Multi-Layer Neural Network with functional inputs: an inverse regression approach

Louis Ferré Nathalie Villa
Université Toulouse Le Mirail, Toulouse, France [*]

Abstract

Functional data analysis is a growing research field since more and more pratical applications involve functional data. In this paper, we focus on the problem of regression and classification with functional predictors: the model suggested combines an efficient dimension reduction procedure (functional SIR, first introduced by Ferré and Yao (2003)), for which we give a regularized version, and the accuracy of a neural network. The consistency of the model is proved and the method is successfully confronted to real life data.

Keywords: Classification, Dimension Reduction, Functional Data Analysis, Multi-layer Perceptron, Prediction.

1 Introduction

Functional regression is now a very important part of statistics as functional variables occur frequently in practical applications. We present two examples that take place in this area. First, we face a regression problem where the regressor are curves (see Figure 1): the Tecator data problem consists in predicting the fat content of pieces of meat from a near absorbance spectrum. This data set has already been studied by Thodberg (1995) and Ferré and Yao (2003).

[Figure 1 about here.]

Secondly, in the phoneme data set, the data are log-periodograms of a 32 ms duration corresponding to recorded speakers and we expect to determine which one of the five phonemes, [sh] as in "she", [dcl] as in "dark", [iy] as in "she", [aa] as in "dark" and [ao] as in "water", corresponds to this recording. It has already been described by Hastie, Buja and Tibschirani (1995) and by Ferraty and Vieu (2003). Clearly, here, functional data is also involved but we face now a classification problem. However, we will see that both - regression and classification - can be tackled via a common modelling.

[*] *Correspondence Address:* N. Villa, Département de mathématiques et informatique, Université Toulouse Le Mirail, 5 allées A. Machado, 31058 Toulouse cedex 1, FRANCE. Email: villa@univ-tlse2.fr

An extensive review of the numerous studies developped for functional data analysis can be found in Ramsay and Silverman (1996) including regression and classification but also many factorial methods. A particularity of functional regression is that it leads to ill-posed problems because of the infinite dimension of the feature space. Then original solutions have been introduced to overcome this problem for the functional linear regression, see e.g. Cardot et al. (1996) or the nonparametric regression, Ferraty and Vieu (2002). At the same time, Dauxois, Ferré and Yao (2001) and then Ferré and Yao (2003 and 2004) have proposed a semi-parametric model for Hilbertian variables which corresponds to the functional version of Li's Sliced Inverse Regression, Li (1991).

On a classification point of view, many solutions have been proposed to overcome ill-posed functional problems including the popular penalization methods. Friedman (1989) presents the RDA model based on regularization and shrinkage. Hastie, Tibshirani and Buja (1994 and 1995) propose a discriminant analysis penalized by smoothing functionals. The idea of penalization was first developped by Ivanov (1962) and Tihonov (1963 a and b) and it has been used by Pezzulli and Silverman (1993) and Silverman (1996) for smoothed Principal Components Analysis and by Leurgans, Moyeed and Silverman (1993) for Canonical Correlation Analysis. Finally, a review of many regularization methods can be found in Tenorio (2001).

In this paper, we propose a new way to achieve functional regression: the idea is to join the efficiency of a dimension reduction method using smoothing penalization, to the strong adaptability of a neural network which can provide highly non linear solutions even if the number of predictors is too large for classical nonparametric methods such as kernels. The functional SIR dimension reduction method is first presented in Section 2. For this penalized version, consistency results are given in Section 3. Section 4 discusses Neural Network and gives consistency results for the proposed model combining FSIR and Neural Networks (which will be called SIR-NNr). Section 5 is devoted to applications: Section 5.1 deals with the Tecator data set and Section 5.2 with the phoneme data set. In Appendix A, we give a sketch of the proofs. All programs have been made using Matlab and are available on request.

2 Sliced Inverse Regression

Let Y be a real random variable and X be a multivariate variable assumed to have a fourth moment. To overcome the curse of dimensionality in the nonparametric regression of Y on X, Li (1991) introduced the Sliced Inverse Regression. He considers the following model

$$Y = f(a_1'X, a_2'X, \dots, a_q'X, \epsilon),$$

where ϵ is centered and independent of X, f is an unknown function and $(a_j)_{j=1,\dots,q}$ are lineary independent vectors.

The space spanned by $(a_j)_{j=1,\dots,q}$ is called EDR (Effective Dimension Reduction) space. SIR deals with the estimation of this EDR space and the aim of sliced inverse regression is to estimate it by means of the eigenvectors of the matrix $Var(X)^{-1}Var(E(X|Y))$.

In the multivariate context, numerous works deal with SIR. Li (1991), Schott (1994), Ferré (1998) and Vellila (1998) have worked to determine the dimensionality q. Then, methods have been proposed to improve SIR: different estimates of the covariance of the conditional mean have been built (in Hsing and Carroll (1992), Zhu and Ng (1995) and Zhu and Fang (1996)) while other methods have been proposed to estimate the EDR space (for example, PHD proposed by Li (1992), SAVE by Cook (1991) or MAVE by Xia, Tong, Li and Zhu (2002)). The main interest of this model is that, once the EDR space is estimated, the estimation of f is obtained very easily with traditional techniques provided that q is not too large.

2.1 Functional SIR

Now consider a real random variable Y and X a random variable taking its values in $\mathcal{L}^2_{\mathcal{T}}$, the space of squared intregrable functions from a compact interval $\mathcal{T}$ into $\mathbb{R}$. With the usual inner product defined by, for all f, g in $\mathcal{L}^2_{\mathcal{T}}$, $< f, g >= \int_{\mathcal{T}} f(t)g(t)dt$, $\mathcal{L}^2_{\mathcal{T}}$ is a Hilbert space. We will assume that the random variable X is centered and has a fourth moment. Then, the covariance operator of X exists and is defined by $\Gamma_X = E(X \otimes X)$ where $X \otimes X$ denotes the operator which associates to any f in $\mathcal{L}^2_{\mathcal{T}}$, $< f, X > X$. We also get that $E(X/Y)$ and $\Gamma_{E(X|Y)} = Var(E(X|Y))$ exist. Ferré and Yao (2003) have proposed to investigate the following model for functional inverse regression:

$$Y = f(< X, a_1 >, \ldots, < X, a_q >, \epsilon) \qquad (1)$$

where f is an unknown function, ϵ a random variable which is centered and independent of X and $(a_j)_{j=1,\ldots,q}$ are lineary independent functions of $\mathcal{L}^2_{\mathcal{T}}$.

We focus on the estimation of $(a_j)_{j=1,\ldots,q}$. The key of the method comes from the following theorem:

Theorem 1 *Writing $A = (< X, a_1 >, \ldots, < X, a_q >)^T$, if*

(A1) *for all u in $\mathcal{L}^2_{\mathcal{T}}$ there exists v in $\mathbb{R}^q$ such that: $E(< u, X > /A) = v^T A$*

then $E(X/Y)$ belongs to the subspace spanned by $\Gamma_X a_1, \ldots, \Gamma_X a_q$.

Remark: Note that Cook and Weisberg (1991) show that elliptically distributed variables satisfy condition **(A1)** in the multidimensional context but this can be transposed in infinite dimensional Hilbert spaces.

By using the result of Dauxois, Ferré and Yao (2001), a consequence of Theorem 1 is that the EDR subspace contains the Γ_X-orthonormed eigenvectors of $\Gamma_X^{-1}\Gamma_{E(X/Y)}$ associated with the q positive eigenvalues. This is the generalization of Li (1991) on SIR to infinite dimensional case.

Unfortunately, Γ_X^{-1} is not defined since we have to assume that Γ_X is a positive definite operator which implies that it is not invertible as defined from $\mathcal{L}^2_{\mathcal{T}}$ to $\mathcal{L}^2_{\mathcal{T}}$. However, if we call $(\delta_i)_{i=1,\ldots,\infty}$ its sequence of eigenvalues and $(u_i)_{i=1,\ldots,\infty}$

those of orthonormed eigenvectors, R_Γ the range of Γ_X and $R_\Gamma^{-1} = \left\{ h \in \mathcal{H} : \exists f \in R_\Gamma, h = \sum_i \frac{1}{\delta_i}(u_i \otimes u_i)(f) \right\}$, Γ_X is a one-to-one mapping from R_Γ^{-1} to R_Γ whose inverse, also called Γ_X^{-1}, is defined by $\Gamma_X^{-1} = \sum_i \frac{1}{\delta_i} u_i \otimes u_i$.

A basis of the EDR space is thus given by the eigenvector of $\Gamma_X^{-1}\Gamma_{E(X/Y)}$ but to ensure that these eigenvectors exist in $\mathcal{L}^2_{\mathcal{T}}$, we have to assume that (see Ferré and Yao (2004) for details) $\sum_i \sum_j \frac{1}{\delta_i \delta_j} E(E(\zeta_i/Y)E(\zeta_j/Y))^2 < +\infty$, where $X = \sum_i \zeta_i u_i$ is the Karunen Loeve decomposition of X.

Thus, in order to estimate the EDR space, we have to choose an estimate for $\Gamma_{E(X/Y)}$. We have two possibilities:

1. A slicing approach. In Ferré and Yao (2003), the estimate is obtained by partitionning the domain of Y in $\{I_h\}_{h=1,\dots,H}$:

$$\Gamma^N_{E(X/Y)} = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \otimes \mu_h - \overline{X} \otimes \overline{X}$$

   where, if $\mathbb{I}$ is the indicator function, $N_h = \sum_{n=1}^N \mathbb{I}_{\{Y^n \in I_h\}}$, $\mu_h = \frac{1}{N_h} \sum_{n=1}^N X^n \mathbb{I}_{\{Y^n \in I_h\}}$ and $\overline{X} = \frac{1}{N} \sum_{n=1}^N X^n$ is the empirical mean.

2. A kernel based approach. In Ferré and Yao (2004), it is assumed that Y has a probability density; thus a kernel estimate (of the Nadaraya-Watson type) is used:

$$\widehat{E(X/Y = y)} = \sum_{n=1}^{N} \frac{X^n K\left(\frac{Y^n - y}{h}\right)}{\sum_{m=1}^N K\left(\frac{Y^m - y}{h}\right)}$$

   and $\Gamma^N_{E(X/Y)} = \frac{1}{N} \sum_{n=1}^N \widehat{E(X/Y = Y^n)} \otimes \widehat{E(X/Y = Y^n)} - \overline{X} \otimes \overline{X}$.

A usual estimate of Γ_X is $\Gamma_X^N = \frac{1}{N} \sum_{n=1}^N X^n \otimes X^n - \overline{X} \otimes \overline{X}$, but this estimate is ill conditionned (because Γ_X is not a bounded operator) so the eigenvectors of $(\Gamma_X^N)^{-1}\Gamma^N_{E(X/Y)}$ do not converge to the eigenvectors of $\Gamma_X^{-1}\Gamma_{E(X/Y)}$. That is the reason why penalization or regularization is needed.

Ferré and Yao (2003) suggest to proceed like Bosq (1991) by considering, instead of Γ_X, a sequence of finite rank operators with bounded inverses and converging to Γ_X. This leads to the estimates $(\hat{a}_j^N)_{j=1,\dots,q}$ of $(a_j)_{j=1,\dots,q}$ that, under some conditions, satisfy the following consistency result:

$$\| \hat{a}_j^N - a_j \| \rightarrow_p 0$$

The authors also suggest a way of estimating the EDR space for functional data without inverting the covariance operator of the regressor (Ferré and Yao, 2004).

We propose, in Section 3, a regularized approach by penalization.

2.2 SIR for classification

Let $\mathcal{C}_1, \ldots, \mathcal{C}_H$ be H groups. When Y is multidimensional, the results of Dauxois and al. (2001) are still available and by setting $Y = (\mathbb{I}_{\mathcal{C}_1}, \ldots, \mathbb{I}_{\mathcal{C}_H})$, where $\mathbb{I}_{\mathcal{C}_h}$ is the indicator function of the hth group, Model (1) remains valid and we get a natural way to include classification problems into FSIR, see Ferré and Villa (2005). Note that, in the functional case, multivariate methods for discrimination have been extended, mainly inspired from Linear Discriminant Analysis (LDA). In this area, let us mention the works of Hastie et al. (1994, 1995) and James et al. (2000).

Now, by estimating $\Gamma_{E(X/Y)}$ by

$$\Gamma^N_{E(X/Y)} = \frac{1}{N}\sum_{h=1}^{H} N_h \widehat{E(X/Y=h)} \otimes \widehat{E(X/Y=h)} - \overline{X} \otimes \overline{X}$$

where $N_h = \sum_{n=1}^{N} \mathbb{I}_{\{Y^n=h\}}$ and $\widehat{E(X/Y=h)} = \frac{1}{N_h}\sum_{n=1}^{N} X^n \mathbb{I}_{\{Y^n=h\}}$, FSIR leads to a discriminant analysis. The estimation of the EDR space is identical to the discriminant space in linear discriminant analysis. However, the estimation of f leads to a natural classification rule. Indeed, since we have, for all x, $f(x) = E(Y|X=x) = (P(C_1|X=x), ..., P(C_H|X=x))$, the estimation of f coincides with the estimations of the probabilities of the groups conditionally to X.

3 Regularized functional SIR

In Section 2, we saw that the EDR space contains the eigenvalues of the operator $\Gamma_X^{-1}\Gamma_{E(X/Y)}$. Thus, as it is the case for the Discriminant Analysis, the estimator of the first direction of the EDR space can be found by maximizing a Rayleigh criterion: $\max_a \frac{<\Gamma_{E(X/Y)}a,a>}{<\Gamma_X a,a>}$. Unfortunately, as Γ_X^N is ill conditionned, the maximization of the empirical Rayleigh expression does not lead to a good estimate of the EDR space: that is the reason why a regularization is needed.

Provided that we have smooth functions, a relevant method for functionnal data is to penalize the covariance operator in the Rayleigh expression by introducing smoothing constraints on the estimated functions. This method has already proved its great efficiency (see Hastie and al. (1995) for an example of the penalized discriminant analysis).

3.1 Main results

Let $\mathcal{S}$ be the subspace of $\mathcal{L}^2_{\mathcal{T}}$ of functions with a squared integrable second derivative. We introduce a penalty through a bilinear form defined on $\mathcal{S} \times \mathcal{S}$ by, for all f, g in $\in \mathcal{S}$,

$$[f,g] = \int_{\mathcal{T}} D^2 f(t) D^2 g(t) dt$$

We also define the penalized bilinear form associated with empirical operator Γ_X^N:

$$Q_\alpha^N(f,g) = <\Gamma_X^N f, g> + \alpha[f,g]$$

where α is a regularization parameter. The solutions of the regularized SIR are given by maximizing, under orthogonal constraints, the function

$$\gamma^N(a) = \frac{<\Gamma_{E(X/Y)}^N a, a>}{<\Gamma_X^N a, a> + \alpha[a,a]}.$$

In order to obtain consistency results for the estimate of $(a_j)_{j=1,\ldots,q}$, we make the following assumptions

(A2) $E(\parallel X \parallel^4) < +\infty$;

(A3) for all $\alpha > 0$,

$$\inf_{\|a\|=1,\ a\in\mathcal{S}} Q_\alpha(a,a) = \rho_\alpha > 0;$$

(A4) $\Gamma_{E(X/Y)}^N$ is a continuous operator which converges in probability to $\Gamma_{E(X/Y)}$ with $\sqrt{N}$ rate;

(A5) $\lim_{N\to+\infty} \alpha = 0$, $\lim_{N\to+\infty} \sqrt{N}\alpha = +\infty$;

(A6) $(a_j)_{j=1,\ldots,q}$ are Γ_X-orthogonal, with Γ_X-norm equal to 1 belonging to $\mathcal{S}$ and verifying, for all u such that $<\Gamma_X u, a_1>= 0$ and that $<\Gamma_X u, u>= 1$,

$$<\Gamma_{E(X/Y)} u, u> \leq <\Gamma_{E(X/Y)} a_2, a_2> = \lambda_2 < \lambda_1;$$

Theorem 2 *Under assumptions* **(A1)-(A6)** *the function γ^N reaches its maximum on $\mathcal{S}$ with probability converging to 1 when N grows to $+\infty$.*
Let then a_1^N be a vector of $\mathcal{S}$ for which γ^N is maximum and which is such that $<\Gamma_X a_1^N, a_1>= 1$. Then,

$$<\Gamma_X(a_1^N - a_1), a_1^N - a_1 > \to_p 0\ .$$

Remarks:

- For an understandable presentation, we introduce a particular type of penalization but previous results can be found for other regularization functionals satisfying the assumptions. For example, we can replace the bilinear form [., .] by another one which is similar to the one used in Ridge-PDA (Hastie et al, 1995).

- Assumptions **(A2)**, **(A3)** and **(A4)** are technical assumptions that ensure the existence and convergence for $(a_j^N)_{j=1,\dots,q}$: **(A2)** implies that Γ_X^N will converge to Γ_X at the $\sqrt{N}$ rate; we can find in Leurgan et al (1993) conditions that involve **(A3)**. This assumption shows the purpose of regularization: it controls the scaling of Q_α and, thanks to **(A5)**, ensures that the denominator of γ^N doesn't go too fast to 0. Finally **(A5)** gives a way of choosing regularization parameter α (for pratical aspects see section 3.2).

3.2 Practical aspects

On a practical point of view, X has been observed at some points t_1, t_2, $\dots$, t_D (for a understandable presentation, we suppose that these observations have been centered). The optimization of the penalized Rayleigh expression described in Section 3.1 can be applied by using, for example, B-Splines $(B_i)_i$ to parametrize a_1^N:

$$a_1^N(t) = \sum_i A_{1i} B_i(t) = A_1 B$$

where B is the matrix containing the values of $(B_i(t))_i$ at the points t_1, t_2, $\dots$, t_D.
Similarly, the matrix of observations X can be written in the form of B-Splines:

$$X = CB$$

with $C = \begin{bmatrix} C^1 \\ \vdots \\ C^N \end{bmatrix}$. Let $B^{(2)}$ be the vector containing the values $D^2 B(t)$. If we use the slicing estimate of $\Gamma_{E(X/Y)}$ for regression, we introduce,
for all $h = 1, \dots, H$,

$$Y_h = \begin{bmatrix} \mathbb{I}_{\{Y^1 \in I_h\}} \\ \vdots \\ \mathbb{I}_{\{Y^N \in I_h\}} \end{bmatrix}.$$

Then the problem of maximizing γ^N is equivalent to maximizing

$$\frac{A' M_e A}{A' M_{X,\alpha} A}$$

where M_e is the estimator of $\Gamma_{E(X/Y)}$ obtained by the slicing approach :

$$M_e = \sum_{h=1}^{H} \frac{N_h}{N} BB'C'Y_h Y_h' CBB'$$

and where

$$M_{X,\alpha} = \frac{1}{N} BB'C'CBB' + \alpha B^{(2)}\,' B^{(2)} .$$

The first solution is the eigenvector, with $M_{X,\alpha}$-norm equal to 1, associated with the largest eigenvalue of the matrix $M_{X,\alpha}^{-1}M_e$. By pursuing the procedure under othogonality constraints, we get that the other solutions are the $M_{X,\alpha}$-orthonormal eigenvectors of $M_{X,\alpha}^{-1}M_e$.

If we deal with classification, the same procedure is achieved by letting

$$Y_h = \begin{bmatrix} \mathbb{I}_{\{Y^1=h\}} \\ \vdots \\ \mathbb{I}_{\{Y^N=h\}} \end{bmatrix}.$$

Finally we have to find the optimal value for α. This can be done, if the sample we have is large enough (which is the case in the applications that we present), by dividing it into two parts: we apply the previous procedure on the first part to fin $(a_j^N)_j$ and evaluate the error committed by Model (1) on the second part; the best parameter is then chosen to minimize this error.

Remark : The estimation of $\Gamma_{E(X/Y)}$ can also be made by a kernel approach ; the efficiency of this approach can even be better than those we have with the slicing estimate (see Ferré and Yao (2004) for practical comparisons).

4 Neural network

4.1 Approximation by neural networks

After the EDR space is estimated, the purpose is to get an estimation of function f in (1): we propose to use a feedforward neural network with one hidden layer. This method (see, e.g., Bishop (1995) for a review on Neural Networks) is an alternative to nonparametric regressions if the dimension of the EDR space is too large. It has the advantage of working in any cases while nonparametric methods, such as kernel or splines, face the curse of dimensionality.

The main interest of neural networks is their ability to approximate any function with the desired precision (universal approximation); see, for instance, Hornik (1991, 1993) for the multivariate context and Stinchcombe (1999) and Rossi, Conan-Guez and Fleuret (2002) in the infinite dimensional one.

4.2 Consistency results

Neural Network approximations of functionals in infinite dimensional spaces have been studied in Chen and Chen (1995), Sandberg and Xu (1996), Rossi, Conan-Guez and Fleuret (2002) and Conan-Guez and Rossi (2002 and 2003). Several strategies are available either by directly using the curves as inputs of the feedforward neural networks or by first projecting the data onto a classical functional basis (such as a spline basis, a Fourier basis, wavelets) or a basis derived from the PCA of X. This latter approach is used by Thodberg (1995).

Our approach is similar but, instead of projecting the data onto a fixed basis or a principal component basis, we project them onto the EDR space. The EDR space behaves as an efficient subspace for the regression of Y on X and it is a way to get a basis which takes into account the relationship between Y and X. In fact, the data are projected onto an estimation of the EDR space, so the accuracy of the projection and then the estimation of the optimal weights for the neural network also depend on how good the EDR space is estimated.

We construct a perceptron (see Figure 2) with one hidden layer having

- as inputs, the coordinates of the projection of X on $(a_j)_{j=1,\dots,q}$: $< X, a_1 >$, $\dots$, $< X, a_q >$;
- q_2 neurons on the hidden layer (where q_2 is a parameter to be estimated);
- as outputs, one neuron for regression and H neurons for classification, representing target Y.

[Figure 2 about here.]

The output of such a neural network is then $\sum_{i=1}^{q_2} w_i^{(2)} g\left(\sum_{j=1}^{q} w_{i,j}^{(1)} < X, a_j > +w_i^{(0)}\right)$ where g is the activation function (for example a sigmoid). The purpose of the training step is then to find w^* which minimizes a loss function L between the output of the neural network with weights $w = \left((w_i^{(2)})_{i=1,\dots,q_2}, (w_{i,j}^{(1)})_{i=1,\dots,q_2}^{j=1,\dots,q}, (w_i^{(0)})_{i=1,\dots,q_2}\right)$, and the target Y:

$$w^* = argmin\left\{E\left[\,L\left(\sum_{i=1}^{q_2} w_i^{(2)} g\left(\sum_{j=1}^{q} w_{i,j}^{(1)} < X, a_j > +w_i^{(0)}\right), Y\right)\right]\right\}. \quad (2)$$

Actually we obtain an estimation w_N^* of w^* by

$$w_N^* = argmin\left\{\sum_{n=1}^{N} L\left(\sum_{i=1}^{q_2} w_i^{(2)} g\left(\sum_{j=1}^{q} w_{i,j}^{(1)} < X^n, a_j^N > +w_i^{(0)}\right), Y^n\right)\right\}.$$

White (1989) gives a consistency theorem for the weights of a neural networks estimated by a set of iid observations. Since $(a_j^N)_j$ is an estimation of the EDR space deduced from the whole data set $(X^n, Y^n)_n$, the inputs of our functional perceptron used to determine w_N^* do not satisfy the iid assumption and then proper consistency result is then needed.

Let us introduce some notations: ζ is the function from $\mathcal{O} \times \mathcal{W}$ ($\mathcal{O}$ is an open set of $\mathbb{R}^{q+1}$ and $\mathcal{W}$ is a compact set of $\mathbb{R}^{(q+2)q_2}$) such as for all $z = (u, y)$ in $\mathcal{O}$, $\zeta(z, w) = L\left(\sum_{i=1}^{q_2} w_i^{(2)} g\left(\sum_{j=1}^{q} w_{i,j}^{(1)} u_j + w_i^{(0)}\right), y\right)$; Z is the couple of random variables $(\{< X, a_j >\}_j, Y)$ and $\{Z_n\}_{n=1,\dots,N}$ are observations of Z; finally, $\{\tilde{Z}_N^n\}_{n=1,\dots,N}$ are the couples of $(\{< X^n, a_j^N >\}_j, Y^n)$. In our context, the consistency of the Multilayer Perceptron is given by the following theorem:

Theorem 3 *Under assumptions* **(A1)-(A6)** *and the following assumptions*

(A7) *for all* z *in* $\mathcal{O}$, $\zeta(z,.)$ *is continuous;*

(A8) *there is a measurable function* $\tilde{\zeta}$ *from* $\mathcal{O}$ *into* $\mathbb{R}$ *such that for all* z *in* $\mathcal{O}$, *for all* w *in* $\mathcal{W}$, $|\zeta(z,w)| < \tilde{\zeta}(z)$ *and* $E(\tilde{\zeta}(Z)) < +\infty$*;*

(A9) *for all* w *in* $\mathcal{W}$, *there exists* $C(w) > 0$ *such that, for all* (x,y) *and* (x',y') *in* $\mathcal{O}$, $|\zeta((x,y),w) - \zeta((x',y),w)| \leq C(w) \parallel x - x' \parallel$

(A10) *for all* w *in* $\mathcal{W}$, $\zeta(.,w)$ *is measurable.*

If $\mathcal{W}^*$ *is the set of minimizers of the problem (2) then*

$$d(w_N^*, \mathcal{W}^*) \xrightarrow{N\rightarrow+\infty}_p 0.$$

Remarks:

- This list of assumptions is, for example, checked by a perceptron with one hidden layer and a sigmoid function $g(x) = \frac{e^x}{1+e^x}$ on the hidden layer associated with the mean squared error $L(\psi, y) = \parallel \psi - y \parallel^2$.
- Assumptions **(A1)-(A6)** ensure the convergence of $(a_j^N)_{j=1,\ldots,q}$ to $(a_j)_{j=1,\ldots,q}$ but they can be replaced by a list of assumptions implying the same result. For example, we would have the same consistency result by projecting the data on the estimated EDR space found by the functional SIR presented in Ferré and Yao (2003 and 2004).

5 Applications

5.1 Tecator data

As already said, the Tecator data problem consists in predicting the fat content of pieces of meat from a near infrared absorbance spectrum. We have $N = 215$ observations of (X, Y) where X is the spectrum of absorbance discretized at one hundred points and Y is the lipid rate.

In order to compute the procedure described in section 3.2, we project the data on a cubic Spline basis. Because of their smoothness, these data are very well projected on a basis with 40 knots (actually, up to 40 knots, the interpolation is exact); then, we used this projection for the computation when needed and used the original data in the other cases. We tried several classical methods in order to test the efficiency of SIR-NNr. The competitors are:

- **SIR-NNr**: the functional SIR regularized by penalization, presented in Section 3, pre-proceeds a neural network. The neural network training step is made by early stopping procedure: the learning sample is divided into 3 samples (training / validation / test); the training sample is used to train the neural network, the validation sample for the early stopping procedure and this training step is performed 10 times. The best performance of the test samples is kept as the optimal weights;

- **SIR-NNk**: here we use the smoothed functional inverse regression method presented in Ferré and Yao (2003) as pre-processing to a neural network; the purpose is to show the benefit of the regularization. The neural network is also trained by early stopping;
- **PCA-NN**: in order to show the advantage of SIR, we compute a principal component analysis (as Thodberg (1995)) before a neural network procedure is used (a classical neural network while Thodberg uses a sophisticated bayesian neural network);
- **NNf**: this method is the functional neural network (the Spline projections are used to represent the functional weights and inputs) described by Rossi and Conan-Guez (2003);
- **SIR-L**: after projecting the data on the EDR space determined by regularized SIR, we compute a linear regression in order to show the efficiency of a neural network compared to a classical parametric method.

We also have to notice that classical nonparametric methods such as kernel can not be used for this data set as the dimensionality of the EDR space is too large (the value of q is given in Table 1).

Before we compare the different methods and in order to limit computational time, we determined the best parameters for each one : our sample is divided into two parts: on the first one, we determine the values of $(a_j^N)_j$ and of the weights of the neural network for various values of α, q and q_2. On the second part, we determine the standard error of prediction (SEP): the "best" parameters are those who minimize this SEP (see Table 1).

[Table 1 about here.]

Then, in order to see not only the error made by each method but also its variability, we randomly build 50 samples divided as follows: the learning sample contains 172 observations and the test sample contains 43. All five methods are first trained on the learning sample (with their optimal parameters pre-determined as described above) and the standard error of prediction (SEP) is then performed on the test sample.

Figure 3 gives the boxplot of the test errors for the 50 samples and Table 2 gives a numerical description of the performances of the different methods.

[Figure 3 about here.]

[Table 2 about here.]

These results show the excellent performances obtained by SIR-NNr: its SEP average over the 50 samples is twice lower than any of the other competitors. Moreover, this method garantees a good stability unlike the others. SIR seems to be a very good pre-processing stage, as SIR-NNk also obtains good performances. Then we have NNf but its rather good results suffer from a very slow computational time. To show this, we give the computational time of each method in Table 3. Clearly NNf is very expensive while SIR-L is very fast but works poorly. Actually, it is closely related to the number of inputs: 42 for NNf, 10 for SIR-NNk, 12 for PCA-NN and 20 for SIR-NNr.

[Table 3 about here.]

5.2 Phoneme data

In this section, we compare our methodology with other approaches on a classification problem, namely the phoneme data. The data are log-periodograms of a 32 ms duration corresponding to recorded speakers; it deals with the discrimination of five speech frames corresponding to five phonemes transcribed as follow: [sh] as in "she", [dcl] as in "dark", [iy] as in "she", [aa] as in "dark" and [ao] as in "water". Finally, the data consist in 4 509 log-periodograms of a 256 length (see Figure 4).

[Figure 4 about here.]

We tried several classical methods in order to test the efficiency of SIR-NNr which is compared with:

- **SIR-NNp**: a classical SIR as presented in Ferré and Yao (2003) as pre-processing of a neural network: the purpose is to show the advantage of regularization compared to a projection on a PCA basis;

- **SIR-K**: a regularized functional SIR where the function f is estimated by a nonparametric kernel method;

- **Ridge-PDA**: the penalized discriminant analysis introduced in Hastie et al. (1995) which uses ridge penalty;

- **NPCD-PCA**: a nonparametric method using kernel and semi-metrics based on Principal Component Analysis and introduced by Ferraty and Vieu (2003).

The optimal parameters for these methods are shown in Table 4.

[Table 4 about here.]

For the SIR stage, the dimension of the EDR space is 4. This can be seen by looking at the projection of the data on the EDR space (for SIR-NNr, for example, see Figure 5). We can see that only the fourth axis is able to separate the phonems [aa] and [ao].

[Figure 5 about here.]

Then we randomly build 50 samples divided as follows: the learning sample contains 1 735 log-periodograms (347 for each class) and the test sample contains also 1 735 (347 for each class). All five methods are first trained on the learning sample and the test error rate is then computed on the test sample. Figure 6 proposes the boxplot of the test error rates and Table 5 gives a description of the performances of test error rates over the 50 samples.

[Figure 6 about here.]

[Table 5 about here.]

The results of SIR-NNr, SIR-NNp and SIR-K are very close. The beneficial aspect of SIR is highlighted since those three methods work better than others based on different projections of data. The advantage of regularization is also revealed since it leads again to the best results. Then comes RPDA and finally NPCD-PCA which provides the poorest performances. On the contrary, in this low dimensional problem, neural networks seem to be less performant than kernels and have a bigger variability (standard deviation is 0,56 for SIR-NNr and only 0,40 for SIR-K): this problem can be removed by increasing the number of training steps or by using more sophisticated architecture, but at the price of a larger computational cost. Finally, if SIR-K obtains the best mean, SIR-NNr is the method which reaches the best minima which shows its great potential.

In conclusion, both on regression and classification problems, regularized SIR-NN is a competitive solution for functional problems: we can explain these good results by noting that the procedure combines an efficient dimension reduction model and the great accuracy of a neural network, which is able to approximate almost every function. Thus this model can be efficient both for ill-posed problems thanks to the penalized functional and for problems with a large dimensionality thanks to the neural network step. Finally it has another great advantage: computational time is rather short and does not increase too much with the number of observation points for the curves.

A Proofs

Here we give main lines of the proofs of Theorems 2 and 3.

A.1 Theorem 2

The proof of this theorem is related to the one of Theorem 1 in Leurgan at al. (1993) and only sketches are given.

Lemma 1: Using Central Limit Theorem, it is easy to show that if $\delta^N = max\{|||\Gamma_X^N - \Gamma_X|||; |||\Gamma_{E(X/Y)}^N - \Gamma_{E(X/Y)}|||\}$ and if the sequence $(k_N)_N$ satisfies $\sqrt{N}k_N \rightarrow +\infty$ then

$$k_N^{-1}\delta^N \rightarrow_p 0.$$

Existence: Let Q_α be the bilinear form satisfying, for all f, g in $\mathcal{S}$, $Q_\alpha(f,g) =< \Gamma_X f, g > +\alpha[f,g]$. We have for α in $[0,1]$, $Q_\alpha = (1-\alpha) < \Gamma_X ., . > +\alpha Q_1$ and then, for all u such that $\| u \|= 1$, $\frac{1}{\alpha} Q_\alpha(u,u) > (\frac{1}{\alpha} - 1) < \Gamma_X u, u > +Q_1 > \rho_1$ by the positiveness of Γ_X. Then, $\sqrt{N}\rho_\alpha > \alpha\sqrt{N}\rho_1$ and we have

$$\sqrt{N}\rho_\alpha \to +\infty \ . \qquad (3)$$

Then, by Lemma 1, noting $\Delta_1^N = \Gamma_X^N - \Gamma_X$,

$$\lim_{N\to+\infty} P\left(\{\omega \in \Omega \quad ||| \Delta_1^N ||| \leq \frac{1}{2}\rho_\alpha\}\right) = 1.$$

But, we have

$$\{\omega \in \Omega \quad ||| \Delta_1^N ||| \leq \frac{1}{2}\rho_\alpha\} \subset \left\{\omega : \forall\ a \in \mathcal{S},\ \| a \|= 1,\ Q_\alpha^N(a,a) \geq \frac{1}{2}\rho_\alpha > 0\right\}$$

and finally the right hand part of the previous equation has a probability converging to 1 when N converges to $+\infty$.

Let $\overline{\mathcal{B}(0,1)}$ be the weak closure of $\{a \in \mathcal{S} \quad Q_\alpha^N(a,a) = 1\}$ and ζ be the functional defined on $\{a \in \mathcal{S} \quad Q_\alpha^N(a,a) = 1\}$ by

$$\zeta(a) =< \Gamma_{E(X/Y)}^N a, a >$$

then ζ can be extended to a uniformly continuous functional $\tilde{\zeta}$ defined on $\overline{\mathcal{B}(0,1)}$ for the weak topology. Finally, provided that $Q_\alpha^N(a,a) \geq \frac{1}{2}\rho_\alpha$, $\tilde{\zeta}$ reaches its maximum on weak compact $\overline{\mathcal{B}(0,1)}$ which concludes the proof of the existence of $(a_j^N)_{j=1,\dots,q}$.

Consistency: For the following we suppose that we stand on the set where γ^N has a maximum on $\mathcal{S}$ and reaches it.

Let λ_1^N be this maximum and λ_1^α be the maximum of

$$\gamma_\alpha(a) = \frac{< \Gamma_{E(X/Y)} a, a >}{< \Gamma_X a, a > +\alpha[a,a]}$$

on $\mathcal{S}$; λ_1^α is well defined thanks to assumption **(A3)**.

Considering $\frac{\gamma_\alpha(a)}{\gamma_0(a)}$ we easily show that

$$\lambda_1^\alpha \to \lambda_1. \qquad (4)$$

Then by proving that $\sup_{a\in\mathcal{S}} |\gamma^N(a) - \gamma_\alpha(a)| \to_p 0$ we can show that

$$\left|\lambda_1^N - \lambda_1^\alpha\right| \to_p 0. \qquad (5)$$

Finally, combining (4) and (5) we conclude that

$$\lambda_1^N \to_p \lambda_1 \qquad (6)$$

Then using (6) we demonstrate that

$$\gamma(a_1^N) \rightarrow_p \lambda_1 = \gamma(a_1). \tag{7}$$

Thanks to the conclusion of Theorem 1 we show that

$$\lim_{N\to+\infty} \mathbb{P}(< \Gamma_{E(X/Y)}a_1, a_1^N - a_1 >=< \Gamma_X a_1, a_1^N - a_1 >= 0) = 1.$$

Let μ_N be $< \Gamma_X(a_1^N - a_1), a_1^N - a_1 >$; if $< \Gamma_{E(X/Y)}a_1, a_1^N - a_1 >= 0$, we have

$$\lambda_1^{-1}\gamma(a_1^N) \leq \frac{1+\lambda_1^{-1}\lambda_2\mu_N}{1+\mu_N} .$$

As $\lambda_1^{-1}\lambda_2 < 1$, the right hand side of the previous inequality is less than 1; but $\lambda_1^{-1}\gamma(a_1^N)$ converges in probability to 1 by (7) so

$$\frac{1+\lambda_1^{-1}\lambda_2\mu_N}{1+\mu_N} \rightarrow_p 1$$

and then we conclude with $\mu_N \rightarrow_p 0$.

A.2 Theorem 3

The proof of this theorem is close to the one found in Rossi and al. (2002) and Conan-Guez and Rossi (2002); the main difference is in the fact that the projection for the data is a random variable. The proof will be divided into two parts:

We first prove that

$$\sup_{w\in\mathcal{W}} \left| \frac{1}{N}\sum_{n=1}^{N} \zeta(\tilde{Z}_N^n, w) - E(\zeta(Z,w)) \right| \rightarrow_p 0. \tag{8}$$

Forall w in $\mathcal{W}$, we have

$$\left| \frac{1}{N}\sum_{n=1}^{N} \zeta(\tilde{Z}_N^n, w) - E(\zeta(Z,w)) \right|$$

$$\leq \left| \frac{1}{N}\sum_{n=1}^{N} \zeta(\tilde{Z}_N^n, w) - \frac{1}{N}\sum_{n=1}^{N} \zeta(Z_n, w) \right| + \left| \frac{1}{N}\sum_{n=1}^{N} \zeta(Z_n, w) - E(\zeta(Z,w)) \right|.$$

By using dominated convergence theorem, the fact that $\mathcal{W}$ is a compact set, that $\zeta(z, .)$ is continuous for all $z \in \mathcal{O}$, and that $\zeta(., w)$ is mesurable for all $w \in \mathcal{W}$, we can show that, for all $\tilde{w} \in \mathcal{W}$,

$$\lim_{\mu\to 0} E\left(\sup_{w\in\mathcal{W}\cap\mathcal{B}(\tilde{w},\mu)} \zeta(Z,w) \right) = E(\zeta(Z,\tilde{w}));$$

Then let ϵ be a real positive number, for all $\tilde{w} \in \mathcal{W}$, there is a $\mu(\tilde{w})$ such that

$$E\left(\sup_{w\in\mathcal{W}\cap\mathcal{B}(\tilde{w},\mu(\tilde{w})}\zeta(Z,w)\right) \leq E(\zeta(Z,\tilde{w}) + \frac{\epsilon}{3} \tag{9}$$

$$E\left(\inf_{w\in\mathcal{W}\cap\mathcal{B}(\tilde{w},\mu(\tilde{w})}\zeta(Z,w)\right) \geq E(\zeta(Z,\tilde{w}) - \frac{\epsilon}{3}. \tag{10}$$

Using the law of large numbers we can deduce from (9) and (10) that for all $\tilde{w} \in \mathcal{W}$, almost surely, there is a $N(\tilde{w}) \in \mathbb{N}$ such that, for all $N \geq N(\tilde{w})$,

$$\sup_{w\in\mathcal{W}\cap\mathcal{B}(\tilde{w},\mu(\tilde{w}))}\left|\frac{1}{N}\sum_{n=1}^{N}\zeta(Z_n,w) - E(\zeta(Z,w))\right| \leq \epsilon.$$

As $\mathcal{W}$ is a compact set, we can find $\tilde{w}_1,\ldots,\tilde{w}_I$ such as $\{\mathcal{B}(\tilde{w}_i,\mu(\tilde{w}_i))\}_{i=1,\ldots,I}$ re-cover $\mathcal{W}$. Using these sets we conclude that

$$\left|\frac{1}{N}\sum_{n=1}^{N}\zeta(Z_n,w) - E(\zeta(Z,w))\right| \leq \epsilon,$$

Using assumption **(A8)** we see that

$$\begin{aligned}&\left|\tfrac{1}{N}\textstyle\sum_{n=1}^{N}\left(\zeta(\tilde{Z}_N^n,w) - \zeta(Z_n,w)\right)\right| \\ &\qquad\leq C(w)\left[\textstyle\sum_{j=1}^{q} < \Gamma_X^N(a_j^N - a_j), a_j^N - a_j >\right]^{1/2}\end{aligned}$$

As $|||\Gamma_X^N - \Gamma_X||| \rightarrow_p 0$ and as, for all $j = 1,\ldots,q$, $< \Gamma_X(a_j^N - a_j), a_j^N - a_j > \rightarrow_p 0$, we then conclude that

$$\sup_{w\in\mathcal{W}}\left|\frac{1}{N}\sum_{n=1}^{N}\left(\zeta(\tilde{Z}_N^n,w) - \zeta(Z_n,w)\right)\right| \rightarrow_p 0,$$

which finally implies (8).

Secondly, let ϵ be a positive real. According to the Dominated Convergence Theorem, $E(\zeta(Z,.))$ is a continuous function which reaches its minimum m on compact set $\mathcal{W}$. Then we can show that there is a $\eta(\epsilon) > 0$ such that, for all w in $\mathcal{W}$,

$$|E(\zeta(Z,w)) - m| \leq \eta \;\Rightarrow\; d(w,\mathcal{W}^*) \leq \epsilon \tag{11}$$

Then let $\Omega_{\eta,N}$ be the following subset of Ω

$$\left\{\omega \in \Omega \quad \left|\frac{1}{N}\sum_{n=1}^{N}\zeta(\tilde{Z}_N^n,w) - E(\zeta(Z,w))\right| \leq \frac{\eta}{3}\right\}.$$

If $\omega \in \Omega_{\eta,N}$ then, as $\mathcal{W}$ is a compact set, we can find $N \in \mathbb{N}$, $w_N^*(\omega) \in \mathcal{W}$ which minimizes $\frac{1}{N}\sum_{n=1}^{N}\zeta(\tilde{Z}_N^n(\omega),w)$. Let w^* be in the closure of $\{w_N^*\}_N$; then by arguments similar to the ones used in the first part of the proof we show that, for all $\omega \in \Omega_{\eta,N}$ and for all $w \in \mathcal{W}$,

$$E(\zeta(Z, w^*)) \leq E(\zeta(z, w)) + \eta,$$

which implies by the use of (11) that

$$\Omega_{\eta,N} \subset \{\omega \ \ d(w^*(\omega), \mathcal{W}^*) \leq \epsilon\}$$

and this concludes the proof as $\lim_{N \to +\infty} P(\Omega_{\eta,N}) = 1$.

References

[1] Bishop (1995) *Neural network for pattern recognition.* New York : Oxford University Press.

[2] Bosq D. (1991) Modelization, non-parametric estimation and prediction for continuous time processes. *Roussas G. (Ed.), Nonparametric Functional estimation and related Topics, NATO, ASI Series*, 509-525.

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

[4] Chen T. and Chen H. (1995) Universal Approximation to Nonlinear Operators by Neural Networks with Arbitrary Activation Functions and Its Application to Dynamical Systems. *IEEE Transactions on Neural Networks*, **6**, 4, 911-917.

[5] Conan-Guez B. and Rossi F. (2002) Multi-Layer Perceptrons for Functional Data Analysis: a Projection Based Approach. *ICANN 2002, Madrid, Spain*, 667-672.

[6] Cook R.D. (1991) Discussion of Li. *Journal of the American Statistical Association*, **86**, 328-332.

[7] Cook R.D. and Weisberg S. (1991) Comment on Sliced Inverse Resgression for dimension reduction by K.C. Li. *Journal of the American Statistical Association*, **86**, 328-332.

[8] Dauxois J., Ferré L. and Yao A.F. (2003) Un modèle semi-paramétrique pour variables aléatoires hilbertiennes. *C.R. Acad. Sci. Paris*, **t. 333**, 947-952.

[9] Ferraty F. and Vieu P. (2002) The functional nonparametric model and application to spectrometric data. *Computational Statistics*, **17**, 515-561.

[10] Ferraty F. and Vieu P. (2003) Curves Discrimination: a Nonparametric Functional Approach. *Computational Statistics and Data Analysis*, **44**, 161-173.

[11] Ferré L. (1998) Determining the dimension in Sliced Inverse Regression and Related Methods. *Journal of the American Statistical Association*, **93**, 132-140.

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

[13] Ferré L. and Yao A.F. (2004) Smoothed functional inverse regression. *To appear in Statistica Sinica.*

[14] Ferré L. and Villa N. (2005) Discrimination de courbes par régression inverse fonctionnelle. *To appear in RSA.*

[15] Friedman J.H. (1989) Regularized Discriminant Analysis. *Journal of the American Statistical Association*, **84**, 405, 165-175.

[16] Hastie T., Buja A. and Tibshirani R. (1994) Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association*, **89**, 428, 1255-1270.

[17] Hastie T., Buja A. and Tibshirani R. (1995) Penalized Discriminant Analysis. *Annals of Statistics*, **23**, 73-102.

[18] Hsing T. and Carroll R.J. (1992) An asymptotic theory for sliced inverse regression. *Annals of Statistics*, **20**, 1040-1061.

[19] Hornik K. (1991) Approximation capabilities of multilayer feedforward networks. *Neural Networks*, **4**, 2, 251-257.

[20] Hornik K. (1993) Some new results on neural network approximation. *Neural Networks*, **6**, 8, 1069-1072.

[21] Ivanov V.K. (1962) On linear problems which are not well-posed. *Soviet Math. Docl.*, **145**, 2.

[22] James G.M., Hastie T.J. and Sugar C.A. (2000) Principal Component models for sparse functional data. *Biometrika*, **87**, 3, 587-602.

[23] Leurgans S.E., Moyeed R.A. and Silverman B.W. (1993) Canonical Correlation Analysis when the Data are Curves. *Journal of the Royal Statistical Society*, **55**, 3, 725-740.

[24] Li K.C. (1991) Sliced Inverse Regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316-342.

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

[26] Pezzulli S. and Silverman B.W. (1993) Some properties of smoothed principal componenets analysis for functional data. *Computational Statistics*, **8**, 1-16.

[27] Ramsay J.O. and Silverman B.W. (1996) *Functional Data Analysis.* New York : Springer Verlag.

[28] Rossi F., Conan-Guez B. and Fleuret F. (2002) Functional Data Analysis with Multi Layer Perceptrons, *IJCNN 2002 (part of WCCI) proceeding, Honolulu, Hawaii*, 2843-2848.

[29] Rossi F. and Conan-Guez B. (2003) Un modèle semi-paramétrique neuronal pour la régression et la discrimination sur données fonctionnelles. preprint 0338 on **http://www.ceremade.dauphine.fr**.

[30] Sandberg I.W. and Xu L. (1996) Network Approximation of Input-Output Maps and Functionals. *Circuits Systems Signal Processing*, **15**, 6, 711-725.

[31] Schott J.R. (1994) Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, **89**, 141-148.

[32] Silverman B.W. (1996) Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, **24**, 1-24.

[33] Stinchcombe M.B. (1999) Neural network approximation of continuous functionals and continuous functions on compactifications. *Neural Networks*, **12**, 3, 467-477.

[34] Tenorio L. (2001) Statistical Regularization of Inverse Problems. *Society for Industrial and Applied Mathematics*, **43**, 2, 347-366.

[35] Tihonov A.N. (1963a) Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Docl.*, **4**, 1036-1038.

[36] Tihonov A.N. (1963b) Regularization of incorrectly posed problems, *Soviet Math. Docl.*, **4**, 1624-1627.

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

[38] Velilla S. (1998) Assessing the number of linear component in a general regression problem. *Journal of the American Statistical Association*, **93**, 1088-1098.

[39] White H. (1989) Learning in Artificial Neural Network: A Statistical Perspective. *Neural Computation*, **1**, 425-464.

[40] Xia Y., Tong H., Li W.K. and Zhu L.X. (2002) An adaptative estimation of dimension reduction space. *Journal of the Royal Statistical Society* B., **64**, 363-410.

[41] Zhu L.X. and Fang K.T. (1996) Asymptotics for kernel estimate of sliced inverse regression. *Annals of Statistics*, **24**, 1053-1068.

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

List of Figures

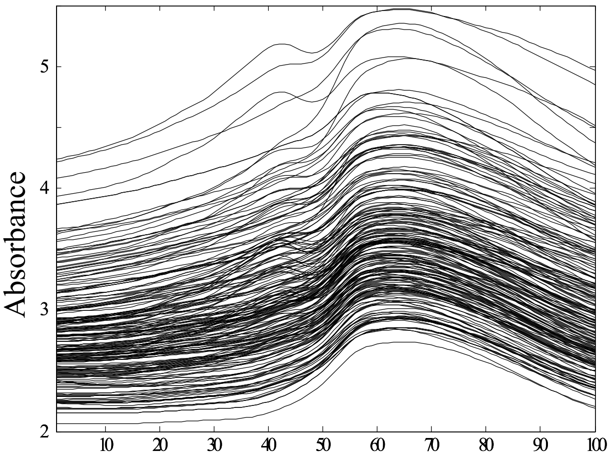

Figure 1: The regressor curves

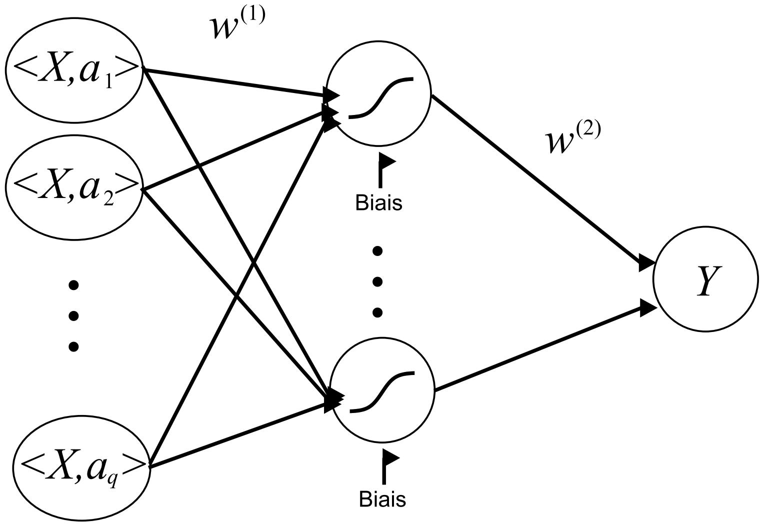

Figure 2: Neural network estimating f

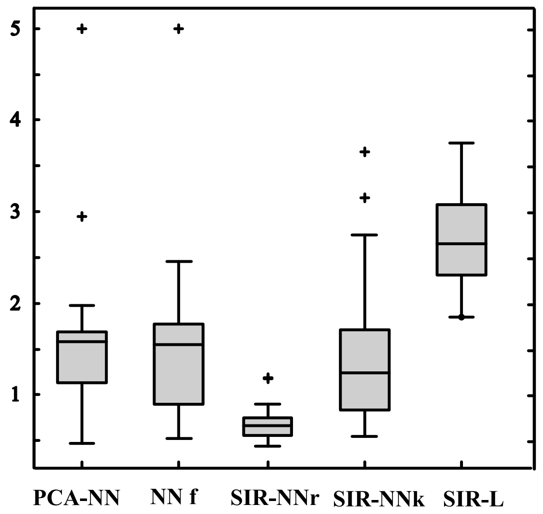

Figure 3: Tecator data set: SEP for 50 samples

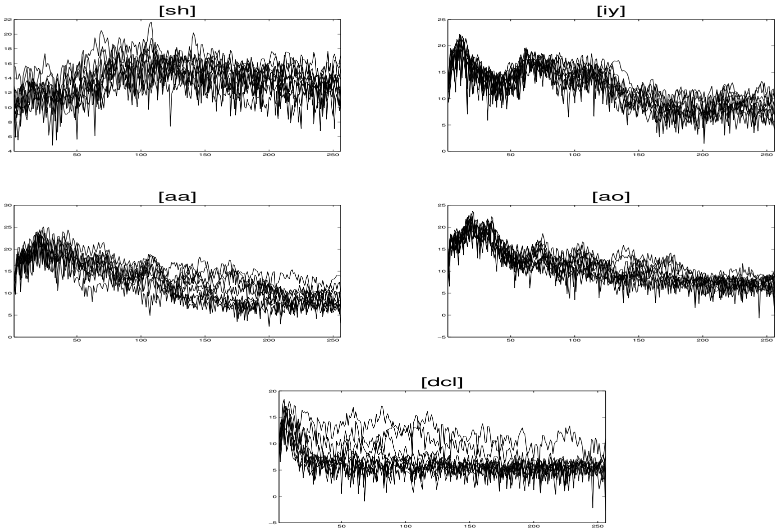

Figure 4: A sample of 10 log-periodograms per class

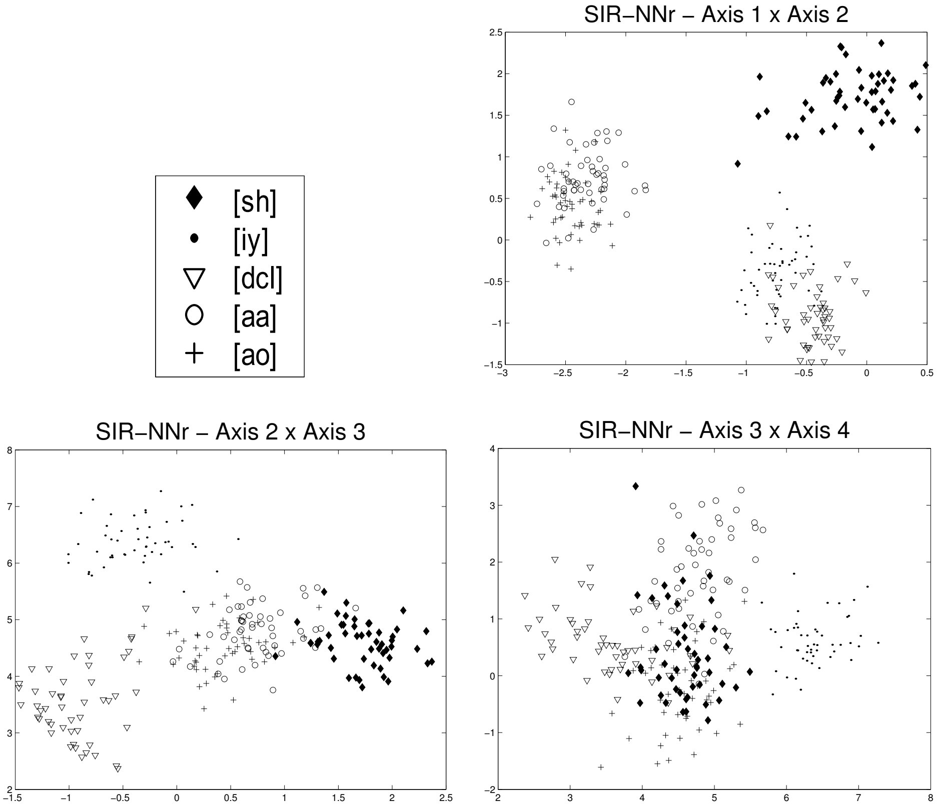

Figure 5: Projection on the EDR space of 50 log-periodograms by class

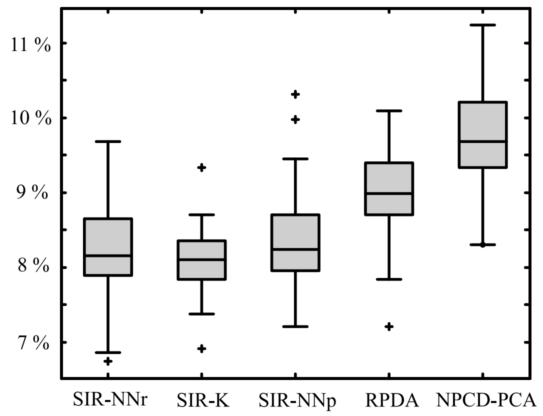

Figure 6: Phoneme Data: Test error rates for 50 samples

List of Tables

	Parameter 1	*Parameter 2*	*Parameter 3*
PCA-NN	$k_n = 25$ (PCA dimension)	$q_2 = 12$ (number of neurons)	
NNf	$q_2 = 18$ (number of neurons)		
SIR-NNr	$\alpha = 5$ (regularization of Γ_X)	$q = 20$ (SIR dimension)	$q_2 = 10$ (number of neurons)
SIR-NNk	$h = 0{,}5$ (kernel window)	q= 10 (SIR dimension)	$q_2 = 15$ (number of neurons)
SIR-L	$\alpha = 0{,}5$ (regularization of Γ_X)	$q = 20$ (SIR dimension)	

Table 1: Best parameters for the five compared methods

	Mean	*Median*	*Standard deviation*	*1st quartile*	*Minimum*
PCA-NN	1,74	1,59	1,82	1,14	0,47
NNf	1,55	1,55	1,13	0,90	0,52
SIR-NNr	0,68	0,66	0,16	0,56	0,44
SIR-NNk	1,40	1,24	0,71	0,84	0,54
SIR-L	2,70	2,64	0,48	2,31	1,84

Table 2: Tecator data set: Description of the performances

Methods	PCA-NN	NNf	SIR-NNr	SIR-NNk	SIR-L
Computational time (number of seconds per sample)	50	350	100	50	1

Table 3: Computational time for the five compared methods

	Parameter 1	*Parameter 2*	*Parameter 3*
SIR-NNr	$\alpha = 10$ (regularization of Γ_X)	$q = 4$ (SIR dimension)	$q_2 = 15$ (number of neurons)
SIR-NNp	$k_n = 17$ (PCA dimension)	$q = 4$ (SIR dimension)	$q_2 = 12$ (number of neurons)
SIR-K	$\alpha = 10^{-3}$ (regularization of Γ_X)	$q = 4$ (SIR dimension)	$h = 1$ (kernel bandwidth)
RPDA	$\alpha = 5$ (regularization of Γ_X)	q= 4 (PDA dimension)	
NPCD-PCA	$k_n = 7$ (PCA dimension)	$h = 25$ (kernel window)	

Table 4: Best parameters for the five compared methods

	Mean	*Median*	*Standard deviation*	*1st quartile*	*Minimum*
SIR-NNr	8,21 %	8,16 %	0,56 %	7,90 %	6,74 %
SIR-K	8,09 %	8,10 %	0,40 %	7,84 %	6,92 %
SIR-NNp	8,38 %	8,24 %	0,59 %	7,95 %	7,20 %
RPDA	8,95 %	8,99 %	0,54 %	8,70 %	7,20 %
NPCD-PCA	9,78 %	9,68 %	0,65 %	9,34 %	8,30 %

Table 5: Phonem Data: Test error rates