - \widetilde{\Gamma}_e) + o_p(1/\sqrt{n})$. But, $\sqrt{n}\|E(\overline{\Gamma}_e - \widetilde{\Gamma}_e)\|_{hs} \leq \sqrt{n}\, E(\|r(Y) \otimes r(Y)\|_{hs} I_{\{f(Y)<e_n\}})$ which tends to zero according to assumption (A-9), and (3) is proved.

Secondly, we prove that

$$\widehat{\Gamma}_e - \overline{\Gamma}_e = o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4}$$

For that, we set $\widehat{\Gamma}_e - \overline{\Gamma}_e = S_1 + S_2 + S_3$, where $S_1 = (1/n)\sum_{i=1}^n (\widehat{r}_{e_n}(Y_i) - r_{e_n}(Y_i)) \otimes (\widehat{r}_{e_n}(Y_i) - r_{e_n}(Y_i))$, $S_2 = (1/n)\sum_{i=1}^n (\widehat{r}_{e_n}(Y_i) - r_{e_n}(Y_i)) \otimes r_{e_n}(Y_i)$ and $S_3 = (1/n)\sum_{i=1}^n r_{e_n}(Y_i) \otimes (\widehat{r}_{e_n}(Y_i) - r_{e_n}(Y_i))$.

But we have

$$S_1 = o_p(\frac{1}{\sqrt{n}}). \tag{5}$$

Indeed, for all $i \in \aleph^*$,

$$\widehat{r}_{e_n}(Y_i) - r_{e_n}(Y_i) = \frac{r_{e_n}(Y_i)}{\widehat{f}_{e_n}(Y_i)}\left(f_{e_n}(Y_i) - \widehat{f}_{e_n}(Y_i)\right) + \frac{1}{\widehat{f}_{e_n}(Y_i)}\left(\widehat{m}(Y_i) - m(Y_i)\right), \tag{6}$$

$|\widehat{f}_{e_n}(Y_i) - f_{e_n}(Y_i)| \leq |\widehat{f}(Y_i) - f(Y_i)| \leq \sup_y |\widehat{f}(y) - f(y)|$ and $||r_{e_n}(Y_i)|| \leq ||r(Y_i)||$. By using the properties of the Hilbert-Schmidt norm of tensor products, we get $||S_1||_{hs} \leq (1/n)\sum_{i=1}^n \|\hat{r}_{e_n}(Y_i) - r_{e_n}(Y_i)\|^2$; thus

$$\begin{aligned}
\sqrt{n}||S_1||_{hs} \leq{}& \frac{\sqrt{n}}{e_n^2}(\frac{1}{n}\sum_{i=1}^n ||r(Y_i)||^2) \sup_y \left|\widehat{f}(y) - f(y)\right|^2 \\
&+ \frac{\sqrt{n}}{e_n^2} \sup_y \|\widehat{m}(y) - m(y)\|^2 \\
&+ \frac{2\sqrt{n}}{e_n^2}(\frac{1}{n}\sum_{i=1}^n ||r(Y_i)||) \sup_y \left|\widehat{f}(y) - f(y)\right| \sup_y \|\widehat{m}(y) - m(y)\|
\end{aligned}$$

and $\sqrt{n}||S_1||_{hs} = O_p((n^{-kc_1} + (\sqrt{\log n}/\sqrt{n}h))^2 n^{2c_2+1/2})$ by assumptions (A-4) to (A-7) and Lemmas 6.1 and 6.2, and (5) is obtained by assumption (A-9).

We also have

$$S_2 = o_p(\frac{1}{\sqrt{n}}). \tag{7}$$

Write $S_2 = (U_1 - \overline{\Gamma}_e) - (U_2 - \overline{\Gamma}_e) - U_3 + U_4$, where

$$\begin{aligned}
U_1 &= \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{m}(Y_i)}{f_{e_n}(Y_i)} \otimes r_{e_n}(Y_i),\\
U_2 &= \frac{1}{n}\sum_{i=1}^{n} r_{e_n}(Y_i) \otimes r_{e_n}(Y_i)\frac{\widehat{f}(Y_i)}{f_{e_n}(Y_i)},\\
U_3 &= \frac{1}{n}\sum_{i=1}^{n} r_{e_n}(Y_i) \otimes r_{e_n}(Y_i)\frac{\widehat{f}_{e_n}(Y_i) - \widehat{f}(Y_i)}{f_{e_n}(Y_i)},\\
U_4 &= \frac{1}{n}\sum_{i=1}^{n} (\widehat{m}(Y_i) - m(Y_i)) \otimes r_{e_n}(Y_i)\frac{\widehat{f}_{e_n}(Y_i) - f_{e_n}(Y_i)}{\widehat{f}_{e_n}(Y_i) f_{e_n}(Y_i)}\\
&\quad + \frac{1}{n}\sum_{i=1}^{n} r_{e_n}(Y_i) \otimes r_{e_n}(Y_i)\frac{(\widehat{f}_{e_n}(Y_i) - f_{e_n}(Y_i))^2}{\widehat{f}_{e_n}(Y_i) f_{e_n}(Y_i)}.
\end{aligned}$$

A proof, similar to the one used to get (5), leads to $U_4 = o_p(1/\sqrt{n})$.

Furthermore, it can easily be shown that $||U_3||_{hs} \leq (2/n)\sum_{i=1}^{n} ||r(Y_i)|| I_{\{\widehat{f}(Y_i)<e_n\}}$ and, since $I_{\{\widehat{f}(Y_i)<e_n\}} \leq I_{\{f(Y_i)<2e_n\}} + (\sup_y |\widehat{f}(y) - f(y)|^2/e_n^2)$, we get

$$||U_3||_{hs} \leq \frac{2}{n}\sum_{i=1}^{n} ||r(Y_i)||(I_{\{f(Y_i)<2e_n\}} + \frac{\sup_y |\widehat{f}(y) - f(y)|^2}{e_n^2}).$$

Hence, $U_3 = o_p(1/\sqrt{n})$ by (A-4) to (A-7), (A-9), and Lemma 6.1. We next show

$$U_1 - \overline{\Gamma}_e = \overline{\Gamma}_e - E(\overline{\Gamma}_e) + o_p(\frac{1}{\sqrt{n}}). \tag{8}$$

Let us set $U_1 - \overline{\Gamma}_e = T + R$ where

$$T = \frac{1}{n^2 h}\sum_{i=1}^{n} X_i \otimes \frac{r_{e_n}(Y_i)}{f_{e_n}(Y_i)} K(0),$$

$$R = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j\neq i} X_j K_h(Y_j - Y_i) \otimes \frac{r_{e_n}(Y_i)}{f_{e_n}(Y_i)} - \overline{\Gamma}_e.$$

Since $||T||_{hs} \leq (K(0)/(nhe_n))[(1/n)\sum_{i=1}^{n} ||X_i||||r(Y_i)||]$, we get $T = O_p(n^{-1+c_1+c_2})$ $= o_p(1/\sqrt{n})$.

It remains to show now that $R = \overline{\Gamma}_e - E(\overline{\Gamma}_e) + o_p(1/\sqrt{n})$. For that, we set $R_1 = ((n-1)/n)(1/(nh))\sum_{i=1}^n m * K_h(Y_i) \otimes ((r_{e_n}(Y_i))/(f_{e_n}(Y_i)))$, $R_2 = (1/(n^2h))\sum_{i=1}^n \sum_{j\neq i} X_j K_h(Y_j - Y_i) \otimes ((r_{e_n}(Y_i))/(f_{e_n}(Y_i)))$ and we verify that

$$R_2 - R_1 - (\overline{\Gamma}_e - E(\overline{\Gamma}_e)) = o_p(\frac{1}{\sqrt{n}}). \tag{9}$$

Take, for all $i \in \aleph^*$, $\alpha_{e_n}(Y_i) = r_{e_n}(Y_i)/f_{e_n}(Y_i)$. Then we have

$$E(||R_2 - R_1||_{hs}^2) = \frac{1}{n^4h^2} E\Big(|| \sum_{i=1}^n (\sum_{j\neq i} (X_j K_h(Y_j - Y_i) - m * K_h(Y_i))\Big) \otimes \alpha_{e_n}(Y_i)||_{hs}^2)$$

and

$$E(||R_2 - R_1||_{hs}^2) = M + Q, \tag{10}$$

with

$$M = \frac{1}{n^4h^2} \sum_{i=1}^n E(||\Big(\sum_{j\neq i} (X_j K_h(Y_j - Y_i) - m * K_h(Y_i))\Big) \otimes \alpha_{e_n}(Y_i)||_{hs}^2),$$

$$Q = \frac{1}{n^4h^2} \sum_{l=1}^n \sum_{p\neq l} E\Big(< \sum_{j\neq l} (X_j K_h(Y_j - Y_l) - m * K_h(Y_l))) \otimes \alpha_{e_n}(Y_l)$$
$$\sum_{j\neq p} (X_j K_h(Y_j - Y_p) - m * K_h(Y_p))) \otimes \alpha_{e_n}(Y_p) >_{hs} \Big).$$

Properties of conditional expectation, the Dominated Convergence Theorem, and the inequality $||\alpha_{e_n}(Y)||^2 \leq ||r(Y)||^2/f(Y)e_n$, yield $M \leq (1/(n^2he_n))$ $E((||r(Y)||^2/f(Y))(((\varphi f) * K_h^2(Y))/h)$. Next, (A-8), the Dominated Converge Theorem and the Cauchy-Schwarz inequality lead to for $y \in \Re$, $(1/h)(||r(y)||^2/$ $f(y))((\varphi f) * K_h^2(y)) \leq (||r(y)||^2/f(y))E(\varphi^2 f(Y))^{1/2}(\int_{-1}^1 K^4(t)dt)^{1/2}$. From the Dominated Convergence Theorem and (A-8) and (A-7), we deduce

$$M = o(\frac{1}{n}). \tag{11}$$

Now one can write

$$Q = I + II, \tag{12}$$

$$I = \frac{2}{n^4h^2} \sum_{l=1}^n \sum_{p\neq l} E(< (X_p K_h(Y_p - Y_l) - m * K_h(Y_l)) \otimes \alpha_{e_n}(Y_l)$$
$$\sum_{j\neq p} (X_j K_h(Y_j - Y_p) - m * K_h(Y_p)) \otimes \alpha_{e_n}(Y_p) >_{hs}),$$

$$II = \frac{1}{n^4h^2} \sum_{l=1}^n \sum_{p\neq l} \sum_{j\neq l, j\neq p} E(< (X_j K_h(Y_j - Y_l) - m * K_h(Y_l)) \otimes \alpha_{e_n}(Y_l)$$
$$\sum_{m\neq p} (X_m K_h(Y_m - Y_p) - m * K_h(Y_p)) \otimes \alpha_{e_n}(Y_p) >_{hs}).$$

Using properties of conditional expectation, we get $|I| \leq (2/(n^2h^2))E((\varphi f) * K_h^2(Y)||\alpha_{e_n}(Y)||^2)$ which leads to

$$I = o(\frac{1}{n}) \tag{13}$$

by (A-1), (A-7) and (A-8).

It can easily be checked that

$$\begin{aligned} II = &\frac{1}{nh^2}(E(||X \otimes (\alpha_{e_n} f) * K_h(Y)||_{hs}^2) \\ &-\frac{1}{nh^2}||E(m * K_h(Y)) \otimes \alpha_{e_n}(Y))||_{hs}^2) + o(\frac{1}{n}). \end{aligned} \tag{14}$$

Thus (10), (11), (13) and (14) yield

$$\begin{aligned} E(||R_2 - R_1||_{hs}^2) = &\frac{1}{nh^2}(E(||X \otimes (\alpha_{e_n} f) * K_h(Y)||_{hs}^2) \\ &-\frac{1}{nh^2}||E(g * K_h(Y)) \otimes \alpha_{e_n}(Y))||_{hs}^2) + o(\frac{1}{n}). \end{aligned} \tag{15}$$

But we also have

$$\begin{aligned} E(<\overline{\Gamma}_e - E(\overline{\Gamma}_e), R_2 - R_1>_{hs}) = &\frac{1}{nh}E(<m(Y) \otimes \alpha_{e_n}(Y), X \otimes (\alpha_{e_n} f) * K_h(Y)>_{hs}) \\ -\frac{1}{nh} <& E(m(Y) \otimes \alpha_{e_n}(Y)),\, E(g * K_h(Y) \otimes \alpha_{e_n}(Y)) >_{hs} + o(\frac{1}{n}). \end{aligned} \tag{16}$$

Thus, $E(||R_2 - R_1 - (\overline{\Gamma}_e - E(\overline{\Gamma}_e))||_{hs}^2) \leq (1/n)E(||X \otimes (\alpha_{e_n} f) * ((1/h)K_h(Y)) - m(Y) \otimes \alpha_{e_n}(Y)||_{hs}^2) + o(1/n)$ and we get $E(||R_2 - R_1 - (\overline{\Gamma}_e - E(\overline{\Gamma}_e))||_{hs}^2) = o(1/n)$, so (9) follows in the same way as (11). Now, since $R_1 - \overline{\Gamma}_e = R_1 - ((n-1)/n)\overline{\Gamma}_e + o_p(1/\sqrt{n})$, by the Dominated Convergence Theorem, we get $R_1 - ((n-1)/n)\overline{\Gamma}_e = o_p(1/\sqrt{n})$. Then we deduce $R = \overline{\Gamma}_e - E(\overline{\Gamma}_e) + o_p(1/\sqrt{n})$ and (8) is proved.

By substituting m by f and $\widehat{m}$ by $\widehat{f}$, we similarly prove that $U_2 - \overline{\Gamma}_e = \overline{\Gamma}_e - E(\overline{\Gamma}_e) + o_p(1/\sqrt{n})$ and finally get (7) by using (8), $U_3 = o_p(1/\sqrt{n})$ and $U_4 = o_p(1/\sqrt{n})$.

To get (4), it is enough to notice that the proof of $S_3 = o_p(1/\sqrt{n})$ is similar to that of (7).

Combining (2), (3) and (4) leads to $\widehat{\Gamma}_e - \Gamma_e = \tilde{\Gamma}_e - \Gamma_e + o_p(1/\sqrt{n})$. But, by the Weak Law of Large Number, we also have $\tilde{\Gamma}_e - \Gamma_e = O_p(1/\sqrt{n})$ and the proof is complete.

6.3. Proof of Theorem 3.2

From the previous proof, we have $\widehat{\Gamma}_e - \Gamma_e = \tilde{\Gamma}_e - \Gamma_e + o_p(1/\sqrt{n})$, and it is enough to use the Central Limit Theorem for Hilbertian variables (Fortet (1995) and Slusky's theorem to complete the proof.

Acknowledgement

The authors are grateful to the Co-Editors and the two referees for their advisable comments.

References

Borggaard, C. and Thodberg, H. H. (1992). Optimal minimal neural interpretation of spectra. *Analytic Chemistry* **64**, 545-551.

Bura E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric model. *J. Roy. Stat. Soc. B.* **63**, 393-410.

Cardot, H., Ferraty, F. and Sarda P.(1999). Functional Linear Model. *Statist. Probab. Lett.* **45**, 11-22.

Conan-Guez, B. and Rossi, F. (2002). Approche régularisée du traitement de données fonctionnelles par un perceptron multicouches. *Actes des neuvièmes journèes de la SFC, Toulouse*, 169-172.

Cook, R. D. (1991). Discussion of Li (1991) *J. Amer. Statis. Assoc.* **86**, 328-332.

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Dauxois, J., Ferré, L. and Yao, A. F. (2001). Un modèle semi-paramétrique pour variable aléatoire hilbertienne. *C. R. Acad. Sci. Paris* **t.327**, série I, 947-952.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist.* **17**, 545-564.

Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *J. Amer. Statist. Assoc.* **93**, 132-140.

Ferré, L. and Villa, N. (2005). Multi-layer neural network with functional Inputs: an inverse regression approach. Preprint.

Ferré, L. and Yao, A. F. (2003). Functional Sliced Inverse Regression analysis. *Statistics* **37**, 475-488.

Fortet, R. (1995). *Vecteurs, Fonctions et Distributions Aléatoires dans les Espaces de Hilbert.* Hermes, Paris.

He G., Müler H. G. and Wang, J. L. (2003). Functional canonical analysis for square integrable stochastic processes. *J. Multivariate Anal.* **85**, 54-77.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20**, 1040-1061.

Kato, T. (1966). *Perturbation Theory for Linear Operators.* 2nd edition. Springer Verglag, New York.

Li, K. C. (1991). Sliced Inverse Regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualisation and dimension reduction: another application of Stein's lemma. *Ann. Statist.* **87**, 1025-1039.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis.* Springer Verlag, New York.

Rao, B. L. S. P. (1983). *Nonparametric Functional Estimation.* Academic Press, Orlando, FL.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89,** 141-148.

Thodberg, H. H. (1995). A review of Bayesian Neural Networks with an application to near infrared spectroscopy. *IEEE Trans. Neural. Networks* **7,** 56-72.

Vapnick, V. N. (1998). *The Nature of Statistical Learning,* 2nd edition. Springer Verglag, New York.

Velilla, S. (1998). Assessing the number of linear components in a general regression problem. *J. Amer. Statist. Assoc.* **93,** 1088-1098.

Xia, Y., Tong H., Li, W. K. and Zhu, L. X. (2002). An adaptative estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 1-28.

Yao, A. F. (2001). *Un Modèle Semi-Paramétrique pour Variables Fonctionnelles: la Régression Inverse Fonctionnelle,* Thèse Université Toulouse III.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment *J. Roy. Statist. Soc. Ser. B* **64**, 159-175.

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053-1068.

Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5**, 727-736.

Equipe GRIMM, Université Toulouse le Mirail, 5, Allées Antonio Machado, 31058 Toulouse, France

E-mail: loferre@univ-tlse2.fr

Centre d'Océanologie de Marseille, Campus de Luminy, 13288 Marseille, France.

E-mail: yao@com.univ-mrs.fr

(Received June 2003; accepted May 2004)