Inverse regression methods based on fuzzy partitions

Sandie Ferrigno[1], Ali Gannoun[2] and Jérôme Saracco[3,4]

[1] Institut Elie Cartan Nancy, Université Henri Poincaré Nancy 1
B.P. 239, F-54506 Vandoeuvre-lès-Nancy Cedex, France
e-mail: Sandie.Ferrigno@iecn.u-nancy.fr

[2] CNAM, Mathématiques CEDRIC
292 rue Saint Martin, 75141 Paris Cedex 03, France
e-mail : ali.gannoun@cnam.fr

[3] Université Bordeaux 1, Institut de Mathématiques de Bordeaux,
UMR CNRS 5251
351 cours de la libération, 33405 Talence Cedex, France
e-mail: Jerome.Saracco@math.u-bordeaux1.fr

[4] GREThA, UMR CNRS 5113, Université Montesquieu - Bordeaux IV
Avenue Léon Duguit, 33608 Pessac Cedex, France

Abstract: We consider a semiparametric regression model such that the dependent variable y is linked to some indices $x'\beta_k$ through an unknown link function. Li (1991) introduced sliced inverse regression methods (SIR-I, SIR-II and SIR$_\alpha$) in order to estimate the effective dimension reduction space spanned by the vectors β_k. These methods computationally fast and simple but are influenced by the choice of slices in the estimation process. In this paper, we suggest to use versions of SIR methods based on fuzzy clusters instead of slices which can be seen as hard clusters and we exhibit the corresponding algorithm. We illustrate the sample behaviour of the fuzzy inverse regression estimators and compare them with the SIR ones on simulation study.

AMS Subject Classification: 62H12, 62H30, 62G99.

Key Words: Dimension Reduction, Fuzzy Partition, Sliced Inverse Regression.

1. Introduction

The general goal of the regression of y on x is to infer about the conditional distribution of $y|x$ as far as possible with the avalaible data. In statistical literature, a lot of parametric and nonparametric regression approaches have been developed and studied. Dimension reduction in semiparametric regression is an important key theme in this area of statistics. The goal is to reduce the dimension of the covariate x without of loss information on the regression of the response y on x. We assume throughout the paper that the predictor x is a p-dimensional ($p \geq 2$) random variable and the response variable y is a scalar. It is also assumed that the data $\{(y_i, x_i),\ i = 1, \ldots, n\}$ are independant and identically distributed observations of (y, x) with finite moments.

In dimension reduction context, let $\mathcal{B}$ denote a fixed $p \times K^*$ matrix (with $K^* \leq p$) such that

$$y \perp x | \mathcal{B}'x. \tag{1}$$

This statement means that y and x are independent given any value for the random vector $\mathcal{B}'x$. In other words, it is equivalent to say that the distribution of $y|x$ is the same as that of $y|\mathcal{B}'x$ for all values of x in its sample space. A straightforward consequence is that the p-dimensional covariate x can be replaced by the K-dimensional predictor $\mathcal{B}'x$ without loss of regression information, thus the goal of dimension reduction is achieved since $K^* < p$. Note that (1) is trivially true when $\mathcal{B} = I_p$. Moreover, as it is mentionned in Li (1991) or Cook (1994), the statement (1) can be viewed as a statement about $S(\mathcal{B})$ the linear subspace of $\mathbb{R}^p$ spanned by the columns of $\mathcal{B}$: $y \perp x | P_{S(\mathcal{B})}x$, where $P_{S(\mathcal{B})}$ denotes the projection operator for $S(\mathcal{B})$. This subspace $S(\mathcal{B})$ is called dimension reduction subspace for the regression of y on x. The knowledge of the smallest dimension reduction subspace will be useful for parsimoniously characterizing the distribution of $y|x$. This subspace is the central dimension reduction subspace. In the following, the central subspace is such that $y \perp x | B'x$, where the columns of the $p \times K$ matrix B form a basis of the subspace. From a regression model point of view, on can mention that the model assumed by Li (1991) is the following

$$y = f(B'x, \varepsilon), \tag{2}$$

where f is the unknown link function, ε is the random error term independent of x.

In this paper, we will consider inverse regression methods in order to estimate a basis of $S(B)$. Many numerical methods have been introduced. Let us mention three of them which are relatively simple and easy to implement: sliced inverse regression, SIR (see for instance Li, 1991, or Saracco, 2001), principal Hessian directions, pHd (see Li, 1992, or Cook, 1998), and sliced average variance estimation, SAVE (see Cook and Weisberg, 1991, or Cook and Lee, 1999). All these methods require a linearity condition: the conditional expectation of $\mathbb{E}[x|B'x]$ is linear in $B'x$. In addition, methods based on second moments also require the constant variance condition that means that the conditional covariance matrix $\mathbb{V}(x|B'x)$ is constant. One can observe that both the linearity condition and the constant variance condition do not involve the response variable y, they are only applied to the distribution of the covariate x. Moreover, when the distribution of x is a p-dimensional normal distribution, these two conditions are satisfied. When x is elliptically distributed, the linearity condition holds. Finally, one can also mention that Hall and Li (1993) show that the linearity condition will hold to a reasonable approximation in many problems since p is large. In addition, Cook and Nachtsheim (1994) proposed to use predictor transformations and predictor weighting in order to induce these conditions.

In the following, we only focus on the SIR$_\alpha$ approach which is based on the property of the first and second moment of the inverse distribution of x given y. We replace the slicing step used in the estimation process by a fuzzy partition step. The paper is organized as follows. In Section 2, we give a brief overview on the SIR$_\alpha$ method. We describe the fuzzy partition method in Section 3. The inverse regression approach based on fuzzy partition is proposed in Section 4. Numerical results based on simulations are exhibited in Section 5 in order to study the efficiency of the fuzzy approach. Finally, concluding remarks are given in Section 6.

2. Brief review of SIR$_\alpha$ method

In SIR terminology, the linear subspace $S(B)$ is called the effective dimension reduction (e.d.r.) space, and any directions is this subspace are called e.d.r. directions.

Inverse regression step. The basic principle of SIR methods (SIR-I, SIR-II or SIR$_\alpha$) is to reverse the role of y and x, that is, instead of regressing the univariate variable y on the multivariate variable x, the covariable x is regressed on the response variable y.

The SIR-I estimates based on the first moment $\mathbb{E}(x|y)$ have been studied extensively (see for instance Duan and Li (1991), Li (1991), Carroll and Li (1992), Hsing and Carroll (1992), Zhu and Ng (1995), Kötter (1996), Saracco (1997), Aragon and Saracco (1997)).

But this approach is “blind” for symmetric dependencies (see Cook and Weisberg (1991) or Kötter (2000)). Then, SIR-II estimates based on the inverse conditional second moment $\mathbb{V}(x|y)$ have been suggested (see for instance Li (1991), Cook and Weisberg (1991) or Kötter (2000)). Hence these two approaches concentrate on the use of the inverse conditional moments $\mathbb{E}(x|y)$ or $\mathbb{V}(x|y)$ to find the e.d.r. space.

The idea of the SIR$_\alpha$ method is to conjugate the information from SIR-I and SIR-II in order to increase the chance of discovering all the e.d.r. directions. If an e.d.r. direction can only be marginally detected by SIR-I or SIR-II, a suitable combination of these two methods may sharpen the result.

Let us now recall the geometric properties of the model 2. Let T denote a monotonic transformation of y. Let $\alpha \in [0,1]$. Let $\mu = \mathbb{E}(x)$ and $\Sigma = \mathbb{V}(x)$. In order to conjugate information from the SIR-I and SIR-II approaches, Li (1991) consider the eigendecomposition of $\Sigma^{-1}M_\alpha$ where $M_\alpha = (1-\alpha)M_I\Sigma^{-1}M_I + \alpha M_{II}$. The matrices M_I and M_{II} are respectively the matrices used in the usual SIR-I and SIR-II approaches. They are defined as follows: $M_I = \mathbb{V}(\mathbb{E}(x|T(y)))$ and $M_{II} = \mathbb{E}\left(Q(y)\Sigma^{-1}Q(y)'\right)$ with $Q(y) = \mathbb{V}(x|T(y)) - \mathbb{E}(\mathbb{V}(x|T(y)))$ It can be shown that, under the linearity condition and the constant variance condition, the eigenvectors associated with the largest K eigenvalues of $\Sigma^{-1}M_\alpha$ are some e.d.r. directions. Let us remark that, when $\alpha = 0$ (resp. $\alpha = 1$), SIR$_\alpha$ is equivalent to SIR-I (resp. SIR-II).

Slicing step. Li (1991) proposed a transformation T, called a slicing, which categorizes the response y into a new response with $H > K$ levels. The support of y is partitioned into H non-overlapping slices $s_1, \ldots, s_h, \ldots, s_H$. With such transformation T, the matrices of interest are now written as $M_I = \sum_{h=1}^{H} p_h(m_h - \mu)(m_h - \mu)'$ and $M_{II} = \sum_{h=1}^{H} p_h \left(V_h - \overline{V}\right) \Sigma^{-1} \left(V_h - \overline{V}\right)$, where $p_h = P(y \in s_h)$, $m_h = \mathbb{E}(x|y \in s_h)$, $V_h = \mathbb{V}(x|y \in s_h)$ and $\overline{V} = \sum_{h=1}^{H} p_h V_h$.

Estimation process. It is straightforward to estimate the matrices Σ, M_I, M_{II} and M_α by substituting empirical versions of the moments for their theoretical counterparts, and therefore to obtain the estimation of the e.d.r. directions. Each estimated e.d.r. direction converges to an e.d.r. direction at rate $\sqrt{n}$, see for instance Li (1991) or Saracco (2001). Asymptotic normality of the SIR$_\alpha$

estimates has been studied by Gannoun and Saracco (2003a).

From a practical point of view, the choice of the slicing is discussed in Li (1991), Kötter (2000) and Saracco (2001). Since the SIR theory makes no assumption about the slicing strategy, the user must choose the number H of slices and a slicing strategy. In practice, there are naturally two possibilities: to fix the width of the slices or to fix the number of observations per slice (this second option is often preferred, and from the sample point of view, the slices are such that the number of observations in each slice is as close to each other as possible). Note that H must be greater than K in order to avoid artificial reduction of dimension. Li (1991) noticed that the choice of the slicing is less crucial than the choice of a bandwidth as in kernel-based methods.

One can mention that, in order to avoid the choice of a slicing, kernel-based estimate of SIR-I has been investigated, see Zhu and Fang (1996) or Aragon and Saracco (1997). However, these methods are hard to implement with regard to basic Slicing one and are computationally slow. Moreover, Bura (1997) and Bura and Cook (2001) proposed a parametric version of SIR-I. Note that determining the number K (of indices) is considered by Li (1991), Schott (1994), Ferré (1998) or Bai and He (2004), for the SIR-I method.

Moreover, the practical choice of the parameter α has been discussed in the literature. A test approach which does not require the estimation of the link function has been proposed by Saracco (2001). Two cross-validation criteria have been also developed by Gannoun and Saracco (2003b). Note that these criteria require a kernel smoothing estimation of the link function.

The aim of this paper is to replace the slicing step by a fuzzy partition step. Let us first introduce the notion of fuzzy partition.

3. Brief presentation of fuzzy partition

In this section, we give an overview of probabilistic fuzzy partition in H clusters. Let $\mathcal{T} = \{t_1, \ldots, t_n\}$ be a set of n elements. In hard (or crisp) partition methods, each element of the data set is assigned to exactly one cluster. For instance, an element lying between two clusters must be assigned to one of them. In fuzzy partition, each observation is given fractional membership in multiple clusters.

More precisely, a fuzzy partition of $\mathcal{T}$ into H clusters is a H-tuple $(u^{(h)})_{h=1,\ldots,H}$ of functions from $\mathcal{T}$ to $[0,1]$, each function describes a fuzzy cluster. The numerical value of $u^{(h)}(t_i) = u_i^{(h)}$ represents the membership degree of elements $t_i \in \mathcal{T}$ in cluster h. This value $u_i^{(h)}$ can be interpreted as the probability that element t_i belongs to the cluster h if for each $t_i \in \mathcal{T}$ the condition

$$\sum_{h=1}^{H} u_i^{(h)} = 1$$

is satisfied. For each observation i and each cluster h, the term $u_i^{(h)}$ indicates how strongly element t_i belongs to cluster h. Probabilistic partitions have been studied by many authors, see for instance Bezdek (1981) and Dumitrescu and Pop (1995, 1998) among others.

Let $\mathcal{M}_{nH}$ denote the set of real $n \times H$ matrices. We suppose that $H \geq 2$. A probabilistic fuzzy H-partition space associated with $\mathcal{T}$ can be defined as:

$$\begin{aligned} P_H &= \Big\{ U = \Big[u_i^{(h)} \Big] \in \mathcal{M}_{nH} \ : u_i^{(h)} \in [0,1], \\ &\qquad \textstyle\sum_{h=1}^{H} u_i^{(h)} = 1 \text{ for all } i, \ \sum_{i=1}^{n} u_i^{(h)} > 0 \text{ for all } h \Big\}. \end{aligned}$$

One can note that a hard H-partition space associated with $\mathcal{T}$ (as a slicing) is defined as

$$\begin{aligned} \widetilde{P}_H &= \Big\{ U = \Big[u_i^{(h)} \Big] \in \mathcal{M}_{nH} \ : u_i^{(h)} \in \{0,1\}, \\ &\qquad \textstyle\sum_{h=1}^{H} u_i^{(h)} = 1 \text{ for all } i, \sum_{i=1}^{n} u_i^{(h)} > 0 \text{ for all } h \Big\}, \end{aligned}$$

that is the values $u_i^{(h)}$ can only be equal to zero or one. It is obvious that $\widetilde{P}_H \subset P_H$.

Let us denote the distance of an element t_i to a cluster h determined by the prototype $\xi^{(h)}$ (generally a vector of the same dimension as the data vectors t_i to be interpreted as the cluster centers) by $d(t_i, \xi^{(h)})$. In order to obtain a fuzzy H-partition, Bezdek (1981) proposed to minimize the objective function

$$\sum_{i=1}^{n} \sum_{h=1}^{H} \left(u_i^{(h)} \right)^m d^2(t_i, \xi^{(h)})$$

subject to the constraints

$$\forall 1 \leq i \leq n, \ \sum_{h=1}^{H} u_i^{(h)} = 1 \quad \text{and} \quad \forall 1 \leq h \leq H, \ \sum_{i=1}^{n} u_i^{(h)} > 0, \qquad (3)$$

where the parameter m is chosen in advance (hard partition as $m \rightarrow 1$, totally fuzzy as $m \rightarrow \infty$; the value 2 for m is the most frequently used one). The objective function is minimized iteratively: in every iteration step, minimization with respect to $u_i^{(h)}$ and $\xi^{(h)}$ is done seperately. When the distance function is the euclidean one, the prototype is $\xi^{(h)} = \sum_{i=1}^n (u_i^{[h)})^m t_i / \sum_{i=1}^n (u_i^{[h)})^m$. Struyf et al. (1997) proposed to consider the memberships $u_i^{(h)}$ defined through the minimization of the objective function

$$\sum_{h=1}^{H} \frac{\sum_{i=1}^n \sum_{j=1}^n (u_i^{(h)})^2 (u_j^{(h)})^2 d(t_i, t_j)}{2 \sum_{i=1}^n (u_i^{(h)})^2}. \tag{4}$$

The minimization is carried out numerically by means of an iterative algorithm taking into account the above conditions (3) that the memberships need to obey. Note that to have an idea of how fuzzy the resulting clustering is, a coefficient (called Dunn's partition coefficient) can be computed: $C_H = \sum_{i=1}^n \sum_{h=1}^H \frac{(u_i^{(h)})^2}{n}$. This coefficient C_H always lies in $[\frac{1}{H}, 1]$. It attains its extreme values in the following situations: $C_H = \frac{n}{n} = 1$ for a hard partition (all $u_i^{(h)} = 0$ or 1), $C_H = nH\frac{1}{nH^2} = \frac{1}{H}$ for an entirely fuzzy partition (all $u_i^{(h)} = \frac{1}{H}$). In the simulation study, we will use the Struyf et al.'s approach in order to produce the probabilistic fuzzy partition needed in the estimation process. To conclude on fuzzy partition, one can mention that these clustering methods were designed to be robust.

4. Inverse regression based on fuzzy partition

Consider a sample $\{(x_i, y_i),\ i = 1, \ldots, n\}$. Let $\bar{x}$ and $\widehat{\Sigma}$ be the sample mean and the sample variance matrix of the x_i's. Let α be a fixed value in $[0, 1]$.

Step 1. Apply a probabilistic fuzzy H-partition on the data set $\{y_1, \ldots, y_n\}$. Let $\{u_i^{(h)},\ i = 1, \ldots, n$ and $h = 1, \ldots, H\}$ denote the corresponding membership.

Step 2. For each h, compute the following quantities:
- the "size" of the cluster h: $\tilde{\eta}^{(h)} = \sum_{i=1}^n u_i^{(h)}$, one can note that $\sum_{h=1}^H \tilde{\eta}^{(h)} = n$;
- the "weight" of the cluster h: $\tilde{p}^{(h)} = \frac{\tilde{\eta}^{(h)}}{n}$, one can note that $\sum_{h=1}^H \tilde{p}^{(h)} = 1$;

- the "center" of the cluster h: $\tilde{m}^{(h)} = \frac{1}{\tilde{\eta}^{(h)}} \sum_{i=1}^{n} u_i^{(h)} x_i = \sum_{i=1}^{n} \tilde{p}_i^{(h)} x_i$ where $\tilde{p}_i^{(h)} = \frac{u_i^{(h)}}{\tilde{\eta}^{(h)}}$;

- the "variance matrix" of the cluster h: $\tilde{V}^{(h)} = \sum_{i=1}^{n} \tilde{p}_i^{(h)} (x_i - \tilde{m}^{(h)})((x_i - \tilde{m}^{(h)})'$.

Step 3. Compute the matrix $\tilde{M}_\alpha = (1-\alpha)\tilde{M}_I \widehat{\Sigma}^{-1} \tilde{M}_I + \alpha \tilde{M}_{II}$, where $\tilde{M}_I = \sum_{h=1}^{H} \tilde{p}^{(h)} (\tilde{m}^{(h)} - \bar{x})(\tilde{m}^{(h)} - \bar{x})'$, $\tilde{M}_{II} = \sum_{h=1}^{H} \tilde{p}^{(h)} (\tilde{V}^{(h)} - \tilde{V}) \widehat{\Sigma}^{-1} (\tilde{V}^{(h)} - \tilde{V})'$ and $\tilde{V} = \sum_{h=1}^{H} \tilde{p}^{(h)} \tilde{V}^{(h)}$.

Step 4. Compute the eigen decomposition of the matrix $\widehat{\Sigma}^{-1} \tilde{M}_\alpha$. Let $\tilde{b}_k,\ k = 1, \ldots, K$ be the eigenvectors associated with the largest eigenvalues. The linear subspace $\tilde{E} = S(\tilde{B})$ is the estimated e.d.r. space, where $\tilde{B} = \left[\tilde{b}_1, \ldots, \tilde{b}_K\right]$.

Remark. If we consider the following "hard" rule: $\forall i = 1, \ldots, n$, $u_i^{*(h)} = 1$ if $h = \arg\max_l u_i^{(l)}$ and 0 otherwise, we come again on the original SIR$_\alpha$ approach with a specific construction of slices.

5. Simulation studies

In this section, simulation studies are carried out to provide evidence for the efficiency of the fuzzy approach in pratice. We first introduce the efficiency measure used as the criterion that measures the distance between the estimated e.d.r. space and the true e.d.r. space. Then we describe the simulated models and the estimation methods. Finally, we comment on the results of the performed simulation studies. All computational work was carried out in Splus.

5.1 Efficiency measure

In order to measure the distance between the estimated e.d.r. space $\tilde{E}$, spanned by the column of $\tilde{B}$, and the true e.d.r. space E, spanned by the column of B, we introduce the Σ-orthogonal projectors on $\tilde{E}$ and E: $P_{\tilde{E}} = \tilde{B}(\tilde{B}'\Sigma\tilde{B})^{-1}\tilde{B}'\Sigma$ and $P_E = B(B'\Sigma B)^{-1}B'\Sigma$. An efficiency measure for the estimates is defined as:

$$m(\tilde{E}, E) = \frac{\mathrm{Tr}(P_{\tilde{E}} P_E)}{K},$$

where K is the dimension of the linear subspaces E and $\tilde{E}$. This measure takes values in the interval $[0, 1]$. Note that

(i) if $\tilde{E} = E$, then $m(\tilde{E}, E) = 1$;

(ii) if $\tilde{E}$ and E are Σ-orthogonal, then $m(\tilde{E}, E) = 0$;

(iii) the closer the measure is to one, the better is the estimation. One can also note that, for $K = 1$, this measure corresponds to the squared cosine of the angle between $\tilde{b}_1$ and β_1.

5.2 Models and estimation methods for simulation

We consider a single index regression model

$$\text{(m1):} \quad y = (x'\beta_1)^2 \exp(x'\beta_1/N) + \varepsilon, \tag{5}$$

and a two indices regression model

$$\text{(m2):} \quad y = (x'\beta_1)^2 + (x'\beta_2)^2 + \varepsilon. \tag{6}$$

In these two models, the variable x and the error term ε are independent and respectively follow the normal distributions $\mathcal{N}_p(0_p, I_p)$ and $\mathcal{N}(0,1)$, where 0_p is the p-dimensional null vector and I_p is the $p \times p$ identity matrix. In performing the simulation, we fix $\beta_1 = (1, 1, -1, -1, 0'_{p-4})'$ and $\beta_2 = (1, -1, 0'_{p-4}, -1, 1)'$. The dimension p will be set at 5 or 10.

We select models (5) and (6) based on the following considerations. In model (5), the parameter N has clearly an influence on the form of the dependence betwen the index $x'\beta$ and the response variable y. When N is small (for instance $N = 1$), the exponential term is preponderant: then model (5) favors methods based the first inverse conditional moments as SIR-I or SIR$_\alpha$ for small values of α, because the regression function is strictly increasing. When N is large (for instance $N = 100$), the influence of the exponential term disappears and the squared polynomial part is preponderant in model (5): hence, the model favors here methods based the second inverse conditional moments as SIR-II or SIR$_\alpha$ for large values of α ($\alpha > 0.5$), because the regression model presents a symmetric dependence. Between these two extreme cases (for instance when $N = 5$), the exponential and polynomial parts carry information on the e.d.r. direction: the one can expect that the inverse regression methods which conjugate information on the first two inverse conditional moments (as SIR$_\alpha$) provide good estimation. In performing the simulation with this model (5), we consider three situations: $N = 1$, $N = 5$ and $N = 100$. Since model (6) clearly presents a symmetric dependence, it favors methods using informations from the inverse conditional variance. It allows us to see the performance of our approaches for a multiple indices model.

Acronyms for the estimation methods used in the simulation studies are given in Table 1.

SIR-I	Sliced Inverse Regression based on the matrix $\hat{M}_I$
SIR-II	Sliced Inverse Regression based on the matrix $\hat{M}_{II}$
SIRa	Sliced Inverse Regression based on the matrix $\hat{M}_\alpha$
FIR-I	Fuzzy Inverse Regression based on the matrix $\tilde{M}_I$
FIR-II	Fuzzy Inverse Regression based on the matrix $\tilde{M}_{II}$
FIRa	Fuzzy Inverse Regression based on the matrix $\tilde{M}_\alpha$

Table 1: *Acronyms for the different estimation methods*

In order to check the impact of the number H of slices or clusters, for the sliced or fuzzy inverse regression methods, several values for H are considered: $H = 2$, 3, 4, 5, 10 or 20. We generate samples of n observations (with $n = 50$, 100, 300 or 500) from models (5) or (6). The slices are built such that the numbers of observations in slices never differ by more than one (these numbers are equal to $[n/H]$ or $[n/H] + 1$ where $[a]$ denotes the integer part of a). The H clusters are obtained with the "fanny" function of Splus using euclidean distances and the objective function (4).

We conduct two simulation studies: the first one focuses on model (5) and the second one concerns models (6). For each study, we generate 100 Monte Carlo samples with the specific values of n and p precised above. For each simulated sample, we estimate the e.d.r. space with the fuzzified methods (FIR-I, FIR-II, FIR$_\alpha$) and the sliced ones (SIR-I, SIR-II, SIR$_\alpha$) for the different values of the parameter H. Then, for each estimated e.d.r. space, we evaluate the corresponding efficiency measure to evaluate the quality of the estimation. Note that, in all simulation studies, the value of α is set at 0.5; the FIR$_\alpha$ method corresponds to a fuzzified version of sliced average variance estimation (SAVE) method of Cook (2000). A discussion about the choice of an optimal α is given in Section 6.

5.3 Results of simulation studies

Results of the first simulation study. We only report the results obtained with the FIR$_\alpha$ and SIR$_\alpha$ methods for illustration since the SIR-I, SIR-II, FIR-I or FIR-II methods can not structurally recover the e.d.r. space when the regression model is symmetric dependent or not. Figure 1 shows boxplots of the effenciency measures calculated with the FIR$_\alpha$ estimates from a total of 100 samples generated from model (5) for several values of N, with $n = 100$ or 300, $p = 5$ or 10 and $H = 2$, 4, 5, 10 and 20. One can observe that:

- When the model (5) does not present a symmetric dependence ($N = 1$), the fuzzy approach works well with reasonable values of H (4, 5 and 10) with respect to $K = 1$. Even if the FIR_α is not the most adequate method for this model (recall that this kind of model favors FIR-I or SIR-I approaches), one can however be satisfied by the numerical performance of this method since the number of fuzzy clusters is reasonable. Except for one specific situation ($p = 10$ and $H = 2$) which is the less favourable one, the FIR_α method seems to perform uniformly better than the SIR_α one.

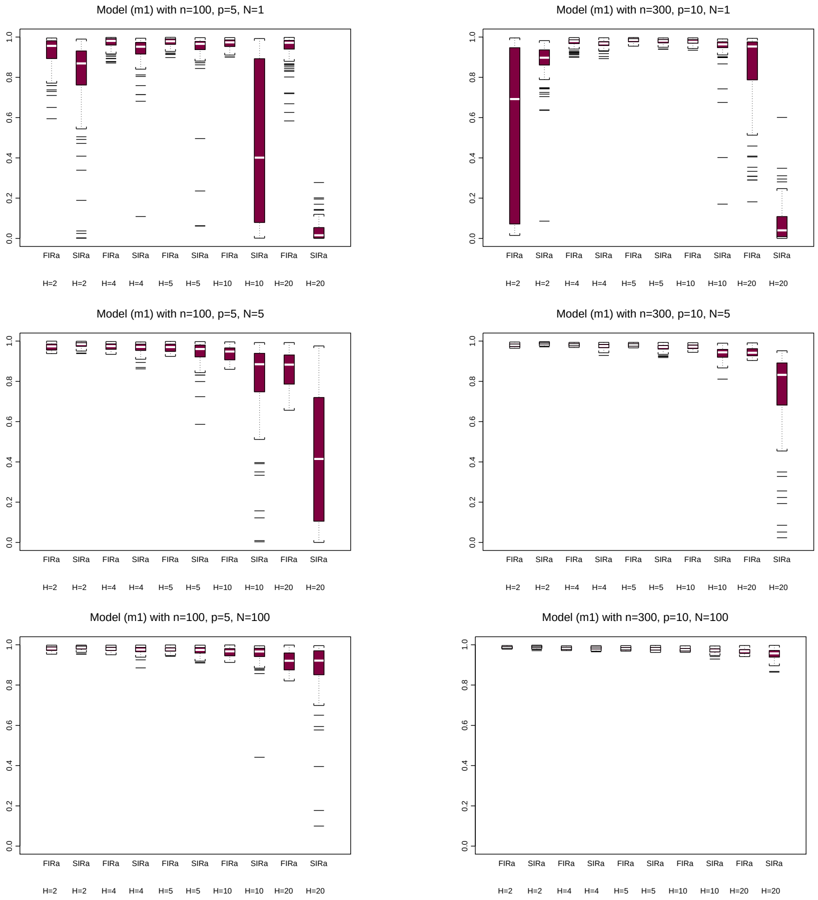

Figure 1: *Model (m1): Boxplots of the efficiency measures for various values of N, n and p.*

- When $N = 5$ (model (5) is “partially” symmetric dependent), FIR$_\alpha$ and SIR$_\alpha$ methods work well since H is not too large. One can observe the significant improvement obtained with FIR$_\alpha$version in all the situations.

• With the symmetric dependent case (with $N = 100$), the results are very similar to the previous ones. Note that the sensitivity to the parameter H is very low, this is particularly true for large values of n.

Results of the second simulation study. We does not report here the results obtained with the FIR-I and SIR-I methods since these two methods are known to be unable to recover the e.d.r. space when the regression model is symmetric dependent. Figure 2 shows scatterplots of the effenciency measures calculated with the FIR-II estimates versus those calculated with the FIR$_\alpha$ estimates, for samples generated from model (6) with $n = 50$, 100 or 300, and $p = 5$ or 10. We clearly observe that these methods give very similar results in terms of this efficiency measure. This is not really surprising since we have already mentioned that model (6) favors the methods using information from the inverse covariance curve. Therefore, we only report the results obtained with the FIR-II and SIR-II methods for illustration in Figure 3.

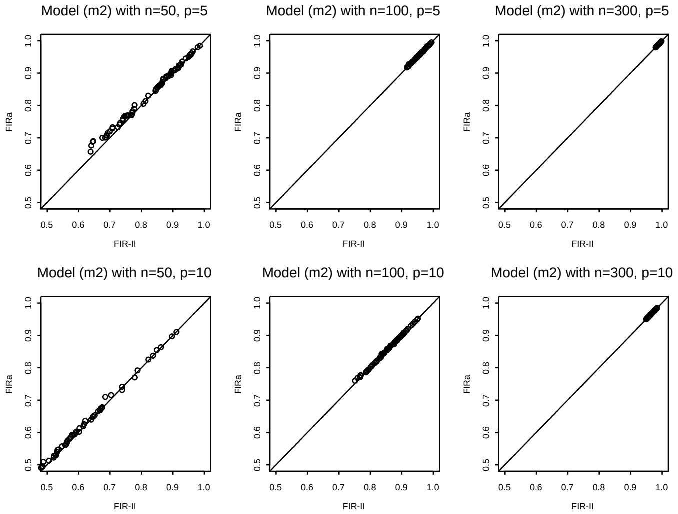

Figure 2: *Model (m2): plots of the efficiency measures of the FIR-II estimates versus the FIR$_\alpha$ estimates for various values of n and p.*

The boxplots of Figure 3 report the distribution of the efficiency

measure of the FIR-II and SIR-II estimates of a total of 100 Monte Carlo from samples generated from the two indices models (6) with different sample sizes (n = 50, 100, 300) and different dimensions of the covariable x (p = 5 or 10). Since in this simulation study, we know that $K = 2$, we systematically consider the two first e.d.r. directions $[\tilde{b}_1, \tilde{b}_2]$ as estimates of a basis of the true e.d.r. space. No choice of the optimal dimension has been implemented at this stage. This point will be discussed in Section 6.
Now we can compare the efficiency of the fuzzy and sliced approaches, FIR-II and SIR-II, and check the impact of the number H of fuzzy clusters or slices for several values of n and p.

• As expected, the two methods perform better as the sample size increases. When $n \geq 300$, the true e.d.r. space is always retrieved by the methods for all values of H. One can notice that SIR-II appears to be more sensitive to high value of this parameter H than FIR-II. We do not report here the results for $n = 500$.

• For reasonable sample sizes (n = 50 or 100), the methods have more difficulties to find the true e.d.r. space; this is particulary true when the dimension of the covariable x become larger ($p = 10$). One can however observe that the FIR-II method works uniformly better than the SIR-II approach. One can also mention that the parameter H seems to have an influence on the quality of the estimates; in particular small values for H seems to be preferable to large values for small sample sizes.

6. Concluding remarks

In this paper, we propose to use a fuzzified version of sliced inverse regression methods. From a practical point of view, we illustrate on simulation the smaple performance of this approach: the FIR methods appears to perfom better than the original SIR methods for most of H (number of fuzzy clusters or slices). Some important issues remain to be discussed.

• The first one concerns the practical number H of fuzzy clusters. For the original SIR methods, to the best of our knowledge, the choice of an “optimal” number H of slices remains unsolved. Here, from a theoretical point of view, the asymptotic convergence of the FIR estimates has been obtained for any $H > K$ where K is the number of indices in the regression model. One can observe on simulation that the methods seems to be not very sensitive to H since the sample size n is relatively large ($n \geq 300$). For “small”

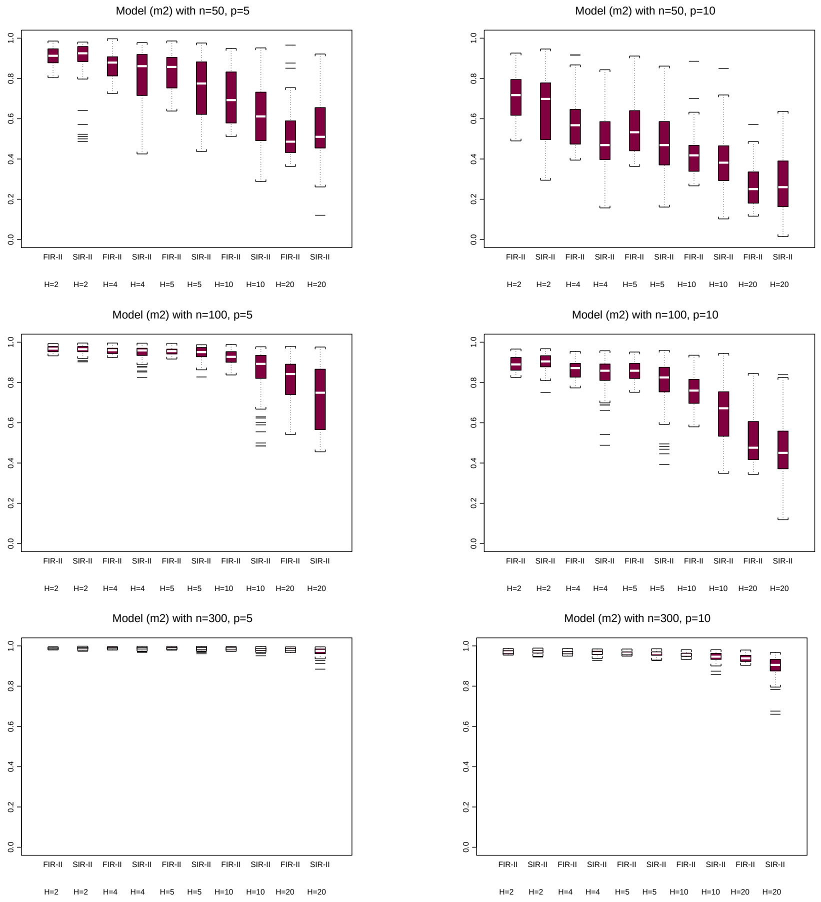

Figure 3: *Model (m2): Boxplots of the efficiency measures for various values of* N*,* n *and* p*.*

sample sizes ($n \leq 100$), we suggest that the parameter H must be chosen nor too large nor too small: for instance, in our simulation studies (with $K = 1$ or 2), $H = 5$ seems to be a good practical choice. Note that, from a computational point of view, the smaller the number H is, the faster is the estimation calculation.

A natural way to smooth the influence of the arbitrary fixed number H of fuzzy clusters could be to combine the results from several fuzzy partitions. Then one can expect a “robust” version of the different estimated matrices and of their estimated eigenvectors. This is idea has already been used in the slicing approach, the main idea of Pooled Slicing methods introduced by Aragon and Saracco (1997) and Saracco (2001). In our context, one could also consider

D different fuzzy partitions with different numbers of clusters. We will have to choose minimum and maximal bounds for H such that $H_{max} - H_{min} = D$. Then, we could consider the following pooled variance-covariance matrices for the pooled version of FIR-I and FIR-II methods: $\tilde{M}_I^P = \frac{1}{D}\sum_{d=1}^{D} \tilde{M}_I^d$ and $\tilde{M}_{II}^P = \frac{1}{D}\sum_{d=1}^{D} \tilde{M}_{II}^d$ where $\tilde{M}_I^d$ (resp. $\tilde{M}_{II}^d$) is the matrix $\tilde{M}_I$ (resp. $\tilde{M}_{II}$) of FIR-I (resp. FIR-II) defined for a fuzzy partitions in d clusters, $d = 1, \ldots, D$. For the FIR$_\alpha$ approach, one can combine these two matrices and we consider the matrix $\tilde{M}_\alpha^P = (1-\alpha)(\tilde{M}_I^P)^2 + \alpha\tilde{M}_{II}^P$. One can proved that the first K eigenvectors of $\tilde{M}_I^P$ (resp. $\tilde{M}_{II}^P$, $\tilde{M}_\alpha^P$) are *e.d.r.* directions. It could be possible to consider another alternative for the pooled version of FIR$_\alpha$: for each $d = 1, \ldots, D$, let us define $\tilde{M}_\alpha^d = (1-\alpha)\left(M_I^d\right)^2 + \alpha M_{II}^d$ and use the pooled matrix $\tilde{M}_\alpha^P = \frac{1}{D}\sum_{d=1}^{D} \widetilde{M_\alpha^d}$.

• The second important issue to discuss concerns the number K of indices. In practice, the user has to choose the parameter. One practical manner to determine the dimension of the e.d.r. space can be to look at the screeplot of the eigenvalues of the matrix of interest and to retain the number of eigenvalues significantly greater than the others. To illustrate this empirical approach, we give in Figure (4) two examples of screeplot. For the first one,we generate $n = 100$ observations from model (5) with $K = 1$. For the second, a $n = 100$ sample has been generated from model (6) with $K = 2$. We clearly observe that the selected dimensions with this empirical criterion are respectively one and two. Note that hypothesis testing procedures or bootstrap procedures could also

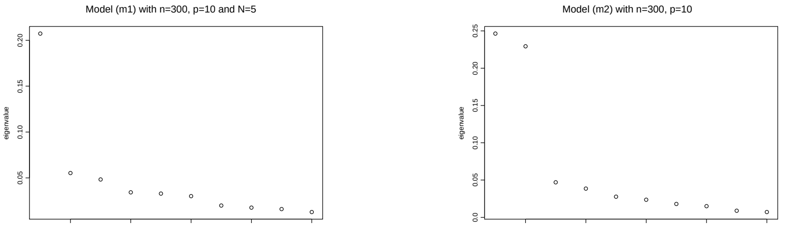

Figure 4: *FIR$_\alpha$ eigenvalues screeplots for samples generated from model (m1) or model (m2) with* $n = 300$*,* $p = 10$*,* $\alpha = 0.5$*,* $H = 5$ *(and* $N = 5$ *for (m1)).*

be developped in order to determine the dimension K. This point is at the moment a challenging issue.

• Another important issue is about the choice of the parameter α for the FIR$_\alpha$ method. In the simulation, the parameter is set at 0.5, in order to have a fuzzified version of sliced average variance estimation (SAVE) which takes into account information from the first two inverse conditional moments. In practice, for instance one can be interested in getting an "optimal" value for α in view of prediction of the dependent variable y. It is also possible to use cross validation (CV) criterion in order to select α, see Gannoun and Saracco (2003b). To be more precise, the procedure is to leave out the ith observation, $i = 1, \ldots, n$, and to predict the corresponding observed value from the remaining subsample of size $n-1$. The idea is then to choose the parameter α that yields the best prediction.

One can note that a "global" approach may be to define a predictive CV criterion to estimate at the same time the parameter α, the dimension K and the number of fuzzy clusters H. The price to pay will be the high computational cost. This point is still under investigation.

• When the dependent variable y is multivariate (see for instance Aragon (1997), Hsing (199) Li et al. (2003), Saracco (2005), or Liquet and Saracco (2007) for a presentation of some inverse regression methods in this multidimensional context), that is $y \in \mathbb{R}^q$, the fuzzy partitions step still works. Therefore, all the results presented in the paper for FIR$_\alpha$ remain respectively true and operational.

To conclude, we presented in this paper an extension of the well-known dimension-reduction method SIR$_\alpha$ to fuzzified version FIR$_\alpha$ which gives interesting results in simulations. Some issues mentioned above remain challenging issue and we leave both of them as future work.

References

Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. *Computational Statistics*, **12**, 355-372.

Aragon, Y. and Saracco, J. (1997). Sliced Inverse Regression (SIR): an appraisal of small sample alternatives to slicing. *Computational Statistics*, **12**, 109-130.

Bai, Z. D. and He, X. (2004). A chi-square test for dimensionality for non-Gaussian data. *Journal of Multivariate Analysis*, **88**, 109-117.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function* *algorithms. With a foreword by L. A. Zadeh. Advanced Applications in Pattern Recognition.* Plenum Press, New York-London.

Bura, E. (1997). Dimension reduction via parametric inverse regression. *L_1-statistical procedures and related topics (Neuchtel, 1997), IMS Lecture Notes Monogr. Ser.*, **31**, 215-228.

Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society, Series B*, **63**, 393–410.

Carroll, R. J. and Li, K. C.(1992). Measurement error regression with unknown link: dimension reduction and data visualization. *Journal of the American Statistical Association*, **87**, 1040-1050.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression". *Journal of the American Statistical Association*, **86**, 328-332.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177-189.

Cook, R. D. (1998). Principal Hessian directions revisited. With comments by Ker-Chau Li and a rejoinder by the author. *Journal of the American Statistical Association*, **93**, 84-100.

Cook, R. D. (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods* , **29**, 2109-2121.

Cook, R. D. and Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association*, **94**, 1187-1200.

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, **89**, 592-599.

Duan, N. and Li, K.C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, **19**, 505-530.

Dumitrescu, D. and Pop, H.F. (1995). Degenerate and non-degenerate convex decomposition of finite fuzzy partitions. I. *Fuzzy Sets and Systems*, **73**, 365-376.

Dumitrescu, D. and Pop, H.F. (1998). Degenerate and non-degenerate convex decomposition of finite fuzzy partitions. II. *Fuzzy Sets and Systems*, **96**, 111-118.

Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association*, **93**, 132-140.

Gannoun, A. and Saracco, J. (2003a). An asymptotic theory for SIR$_\alpha$ method. *Statistica Sinica*, **13**, 297-310.

Gannoun, A. and Saracco, J. (2003b). Two Cross Validation Criteria for SIR$_\alpha$ and PSIR$_\alpha$ methods in view of prediction. *Computational Statistics*, **18**, 585-603.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**, 867-889.

Hsing, T. and Carroll, R. J. (1992). An asympotic theory for Sliced Inverse regression. *The Annals of Statistics*, **20**, 1040-1061.

Hsing, T. (1999). Nearest neighbor inverse regression. *The Annals of Statistics*, **27**, 697-731.

Kötter, T. (1996). An asymptotic result for Sliced Inverse Regression. *Computational Statistics*, **11**, 113-136.

Kötter, T. (2000). Sliced Inverse Regression. In *Smoothing and Regression. Approaches, Computation, and Application* (Edited by M. G. Schimek), 497-512. Wiley, New York.

Li, K. C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association*, **86**, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, **87**, 1025-1039.

Li, K. C., Aragon, Y., Shedden, K. and Thomas Agnan, C. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, **98**, 99-109.

Liquet, B. and Saracco, J. (2007). Pooled marginal slicing approach via SIRa with discrete covariables. To appear in *Computational Statistics.*

Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics - Theory and methods*, **26**, 2141-2171.

Saracco, J. (2001). Pooled Slicing methods versus Slicing methods. *Communications in Statistics - Simulation and Computation*, **30**, 489-511.

Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIRa approach. *Journal of Multivariate Analysis*, **96**, 117-135.

Schott, J. R. (1994). Determining the dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association*, **89**, 141-148.

Struyf, A., Hubert, M. and Rousseeuw, P. J. (1997). Integrating robust clustering techniques in S-PLUS. *Computational Statistics & Data Analysis*, **26**, 17-37 .

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

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of Sliced Inverse Regression. *The Annals of Statistics*, **24**, 1053-1068.