Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Cluster-based Sliced Inverse Regression

Vanessa Kuentz [a,b], Jérôme Saracco [a,b,c,∗]

[a] *Université de Bordeaux, IMB, UMR CNRS 5251, 351 Cours de la Libération, 33405 Talence Cedex, France*

[b] *INRIA Bordeaux Sud-Ouest, CQFD team, France*

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

ARTICLE INFO

Article history:
Received 20 February 2009
Accepted 24 August 2009
Available online 11 September 2009

AMS 2000 subject classifications:
primary 62G05
secondary 62D05

Keywords:
Sliced Inverse Regression (SIR)
Effective dimension reduction (e.d.r.) space
Clustering
Linearity condition

ABSTRACT

In the theory of sufficient dimension reduction, Sliced Inverse Regression (SIR) is a famous technique that enables us to reduce the dimensionality of regression problems. This semiparametric regression method aims at determining linear combinations of a p-dimensional explanatory variable $\mathbf{x}$ related to a response variable y. However it is based on a crucial condition on the marginal distribution of the predictor $\mathbf{x}$, often called the linearity condition. From a theoretical and practical point of view, this condition appears to be a limitation. Using an idea of Li, Cook, and Nachtsheim (2004) in the Ordinary Least Squares framework, we propose in this article to cluster the predictor space so that the linearity condition approximately holds in the different partitions. Then we apply SIR in each cluster and finally estimate the dimension reduction subspace by combining these individual estimates. We give asymptotic properties of the corresponding estimator. We show with a simulation study that the proposed approach, referred as cluster-based SIR, improves the estimation of the e.d.r. basis. We also propose an iterative implementation of cluster-based SIR and show in simulations that it increases the quality of the estimator. Finally the methodology is applied on the horse mussel data and the comparison of the prediction reached on test samples shows the superiority of cluster-based SIR over SIR.

1. Introduction

Parametric regression models are used to highlight the relationship between one response variable y and a p-dimensional explanatory variable $\mathbf{x} = (x^1, \dots, x^p)'$, with $\mathbb{E}(\mathbf{x}) = \mu$ and $\mathbb{V}(\mathbf{x}) = \Sigma$, via for instance the following model:

$$y = f_\theta(\mathbf{x}) + \varepsilon,$$

where f_θ belongs to a family of parametric functions described by θ (vector of real parameters) and ε is a random error. The aim is the estimation of θ, which can be reached for example by maximum likelihood or least squares methods. These techniques work well if the family of f_θ is well specified. However this identification can turn out very difficult in some applications. Nonparametric regression models are then a possible solution. They offer a larger flexibility since they do not formulate any parametric assumption on the link function f. For instance, the model can be written:

$$y = f(\mathbf{x}) + \varepsilon.$$

These methods are essentially based on the property of continuity and derivability of the unknown regression function f. However they only provide good numerical results with low-dimensional explanatory variable. Indeed with high-dimensional problems, the number of observations needed to get information about the local behaviour of f becomes enormous. This is the well-known curse of dimensionality that can be challenged by dimension reduction models.

∗ Corresponding author at: Université de Bordeaux, IMB, UMR CNRS 5251, 351 Cours de la Libération, 33405 Talence Cedex, France.
E-mail addresses: vanessa.kuentz@math.u-bordeaux1.fr (V. Kuentz), jerome.saracco@math.u-bordeaux1.fr (J. Saracco).

1226-3192/$ – see front matter © 2009 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jkss.2009.08.004

Many dimension reduction tools assume that the features of $\mathbf{x}$ can be captured in a lower K-dimensional projection subspace (with $K < p$), such as Sliced Inverse Regression (SIR) methods introduced by Li (1991). They enable us to estimate a basis of this linear subspace. The corresponding model assumes that the dependency between the predictors and the response variable is described by linear combinations of the predictors. The underlying semiparametric model is written:

$$y = f(\mathbf{x}'\beta_1, \dots, \mathbf{x}'\beta_K, \varepsilon), \tag{1}$$

where f is an unknown function, ε is an unknown random error independent of $\mathbf{x}$, and $\beta_1, \dots, \beta_K$ are K unknown vectors in $\mathbb{R}^p$, assumed to be linearly independent. As none condition on the form of f is imposed, the vectors β_k are not identifiable. It is only possible to estimate the space spanned by these vectors, called the effective dimension reduction (e.d.r.) space, which will be denoted by E. When K is small ($K \ll p$), the goal of reduction theory is achieved and we can project the p-dimensional regressor $\mathbf{x}$ onto this K-dimensional space without loss of information on the feature of y given $\mathbf{x}$. Then it will be easier to study the relationship between $\mathbf{x}$ and y via a nonparametric estimation of the regression of y on the corresponding K-dimensional variable.

The basic principle of SIR methods is to reverse the role of y and $\mathbf{x}$ and to study the property of the conditional moments of $\mathbf{x}$ given y. In this paper, we will only focus on the SIR-I method (denoted by SIR hereafter) which is based on the first conditional moment. To facilitate the estimation of the inverse conditional mean, a slicing on the response variable y is realized. Let us denote by T this transformation of y. SIR is a method based on geometrical properties. Indeed, Li (1991) has shown that the centered inverse regression curve, $\mathbb{E}(\mathbf{x}|T(y)) - \mathbb{E}(\mathbf{x})$ as y varies, is contained in the linear subspace of $\mathbb{R}^p$ spanned by the vectors $\Sigma\beta_1, \dots, \Sigma\beta_K$. A straightforward consequence is that the covariance matrix $M_I = \mathbb{V}(\mathbb{E}(\mathbf{x}|T(y)))$ is degenerated in any direction Σ-orthogonal to the β_k's. Therefore the eigenvectors associated with the nonnull K eigenvalues of the matrix $\Sigma^{-1}M_I$ are e.d.r. directions, that is are in the e.d.r. space. One important point in SIR theory is the underlying crucial linearity condition:

$$\mathbb{E}(\mathbf{x}'b|\mathbf{x}'\beta_1, \dots, \mathbf{x}'\beta_K) \text{ is linear in } \mathbf{x}'\beta_1, \dots, \mathbf{x}'\beta_K \quad \text{for any } b. \tag{2}$$

This condition is hard to verify in practice since it involves the unknown directions of the e.d.r. space. However it can be proved that (2) is verified when $\mathbf{x}$ follows an elliptically symmetric distribution, condition which is stronger in theory but easier to verify in practice. A special case is the multinormality of $\mathbf{x}$.

If the collected data set does not follow an elliptically distribution, solutions exist to force data to behave as if they were issued from such a distribution. For instance, if the dimension of $\mathbf{x}$ is small, two alternatives appear. The first one is the normal resampling of the data proposed by Brillinger (1983). The idea is to simulate a normal sample of same size as the original data. These simulated points are called "attractors". Then the principle is to select for each attractor the nearest point of the original sample. Note that some points of the original data set can be chosen several times while others will never be selected. Then the distribution of the selected points is more "normal" than the one of the original observations. The second solution is the re-weighting and trimming scheme of Cook and Nachtsheim (1994). The principle is to define a discrete probability measure that will assign weights to the observations and that is close to one elliptically symmetric distribution. This target elliptical distribution is based on Minimum Volume Ellipsoid (MVE), which enables us to trim a specified proportion of extreme points. Then the choice of one discrete distribution that approaches the target one is reached by Voronoi weights and Dirichlet cells. A problem with this technique is that it can severely reduce the sample size. Moreover these two methods are difficult to put into practice if $\mathbf{x}$ is high-dimensional. However Hall and Li (1993) showed with a bayesian argument that (2) approximately holds for high-dimensional data sets. So they argued that it is not a severe restriction in practice, implying that a blind application of these methods is not dangerous and will produce "good" estimations of the e.d.r. directions, when the dimension p is large.

In this paper we propose to cluster the predictor space, which will force the linearity condition to hold approximately in each cluster. The idea is inspired by the work of Li et al. (2004), who proposed a cluster-based Ordinary Least Squares (OLS) approach for single index models ($K = 1$). It consists in partitioning the predictor space with a k-means algorithm, evaluating the OLS estimate of each cluster and finally pooling them so as to provide an efficient estimation of the central mean subspace. In our approach, we also partition the predictor space into disjoint clusters with a k-means algorithm, which aims at constructing approximately elliptical clusters. Then we estimate the e.d.r directions in each cluster and combine them to produce an efficient estimation of the e.d.r space of model (1). The proposed approach will be referred in the rest of the paper as cluster-based SIR.

Note that for some special data structure cases, k-means can have relative poor performance on elliptical clusters: for instance, a mixture of two elongated components or the some special data structures such as Swiss roll or S-curve, see Cheung (2003) or Everitt, Landau, and Leese (2001) for details and discussion.

In Section 2, we consider the case of single index model, we describe the population and sample approaches of the cluster-based SIR. We show the convergence in probability and the asymptotic distribution of the corresponding estimator of the e.d.r. direction. We extend this approach to multiple indices models in Section 3. A simulation study is carried out in Section 4 in order to show the numerical performance of the approach and to compare it with SIR and cluster-based OLS. We also propose an iterative implementation and show with simulations that the quality of the estimated e.d.r. basis is improved. A real data application is reported in Section 5 to show the predictive performance of cluster-based SIR versus SIR. Finally concluding remarks are given in Section 6.

2. Approach for single index model

We consider in this section single index model ($K = 1$). The corresponding model is:

$$y = f(\mathbf{x}'\beta, \varepsilon). \tag{3}$$

So we focus on the estimation of only one e.d.r. direction b colinear to β. The idea of the proposed approach is to partition the predictor space into a fixed number c of clusters. By doing that, the linearity condition will approximately hold in each cluster. For each one, we compute the e.d.r. direction with SIR. Finally we combine these directions to find the e.d.r direction of model (3) taking into account the whole space.

2.1. Population version

Let us consider a fixed number c of clusters and let us assume that $\mathbf{x}$ is partitioned into c clusters $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(c)}$. Accordingly to the partitioning scheme of $\mathbf{x}$, we get the partition $(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \ldots, c$ of $(\mathbf{x}, y)$. Let us assume that the linearity condition holds in each cluster:

(LC) For $j = 1, \ldots, c$, $\quad \mathbb{E}(\mathbf{x}^{(j)\prime}b|\mathbf{x}^{(j)\prime}\beta)$ is linear in $\mathbf{x}^{(j)\prime}\beta$ for any b.

In each cluster j, let $T^{(j)}$ be the slicing of $y^{(j)}$ into $H^{(j)}$ fixed slices, $s_1^{(j)}, \ldots, s_{H^{(j)}}^{(j)}$, with $H^{(j)} > 1$. From this slicing, the matrix $M_I^{(j)}$ can be written as $M_I^{(j)} = \sum_{h=1}^{H^{(j)}} p_h^{(j)}(m_h^{(j)} - \mu^{(j)})(m_h^{(j)} - \mu^{(j)})'$, where $p_h^{(j)} = P(y^{(j)} \in s_h^{(j)})$, $m_h^{(j)} = \mathbb{E}(\mathbf{x}^{(j)}|y^{(j)} \in s_h^{(j)})$ and $\mu^{(j)} = \mathbb{E}(\mathbf{x}^{(j)})$. Let $\Sigma^{(j)} = \mathbb{V}(\mathbf{x}^{(j)})$. The eigenvector $b^{(j)}$ associated with the largest eigenvalue of the matrix $(\Sigma^{(j)})^{-1}M_I^{(j)}$ is an e.d.r. direction. We define the matrix $B = [b^{(1)}, \ldots, b^{(c)}]$ and we note b the first left singular vector of this matrix. Then Theorem 1 guarantees that this vector is an e.d.r. direction.

Theorem 1. *Assuming the linearity condition* (LC) *and model* (3)*, the major eigenvector b of the matrix BB' is colinear with β.*

Proof. For each $j = 1, \ldots, c$, $b^{(j)}$ is colinear with β, i.e. $b^{(j)} = \alpha_j\beta$, where α_j is a nonnull real. As $B = [\alpha_1\beta, \ldots, \alpha_c\beta]$, we have:

$$BB' = \sum_{j=1}^{c} \alpha_j^2 \beta\beta' = \|\alpha\|^2 \beta\beta',$$

where $\alpha = (\alpha_1, \ldots, \alpha_c)'$ and $\|.\|$ is the norm associated to usual scalar product. Then the eigenvector b associated with the strictly positive eigenvalue of BB' is colinear with β. □

2.2. Sample version

Let $S = \{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ be a sample from the reference model (3). We partition these observations into c clusters using a k-means approach. Note that one hundred initial random sets are chosen and the best partitioning is retained, i.e. the one that provides the minimum sum of squares from points to the assigned cluster centers. By this way, it stabilizes the clustering step and then cluster-based SIR approach. Moreover our k-means algorithm is constrained to avoid sparse clusters. We set the minimum number of points in a slice at $n_{h,\min}$, which implies that the minimum number of observations in the jth cluster is $n_{\min}^{(j)} = n_{h,\min} \times H^{(j)}$. If one cluster obtained with classical k-means contains less than $n_{\min}^{(j)}$ observations, it is merged with the nearest cluster, in sense of Ward criterion. Finally we advise a maximum number of clusters for cluster-based SIR defined as $C_{\max}^n = \frac{n}{H \times n_{h,\min}}$. So for $j = 1, \ldots, c$, we get samples $S^{(j)} = \{(y_i^{(j)}, \mathbf{x}_i^{(j)\prime}), i = 1, \ldots, n^{(j)}\}$. In each cluster, the empirical mean and covariance matrix of the $\mathbf{x}_i$'s are respectively given by $\overline{\mathbf{x}}^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbf{x}_i^{(j)}$ and $\widehat{\Sigma}^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} (\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}^{(j)})(\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}^{(j)})'$. The matrix $M_I^{(j)}$ is estimated by $\widehat{M_I}^{(j)} = \sum_{h=1}^{H^{(j)}} \hat{p}_h^{(j)}(\hat{m}_h^{(j)} - \overline{\mathbf{x}}^{(j)})(\hat{m}_h^{(j)} - \overline{\mathbf{x}}^{(j)})'$ with $\hat{p}_h^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbb{I}_{[y_i \in s_h^{(j)}]}$ and $\hat{m}_h^{(j)} = \frac{1}{n^{(j)}\hat{p}_h^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbf{x}_i^{(j)} \mathbb{I}_{[y_i \in s_h^{(j)}]}$, where the notation $\mathbb{I}$ designates the indicator function. Then the eigenvector $\hat{b}^{(j)}$ associated with the largest eigenvalue of $(\widehat{\Sigma}^{(j)})^{-1}\widehat{M_I}^{(j)}$ is the estimated e.d.r. direction in the jth cluster. We construct the matrix $\widehat{B} = [\hat{b}^{(1)}, \ldots, \hat{b}^{(c)}]$. The major eigenvector $\hat{b}$ of the matrix $\widehat{BB'}$ is then the e.d.r. estimated direction in model (3).

2.3. Asymptotic theory

In what follows, the notation $Z_n \to_d Z$ means that Z_n converges in distribution to Z as $n \to \infty$. The assumptions that are necessary to state our results are gathered below for easy reference.

(A1) The sample S is a sample of independent observations from the single index model (3) or the multiple indices model (1).
(A2) $\mathbf{x}$ is partitioned into c fixed clusters $\mathbf{x}^{(j)}, j = 1, \ldots, c$, such that $\cup_{j=1}^{c} \mathcal{S}^{(j)} = \mathcal{S}$ and $\forall j \neq l$, $\mathcal{S}^{(j)} \cap \mathcal{S}^{(l)} = \emptyset$.

(A3) The support of $y^{(j)}$ is partitioned into a fixed number $H^{(j)}$ of slices such that $p_h^{(j)} \neq 0$, $h = 1, \dots, H^{(j)}$.
(A4) For $j = 1, \dots, c$, $n^{(j)} \to \infty$ as $n \to \infty$.

Comment on (A4). With the proposed k-means step avoiding sparse clusters, we have for each cluster $j = 1, \dots, c$, $n_h^{(j)} \geq n_{h,\min}$ and then $n^{(j)} \geq n_{\min}^{(j)} = H^{(j)} \times n_{h,\min}$, where $n_{h,\min} = \frac{n}{C_{\max}^n \times H^{(j)}}$. In order to get $n^{(j)} \to \infty$ as $n \to \infty$, we can choose for instance $C_{\max}^n = O\left((n/H^{(j)})^{1/2}\right)$.

We show in Theorem 2 the convergence in probability of the cluster-based SIR estimator and give its asymptotic distribution in Theorem 3.

Theorem 2. *Under the linearity condition* (LC) *and the assumptions* (A1)–(A4), *we have* $\hat{b} = b + O_p(n^{-1/2})$, *where b is an e.d.r. direction (colinear with* β*).*

Proof. Li (1991) has shown for SIR that the estimated e.d.r. direction converges to an e.d.r. direction at root n rate. So under the assumptions of the theorem, for each $\mathcal{S}^{(j)}$, $j = 1, \dots, c$, we have

$$\hat{b}^{(j)} = b^{(j)} + O_p(n^{-1/2}).$$

Then we get $\widehat{B} = B + O_p(n^{-1/2})$ and $\widehat{B}\widehat{B}' = BB' + O_p(n^{-1/2})$. Thus the major eigenvector of $\widehat{\widehat{B}\widehat{B}'}$ converges to the corresponding one of BB' at the same rate: $\hat{b} = b + O_p(n^{-1/2})$. From Theorem 1, b is colinear with β. So the estimated e.d.r. direction obtained with cluster-based SIR converges to an e.d.r. direction at root n rate. □

Theorem 3. *Under the linearity condition* (LC) *and the assumptions* (A1)–(A4), *we have:*

$$\sqrt{n}(\hat{b} - b) \longrightarrow_d U \sim \mathcal{N}(0, \Gamma_U),$$

where the expression of Γ_U *is given in* (18)*.*

The proof of Theorem 3 is given in the Appendix.

2.4. Optimal number of clusters

In practice, a crucial step in the proposed method is the choice of the number c of clusters for the partitioning of the predictor space. The choice of an optimal number c^* of clusters can be defined through the following optimization problem:

$$c^* = \arg \min_{c=1,\dots,C} \mathbb{E}((y - \mathbb{E}(y|\mathbf{x}'\hat{b}_{[c]}))^2), \tag{4}$$

where $\hat{b}_{[c]}$ denotes the estimator of the e.d.r. direction when the number of clusters is c.
From a practical point of view, we consider an empirical smoothed version of this minimization problem:

$$\hat{c}^* = \arg \min_{c=1,\dots,C} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{i,[c]})^2, \tag{5}$$

where $\hat{y}_{i,[c]} = \sum_{j=1}^n y_j \mathcal{K}((\mathbf{x}_i'\hat{b}_{[c]} - \mathbf{x}_j'\hat{b}_{[c]})/h_c) / \sum_{j=1}^n \mathcal{K}((\mathbf{x}_i'\hat{b}_{[c]} - \mathbf{x}_j'\hat{b}_{[c]})/h_c)$ is a kernel estimation of $\mathbb{E}(y|\mathbf{x}_i'\hat{b}_{[c]})$, for which h_c is the bandwidth parameter for a partitioning into c clusters and $\mathcal{K}$ is a kernel (the density of the standard univariate normal distribution for instance). The bandwidth parameters h_c, $c = 1, \dots, C$, can be chosen by cross validation.

3. Extension to multiple indices model

In this section, we extend the proposed approach to multiple indices model ($K > 1$). The corresponding model is given in (1). We search for a basis that spans the e.d.r. space $E = \text{Span}(\beta_1, \dots, \beta_K)$.

3.1. Population version

As for the single index model, we partition the predictor space $\mathbf{x}$ into c clusters. We get the partitions $(\mathbf{x}^{(j)}, y^{(j)})$, $j = 1, \dots, c$. For each cluster, we seek with SIR a basis of the e.d.r. space. Let us assume that the following linearity condition (LC*) holds:

(LC*) For $j = 1, \dots, c$, $\mathbb{E}(\mathbf{x}^{(j)\prime}b|\mathbf{x}^{(j)\prime}\beta_1, \dots, \mathbf{x}^{(j)\prime}\beta_K)$ is linear in $\mathbf{x}^{(j)\prime}\beta_1, \dots, \mathbf{x}^{(j)\prime}\beta_K$ for any b.

The eigenvectors $b_1^{(j)}, \dots, b_K^{(j)}$ associated with the largest K eigenvalues of the matrix $(\Sigma^{(j)})^{-1}M_I^{(j)}$ are e.d.r. directions, where matrices $\Sigma^{(j)}$ and $M_I^{(j)}$ have been defined in Section 2. We define the matrix $B^{(j)} = [b_1^{(j)}, \dots, b_K^{(j)}]$ containing these e.d.r. directions, which form a $\Sigma^{(j)}$-orthogonal basis of E. Then the first K eigenvectors of the matrix $B^{(j)}B^{(j)\prime}$, denoted by $\tilde{b}_1^{(j)}, \dots, \tilde{b}_K^{(j)}$, form an I_p-orthonormal basis of E. We store these vectors in the matrix $\tilde{B}^{(j)} = [\tilde{b}_1^{(j)}, \dots, \tilde{b}_K^{(j)}]$. We can now pool the matrices $\tilde{B}^{(j)}$ in the matrix $\mathbb{B}^{(c)} = [\tilde{B}^{(1)}, \dots, \tilde{B}^{(c)}]$. The first K eigenvectors of the matrix $\mathbb{B}^{(c)}\mathbb{B}^{(c)\prime}$ are denoted by $\tilde{\tilde{b}}_1, \dots, \tilde{\tilde{b}}_K$.

Theorem 4. *Assuming the linearity condition* (LC*) *and model* (1), *the vectors* $\tilde{\tilde{b}}_1, \ldots, \tilde{\tilde{b}}_K$ *form an* I_p*-orthogonal basis of the e.d.r. space* E.

Proof. Since $\tilde{b}_1^{(j)}, \ldots, \tilde{b}_K^{(j)}$ form an I_p-orthonormal basis of E, we have $\text{Span}(\mathbb{B}^{(c)}) = E$. Then the eigenvectors associated with the K largest eigenvalues of $\mathbb{B}^{(c)}\mathbb{B}^{(c)'}$ form an I_p-orthonormal basis of E. □

3.2. Sample version

As for the single index model, we estimate in each cluster a basis of the e.d.r space: the first K eigenvectors of the matrix $(\widehat{\Sigma}^{(j)})^{-1}\widehat{M_I}^{(j)}$ defined in Section 2. These vectors form a $\widehat{\Sigma}^{(j)}$-orthogonal basis of the estimated e.d.r. space. We store them in the matrix $\widehat{B}^{(j)} = [\hat{b}_1^{(j)}, \ldots, \hat{b}_K^{(j)}]$. Then the first K eigenvectors of the matrix $\widehat{B}^{(j)}\widehat{B}^{(j)'}$, denoted by $\hat{\tilde{b}}_1^{(j)}, \ldots, \hat{\tilde{b}}_K^{(j)}$, form an I_p-orthogonal basis of the estimated e.d.r. space. We store them in the matrix $\hat{\tilde{B}}^{(j)} = [\hat{\tilde{b}}_1^{(j)}, \ldots, \hat{\tilde{b}}_K^{(j)}]$. Let $\hat{\mathbb{B}}^{(c)} = [\hat{\tilde{B}}^{(1)}, \ldots, \hat{\tilde{B}}^{(c)}]$. Finally the first K eigenvectors of the matrix $\hat{\mathbb{B}}^{(c)}\hat{\mathbb{B}}^{(c)'}$, denoted by $\hat{\tilde{\tilde{b}}}_1, \ldots, \hat{\tilde{\tilde{b}}}_K$, form an I_p-basis of the estimated e.d.r. space.

3.3. Asymptotic theory

Under the linearity condition (LC*) and the assumptions (A1)–(A4), SIR theory provides $\widehat{B}^{(j)} = B^{(j)} + O_p(n^{-1/2})$. Then the first K eigenvectors of the matrix $\widehat{B}^{(j)}\widehat{B}^{(j)'}$ converge at same rate to the corresponding ones of $B^{(j)}B^{(j)'}$. Analogously $\hat{\mathbb{B}}^{(c)} = \mathbb{B}^{(c)} + O_p(n^{-1/2})$ and $\hat{\mathbb{B}}^{(c)}\hat{\mathbb{B}}^{(c)\prime} = \mathbb{B}^{(c)}\mathbb{B}^{(c)\prime} + O_p(n^{-1/2})$. Finally $\hat{\tilde{\tilde{b}}}_k = \tilde{\tilde{b}}_k + O_p(n^{-1/2})$, $k = 1, \ldots, K$, then the estimated e.d.r. basis converges to an e.d.r. basis at root n rate.
As for the single index model, using Delta method and asymptotic results of Saracco (1997) and Tyler (1981), the asymptotic normality of the eigenprojector onto the estimated e.d.r. space can be obtained, as well as the asymptotic distribution of the estimated e.d.r. direction, associated with eigenvalues assumed to be different.

3.4. Choice of dimension K and number of clusters

Until now we have supposed that the dimension K of the reduction model is known. However in most applications this dimension is a priori unknown and hence must be estimated from the data. From a practical point of view, we recommend to choose the dimension K using classical SIR. Several approaches have been proposed in the literature for SIR. The first type of approaches are hypothesis tests based on the nullity of the last $(p-K)$ eigenvalues, see Li (1991), Schott (1994) or Barrios and Velilla (2007). Another approach relies on a quality measure based on the square trace correlation between the true e.d.r. space and its estimate, see for instance Ferré (1998) or Liquet and Saracco (2008).

With this choice $\hat{K}$ of K, we can determine the optimal number of clusters using the kernel estimator method proposed in (5). From a theoretical point of view, when $\hat{K} > 1$ the kernel $\mathcal{K}$ can be replaced by a multidimensional one. From a practical point of view, as soon as $\hat{K} > 2$, this method is no more appropriate due to the curse of dimensionality. However in real applications, this dimension is seldom larger than 2.

Finally for the chosen couple of parameters $(\hat{K}, \hat{c}^*)$, it is important to check if there is a relevant structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{\tilde{\tilde{b}}}_1, \ldots, \mathbf{x}_i'\hat{\tilde{\tilde{b}}}_{\hat{K}}), i = 1, \ldots, n\}$ and to verify that there is no structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{\tilde{\tilde{b}}}_{\hat{K}+1}), i = 1, \ldots, n\}$ if we assume that the dimensions is $\hat{K} + 1$. This methodology is that used for the real data application in Section 5.

4. Simulation study

A simulation study is carried out to evaluate the numerical performance of the proposed method. First we recall the definition of the efficiency measure. Then, in a first stage we consider a single index model and compare the results obtained with cluster-based SIR with those provided by classical SIR and cluster-based OLS (Li et al., 2004). In a second stage, we compare the results obtained with SIR and cluster-based SIR on a multiple indices model (with $K = 2$). Finally we present an iterative version of the cluster-based SIR approach.

In the simulation study, we set $H = H^{(j)} = 4$ for SIR and cluster-based SIR. Moreover according to comment on (A4) we used for instance for sample size $n = 200$ (resp. 500), $C_{\max}^n = 7$ (resp. 12) and $n_{h,\min} = 6$ (resp. 11) in the selection of the number of clusters and in our modified k-means approach.

4.1. Efficiency measure

Let $\check{b}_1, \ldots, \check{b}_K$ be the K estimated e.d.r. directions. We note $\check{B} = [\check{b}_1, \ldots, \check{b}_K]$ and $\check{E} = \text{Span}(\check{B})$ the linear subspace spanned by the $\check{b}_k$'s. Let $B = [\beta_1, \ldots, \beta_K]$ be the matrix of the true directions and let $E = \text{Span}(B)$. Let P_E (resp. $P_{\check{E}}$) be the I_p-orthogonal projector onto E (resp. $\check{E}$) defined as follows: $P_E = B(B'B)^{-1}B'$ and $P_{\check{E}} = \check{B}(\check{B}'\check{B})^{-1}\check{B}'$. Since with cluster-based

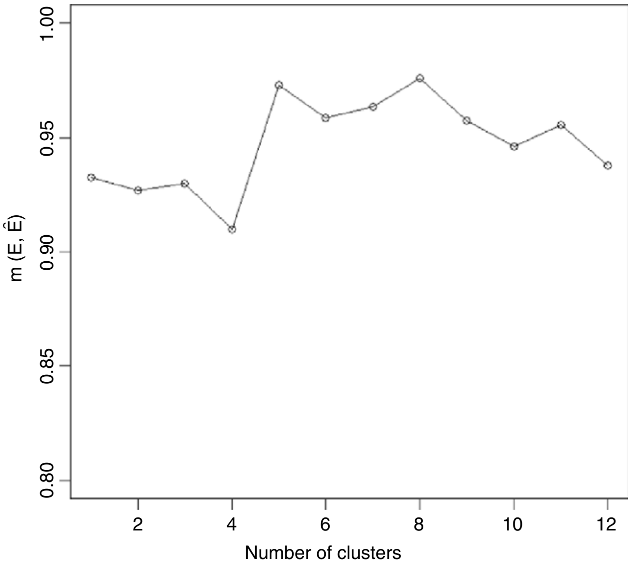

Fig. 1. Efficiency measure (6) of the estimation obtained with SIR ($c = 1$) and cluster-based SIR ($c > 1$) for model (7) with $n = 500$ and $\theta = 0$.

SIR approach we construct an I_p-orthonormal basis of E (that is $\check{B}'\check{B} = I_K$) and if we choose the β_k's such that $B'B = I_K$, the expression of the projectors reduces to: $P_E = BB'$ and $P_{\check{E}} = \check{B}\check{B}'$.
The quality of the estimate $\check{E}$ of E is measured by:

$$m(E, \check{E}) = \text{Trace}(P_E P_{\check{E}})/K. \tag{6}$$

This measure belongs to [0, 1] with $m(E, \check{E}) = 0$ if $\check{E} \perp E$ and $m(E, \check{E}) = 1$ if $\check{E} = E$. Therefore the closer this value is to one, the better is the estimation. When $K = 1$ (single index model), this measure is the squared cosine of the angle formed by the vectors β and $\check{b}$.

4.2. Single index model

First we define the simulated model, then we describe our approach on a sample when the linearity condition is not verified. Finally we generate multiple data replications for which the linearity condition may be seriously violated or not.

4.2.1. Simulated model

We consider the following regression model:

$$y = \exp(x_1 - x_2) + \varepsilon, \tag{7}$$

with $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)'$, where $x_j \sim (1 - \theta) \times \text{Exp}(1) + \theta \times \mathcal{N}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. The variables x_j are mutually independent and the error term ε is independent of $\mathbf{x}$. In this model, the true normalized direction is $\beta = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, 0, 0, 0, 0)'$. In our simulations, the parameter θ will belong to [0, 1]. The value 0 corresponds to non-elliptical distribution (in this case the linearity condition is not verified) and 1 to multinormal distribution.

4.2.2. Single data replication

We exhibit a sample of size $n = 500$ of model (7) with $\theta = 0$. Fig. 1 shows the evolution of the quality criterion (6) as the number of clusters varies between 1 and 12. The case with one cluster matches with classical SIR. The maximum is reached at 8 clusters with an efficiency measure of 0.98, against 0.93 for classical SIR. The estimation of the e.d.r. direction is then improved by the use of cluster-based SIR. Fig. 2 shows the evolution of the empirical criterion (5) as the number of clusters increases. With this measure, we would choose the optimal number of clusters in sense of the quality measure (6). However we have observed in the simulation study that we do not always choose the same number of clusters as with (6). But the chosen number of clusters always provides a measure $m(E, \widehat{E})$ equal or very close to the maximum. Clearly criterion (5) is the only one appropriate criterion from a practical point of view, since criterion (6) requires knowledge of the true basis of the e.d.r. space, which is unknown in real applications. By contrast, criterion (5) is always estimable in practice but it involves a kernel estimation which is computationally expensive because it needs to introduce a tuning parameter (the bandwidth) chosen by cross validation. To reduce the computational cost of the rest of the simulation study (as we know the true e.d.r. direction) the optimal number of clusters will be chosen with the efficiency measure (6).

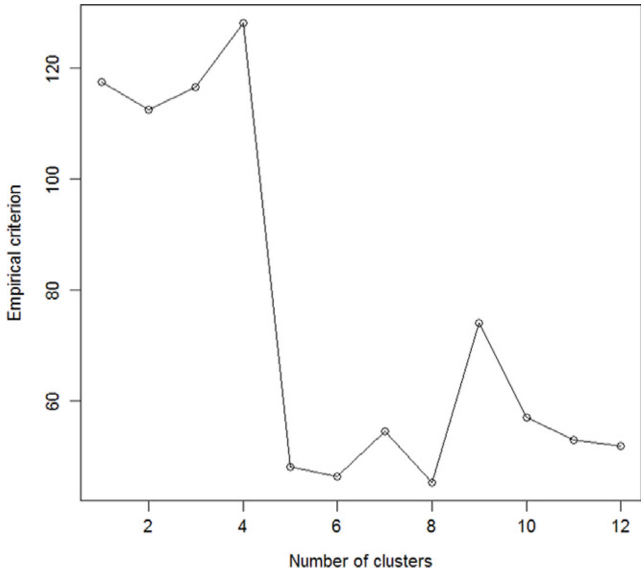

Fig. 2. Empirical criterion (5) obtained with SIR ($c = 1$) and cluster-based SIR ($c > 1$) for model (7) with $n = 500$ and $\theta = 0$.

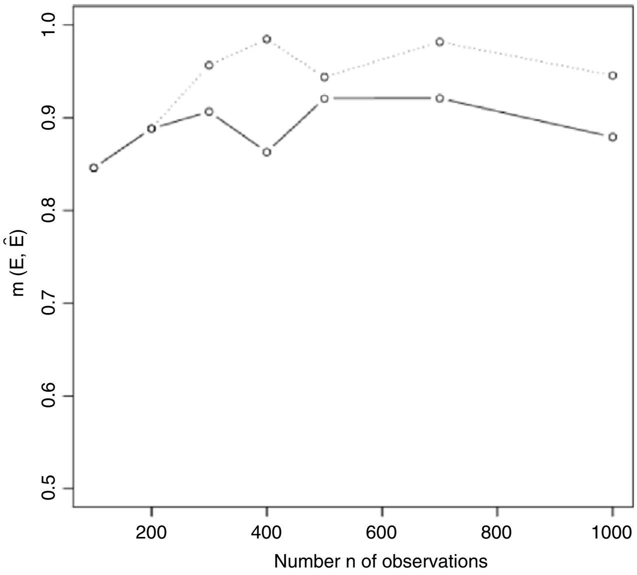

Fig. 3. Quality measure for model (7) with $\theta = 0$ as the number of observations increases (solid line: SIR, dotted line: cluster-based SIR).

Fig. 3 shows the quality of the estimations obtained respectively with SIR and cluster-based SIR for model (7) when $\theta = 0$ and the sample size n takes values 100, 200, 300, 400, 500, 700 and 1000. For each sample size, we have determined the optimal number of clusters. Not surprisingly, we observe that the performances of the two methods globally tend to increase as the number of observations becomes higher. We also observe that in this simulation cluster-based SIR is always better than SIR, except for $n = 100$ and $n = 200$ observations, where the two methods are similar. Indeed for these small sample sizes, cluster-based SIR does not always improve classical SIR because of the small number of observations in some clusters. The cluster-based SIR approach chooses in this case an optimal number of cluster equal to 1, corresponding then to classical SIR. Remark that we have not generated samples with a smaller size than 100, because the use of SIR is then inappropriate. In this case, one should replace the slicing step with pooled-slicing one, see Aragon and Saracco (1997) or Saracco (2001) for details.

4.2.3. Multiple data replications

In this section, we compare SIR and cluster-based SIR on $N = 100$ data replications of model (7). The parameter θ will belong to the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and the number n of observations will be 100, 200, 500 and 1000. For each

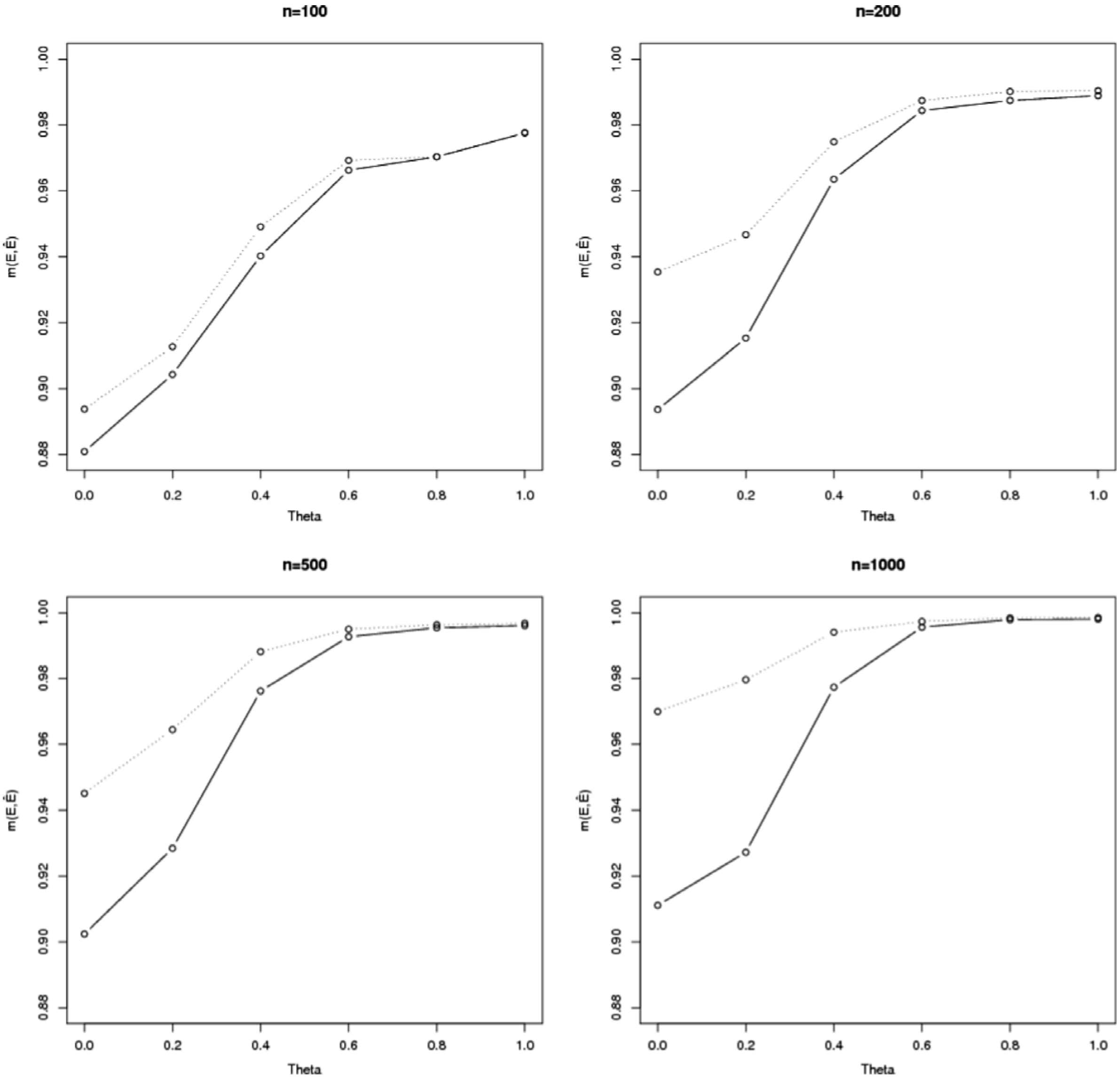

Fig. 4. Plots of the mean of the squared cosines for model (7) with different values of θ and n (solid line: SIR, dotted line: cluster-based SIR).

simulated sample, the e.d.r. direction is estimated with SIR and cluster-based SIR. Cluster-based SIR was implemented with a number of clusters c varying from 1 to 10 (or 20 for $n = 1000$). In this simulation study, the optimal number of clusters was chosen according to criterion (6). We could also use the criterion (5) which gives very similar results but is computationally expensive. The quality measure presented for cluster-based SIR is the one obtained with the optimal number of clusters. Note that the best number may sometimes be equal to 1 (especially for $n = 100$), corresponding then to classical SIR.

Comments on the plots of the means of the squared cosines. Fig. 4 shows the mean of the $N = 100$ squared cosines obtained for each value of θ (from 0 to 1) with SIR and cluster-based SIR estimation methods.

- In each case, both methods give reliable results.
- For the four sample sizes, the performances of both methods increase as θ increases, that is as the data are close to be elliptically distributed ($\theta = 1$). For instance, for cluster-based SIR with $n = 500$, we obtain a mean squared cosine of 0.94 with $\theta = 0$ and 0.99 with $\theta = 1$. For classical SIR, we observe of course the same phenomenon, the mean squared cosine is 0.90 with $\theta = 0$ and 0.99 with $\theta = 1$. This shows that cluster-based SIR is above all helpful in the case of non-ellipticity, which was the aim of the proposed work. Moreover nothing is lost in the case of elliptical distribution. The quality of the results of cluster-based SIR are as good as the ones obtained with SIR.
- As already seen in Fig. 3, we have here a confirmation that the performances of both methods increase as the sample size gets higher. Larger the sample size is, better are the cluster-based SIR results. Indeed, with $n = 1000$ observations, the mean of the squared cosine increases from 0.91 with classical SIR to 0.97 with cluster-based SIR. However with a small

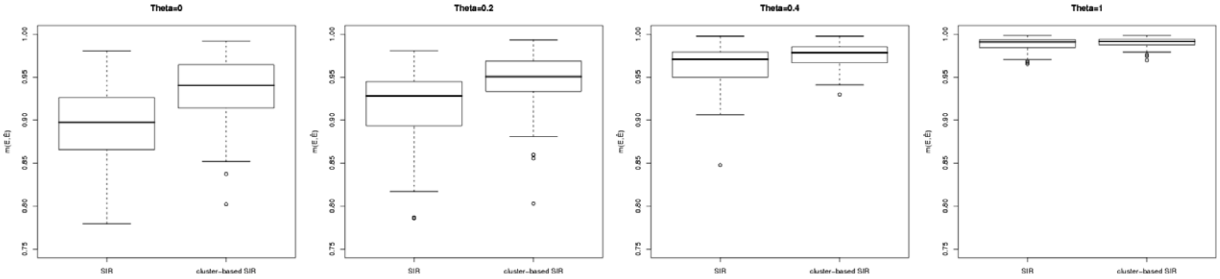

Fig. 5. Boxplots of the squared cosines for model (7) with different values of θ and $n = 200$.

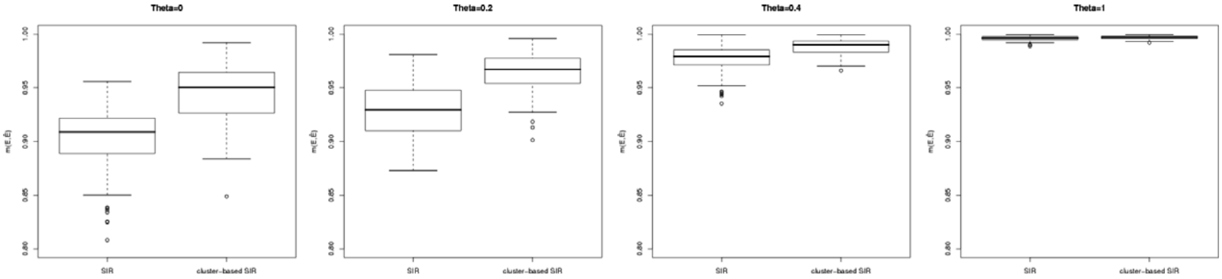

Fig. 6. Boxplots of the squared cosines for model (7) with different values of θ and $n = 500$.

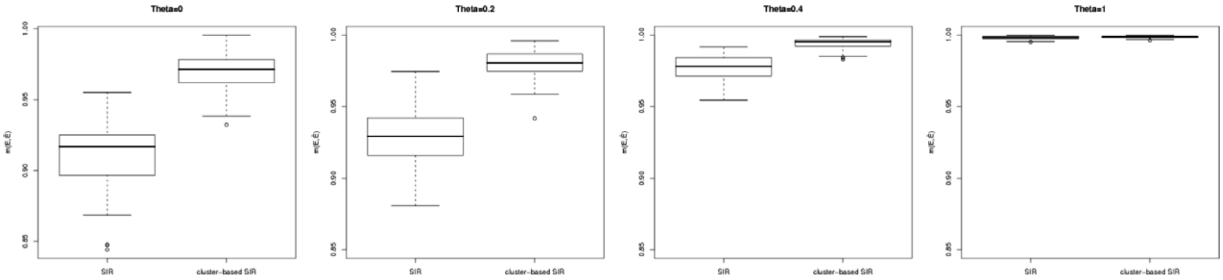

Fig. 7. Boxplots of the squared cosines for model (7) with different values of θ and $n = 1000$.

sample ($n = 100$), the improvement is not so high, the mean squared cosine is 0.88 with SIR and a little more than 0.89 with cluster-based SIR. This comes from the fact that the proposed approach partitions the predictor space. Indeed with large samples, the clustering is better: clusters are better defined and bigger. Therefore the slicing in SIR step occurs on a large number of observations. On the contrary with a small number of observations, the clustering is not so clear and provides sometimes clusters too small for the slicing to be computed. Then cluster-based SIR chooses one cluster, that is the partitioning does not improve the results.

Comments on the boxplots of the squared cosines. Figs. 5–7 show the boxplots of the squared cosines based on the $N = 100$ data replications for $n = 200, 500, 1000$ (with $\theta = 0, 0.2, 0.4, 1$). Both methods give reliable results with a quality measure increasing as the sample size gets higher. For $\theta = 0$ (non-elliptical distribution) and the three sample sizes, cluster-based SIR is better than classical SIR. However in the case of elliptical distribution, the two methods are as effective. The benefits of the use of cluster-based SIR approach is obvious in the case of non-elliptical distribution and big sample size ($n = 1000$).

4.2.4. Comparison with cluster-based OLS

In this section, we compare SIR, cluster-based SIR and cluster-based OLS introduced by Li et al. (2004), on $N = 100$ data replications of model (7) with $\theta = 0$ and $n = 500$. For cluster-based SIR, the optimal number of clusters is chosen according to criterion (5) and for cluster-based OLS, the optimal number of clusters is obtained with a kernel estimator proposed by the authors in their paper. Fig. 8 shows the boxplots of the efficiency measures obtained with SIR, cluster-based SIR and cluster-based OLS. We see that cluster-based SIR is more efficient than SIR and cluster-based OLS. The width of the boxplots of the quality measures is smaller with cluster-based SIR than with the other two methods.

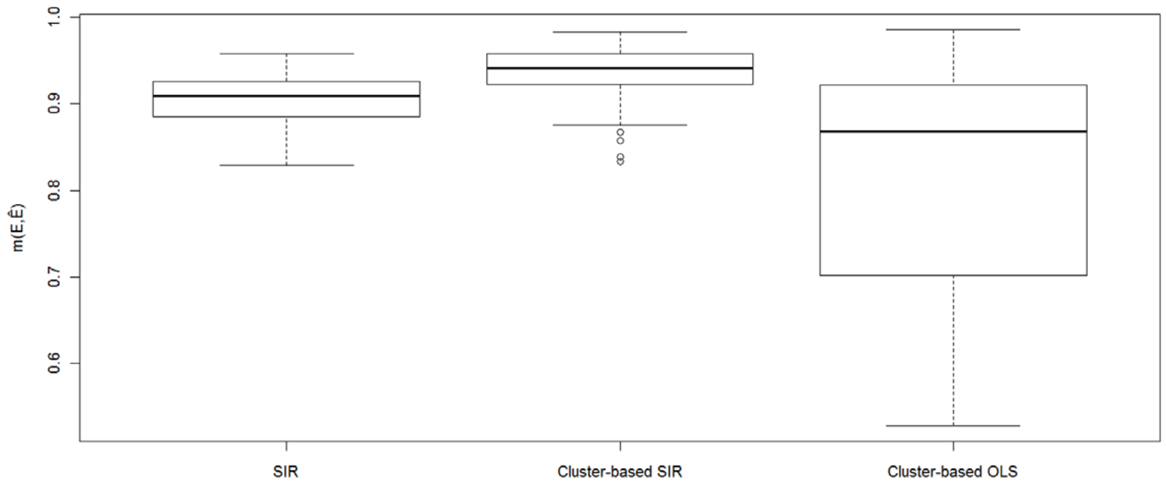

Fig. 8. Boxplots of the efficiency measures obtained with SIR, cluster-based SIR and cluster-based OLS for model (7) where $\theta = 0$ and $n = 500$.

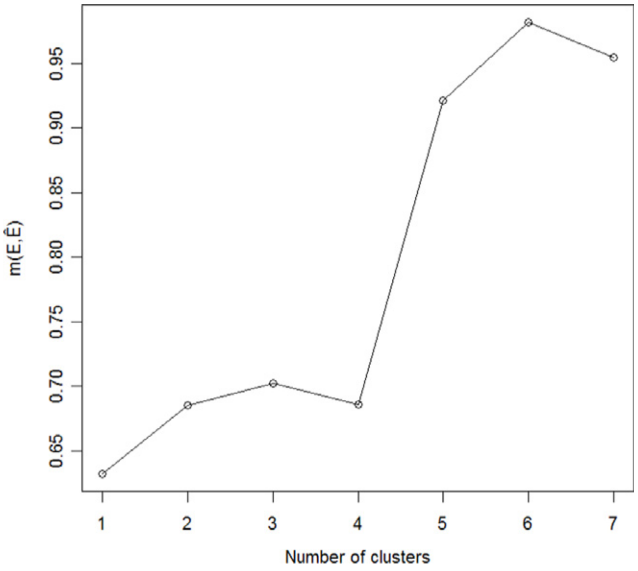

Fig. 9. Efficiency measure of the estimation obtained with SIR ($c = 1$) and cluster-based SIR ($c > 1$) for model (8) where $n = 200$ and $\theta = 0$.

4.3. Two indices model

4.3.1. Simulated model

We consider the following two indices model:

$$y = (\mathbf{x}'\beta_1 + \varepsilon_1)\mathbb{I}_{[\mathbf{x}'\beta_2+\varepsilon_2>0]}, \tag{8}$$

with $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)'$, where $x_j \sim (1-\theta) \times \text{Exp}(1) + \theta \times \mathcal{N}(0, 1)$, $\varepsilon_1 \sim \mathcal{N}(0, 0.1^2)$ and $\varepsilon_2 \sim \mathcal{N}(0, 0.1^2)$. The variables x_j are independent, the error terms ε_1 and ε_1 are independent from each other and from $\mathbf{x}$. We choose as true normalized e.d.r. directions $\beta_1 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, 0)'$ and $\beta_2 = (0, 0, 0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})'$. This model is known as sample selection model. With SIR or cluster-based SIR, the slope parameters β_1 for the observation part of y and β_2 for the selection part (that is the state of y: non-observed (0) or observed) are not individually identifiable. In Chavent, Liquet, and Saracco (2009), a method based on SIR and canonical analysis provides estimates of the directions of β_1 and β_2 for multivariate semiparametric sample selection model, but the price to pay is to add identifiability conditions. Here only the e.d.r. space $E = \text{Span}(\beta_1, \beta_2)$ is identifiable and the quality can only be measured in simulation by $m(E, \hat{E})$ and by $\cos^2(\hat{\hat{\tilde{b}}}_1, \beta_1)$ and $\cos^2(\hat{\hat{\tilde{b}}}_2, \beta_2)$.

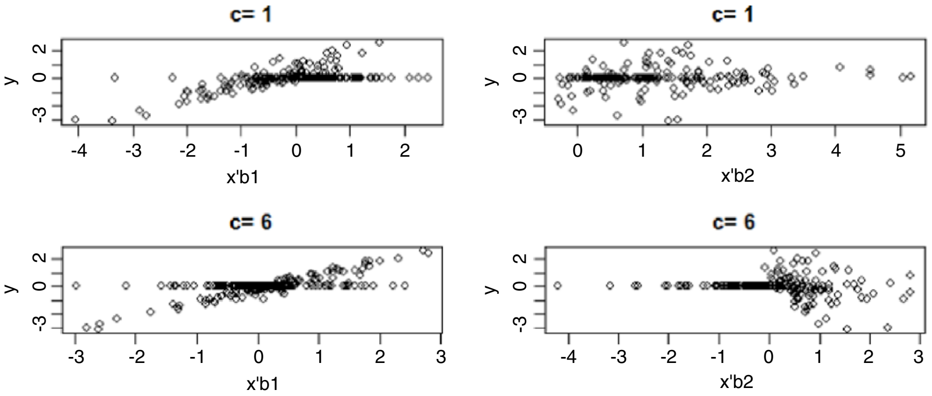

Fig. 10. Scatter plots of $\{(y_i, \mathbf{x}_i'\hat{b}_1), i = 1, \ldots, n\}$ and $\{(y_i, \mathbf{x}_i'\hat{b}_2), i = 1, \ldots, n\}$ for SIR ($c = 1$) at the top and scatter plots of $\{(y_i, \mathbf{x}_i'\hat{\tilde{\tilde{b}}}_1), i = 1, \ldots, n\}$ and $\{(y_i, \mathbf{x}_i'\hat{\tilde{\tilde{b}}}_2), i = 1, \ldots, n\}$ for cluster-based SIR ($\hat{c}^* = 6$) at the bottom.

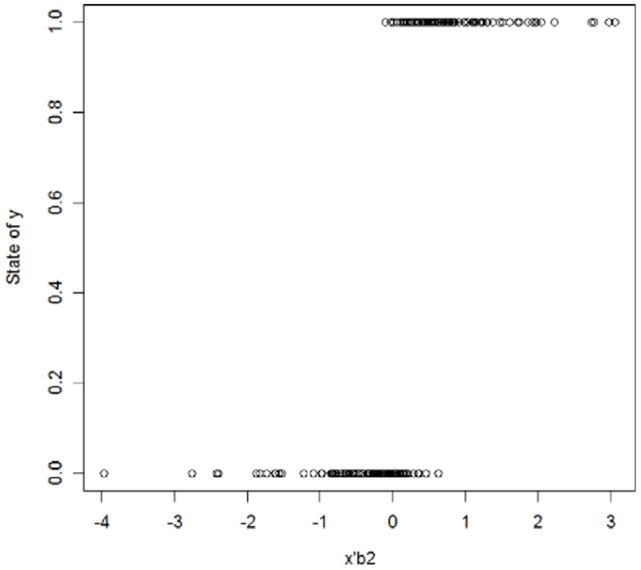

Fig. 11. Scatter plot of the state of y_i (0 when y_i is null and 1 when y_i is nonnull) versus $\mathbf{x}_i'\hat{\tilde{\tilde{b}}}_2, i = 1, \ldots, n$.

4.3.2. Single data replication

We exhibit a sample of size $n = 200$ of model (8) with $\theta = 0$. Fig. 9 shows the evolution of the quality criterion (6) as the number of clusters varies between 1 and 7. The case with one cluster matches with classical SIR. We see that the estimation of the e.d.r. space is improved by clustering the predictor variable $\mathbf{x}$. For cluster-based SIR, the maximum of the quality measure is reached at 6 clusters with an efficiency measure of 0.98 versus 0.63 with classical SIR.

Fig. 10 plots the response variable versus the two estimated e.d.r. directions for SIR and cluster-based SIR (with $\hat{c}^* = 6$). Note that with this simulated sample and the corresponding estimates, the indices $x'\hat{\tilde{\tilde{b}}}_1$ and $x'\beta_1$ (resp. $x'\hat{\tilde{\tilde{b}}}_2$ and $x'\beta_2$) are highly correlated, this is only due to chance (since β_1 and β_2 are not theoretically individually identifiable). Thereby we can observe that the structure is more relevant with cluster-based SIR than with SIR, which emphasizes that the partitioning of the predictor variable is helpful. The bottom left of Fig. 10 shows that the first true e.d.r. direction of model (8) is well recovered. At the bottom right of Fig. 10 we see that the second true e.d.r. direction is also found. Fig. 11 of the scatter plot of the state of y_i (null or nonnull) versus $\mathbf{x}_i'\hat{\tilde{\tilde{b}}}_2, i = 1, \ldots, n$ confirms that the second estimated e.d.r. direction differentiates between null and nonnull values of the response variable.

4.3.3. Multiple data replications

In this section, we compare SIR and cluster-based SIR on $N = 100$ data replications of model (8).

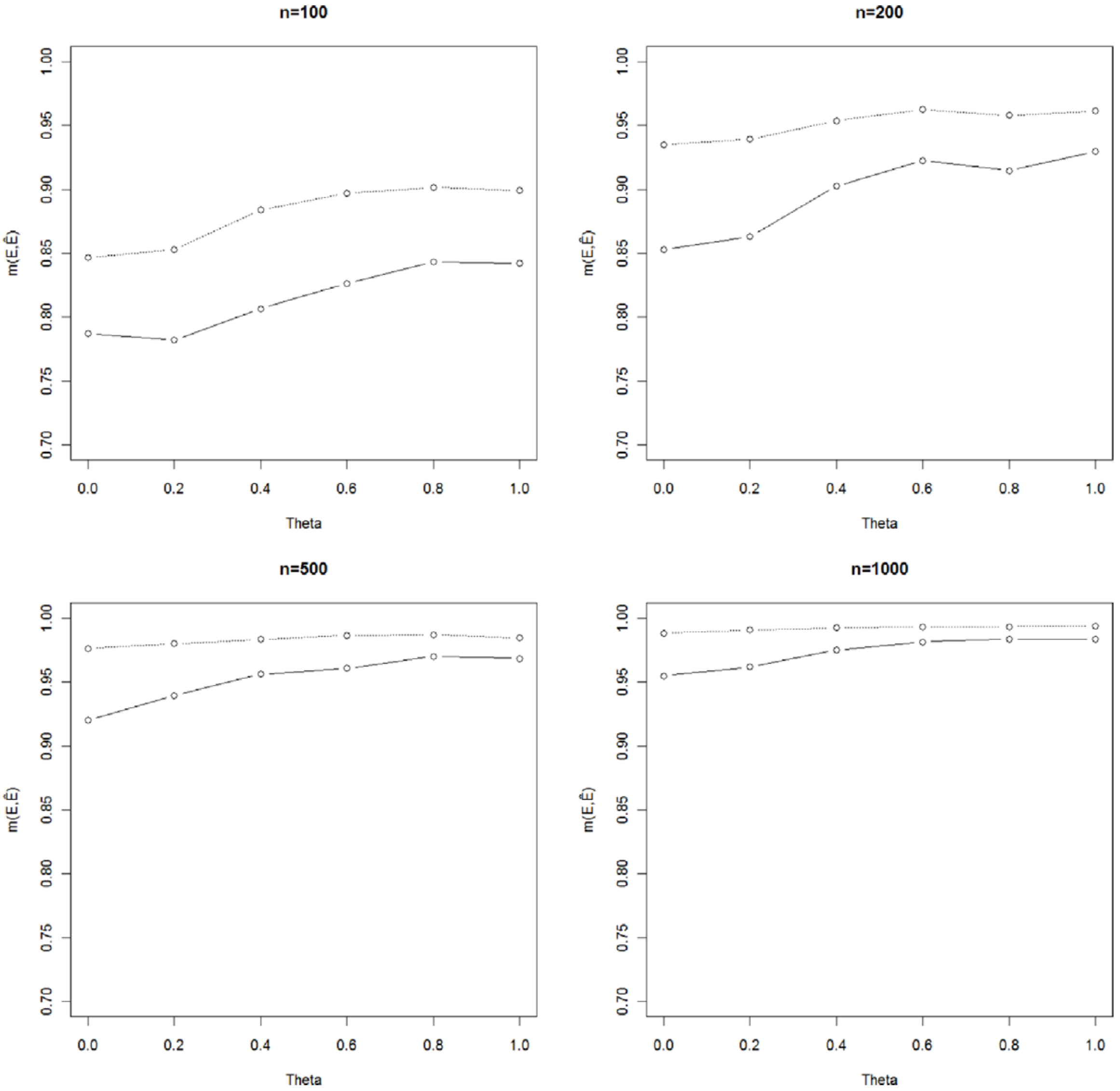

Fig. 12. Plots of the mean of the squared cosines for model (8) with different values of θ and n (solid line: SIR, dotted line: cluster-based SIR).

Comments on the plots of the means of the efficiency measures. Fig. 12 shows the mean of the $N = 100$ quality measures obtained for each value of θ (from 0 to 1) with SIR and cluster-based SIR. Both methods provide very good results and the quality of the estimation increases as the data are close to be elliptical ($\theta = 1$). Cluster-based SIR is always better than classical SIR, especially for sample size of 200 or 500. Indeed with small samples ($n = 100$), the improvement due to clustering is not so high because clusters are sometimes bad defined or too small, preventing then the use of SIR. With large samples ($n = 1000$), the performances of the two methods are similar (with a slight advantage for cluster-based SIR).

Comments on the boxplots of the efficiency measures. Figs. 13–15 show the boxplots of the $N = 100$ efficiency measures obtained for $\theta = 0, 0.2, 0.4, 1$ with SIR and cluster-based SIR estimation methods. The efficiency of both methods is improved when the sample size n increases. Cluster-based SIR always provides better estimations than classical SIR. For both methods, the quality measure increases as θ increases. Compared to SIR, cluster-based SIR is less sensitive to violation of the linearity condition (or elliptical distribution). This is true for any sample size.

4.4. An iterative implementation

For simplicity in the presentation of the iterative implementation we consider the single index model (3). The iterative cluster-based SIR is based on the following implementation. We compute with cluster-based SIR the estimated e.d.r. direction $\hat{b}^{(0)}$. Then we cluster the sample $\{\mathbf{x}_i'\hat{b}^{(0)}, i = 1, \ldots, n\}$, and according to this partition, we compute, with cluster-

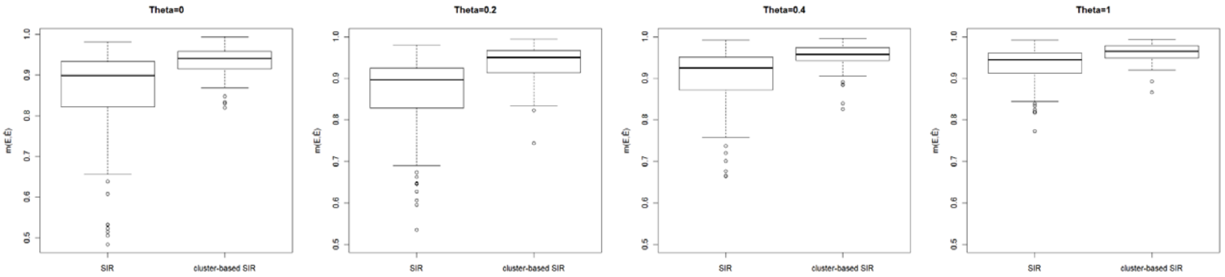

Fig. 13. Boxplots of the efficiency measures for model (8) with different values of θ and $n = 200$.

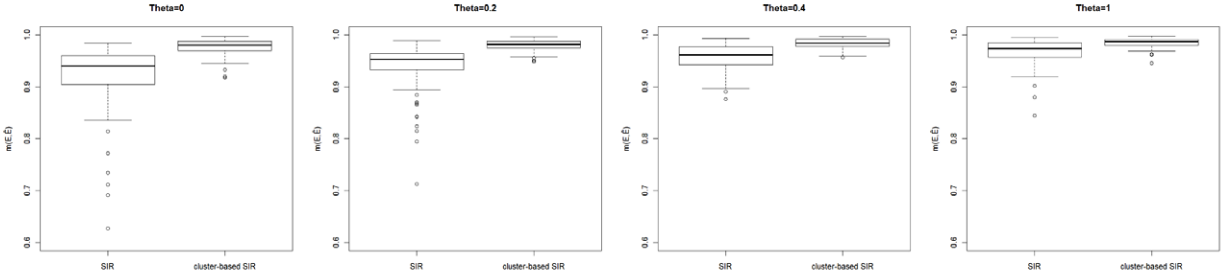

Fig. 14. Boxplots of the efficiency measures for model (8) with different values of θ and $n = 500$.

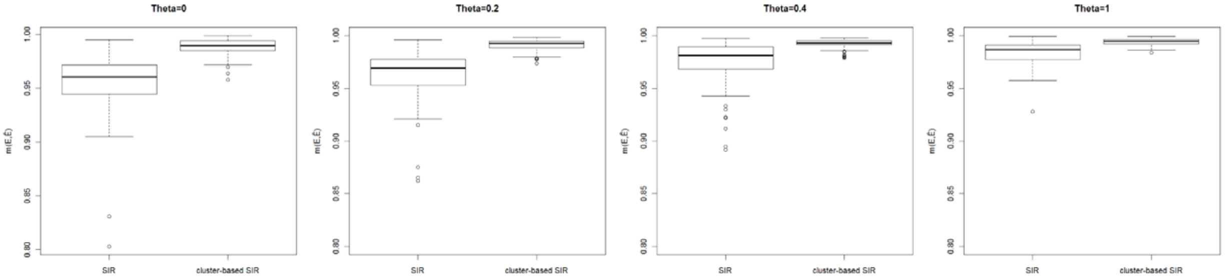

Fig. 15. Boxplots of the efficiency measures for model (8) with different values of θ and $n = 1000$.

based SIR, a new estimation $\hat{b}^{(1)}$ of the e.d.r. direction. As the clustering occurs in a lower-dimensional space (here in $\mathbb{R}$), the partitioning may be better defined and the estimation of the new e.d.r. direction may be improved. We iterate this principle until a stopping criterion, based on the work of Li et al. (2004), is reached: the iteration procedure stops when the correlation between $\mathbf{x}'\hat{b}^{(m)}$ and $\mathbf{x}\hat{b}^{(m+1)}$ reaches a specified threshold (fixed at 0.9 in our simulations). Note that in practice we have often observed that only one iteration is necessary.

In the following simulation results we consider the simulated model (7) with the same parameters given in Section 4.2. We estimate the e.d.r. direction with the cluster-based SIR method and its iterative implementation version on $N = 100$ data replications with $n = 200, 500, 1000$ and $\theta = 0$ (which corresponds to non-elliptical distribution). Fig. 16 shows the boxplots of the quality measure for the different sample sizes. We can clearly observe the improvement of the quality of the estimated e.d.r. basis with the iterative cluster-based SIR approach, particularly for large sample size. The same phenomenon was observed for the other values of θ (0.2,0.4,0.6,0.8,1). We did not report the corresponding results here.

5. Real data application

We consider the data example of horse mussels described in Camden (1989) or Cook and Weisberg (1999). The observations correspond to $n = 201$ horse mussels captured in the Marlborough Sounds at the Northeast of New Zealand's South Island. The response variable y is the muscle mass, the edible portion of the mussel, in grams. The predictor $\mathbf{x}$ is of dimension $p = 4$ and measures numerical characteristics of the shell: length, width, height, each in mm, and mass in grams. In this problem, as the response is discrete, it is slightly transformed as follows $y = y + \epsilon, \epsilon \sim \mathcal{N}(0, 0.01^2)$. Thus we get a continuous variable, which improves the slicing step of SIR. Note that the number of slices used for SIR and cluster-based SIR is $H^{(j)} = 4$ for each cluster $j = 1, \ldots, c$. We used $C^n_{\max} = 8$ and $n_{h,\min} = 6$ in the selection of the number of clusters and in our modified k-means approach. The various studies on this data example have reached to a one-dimensional structure.

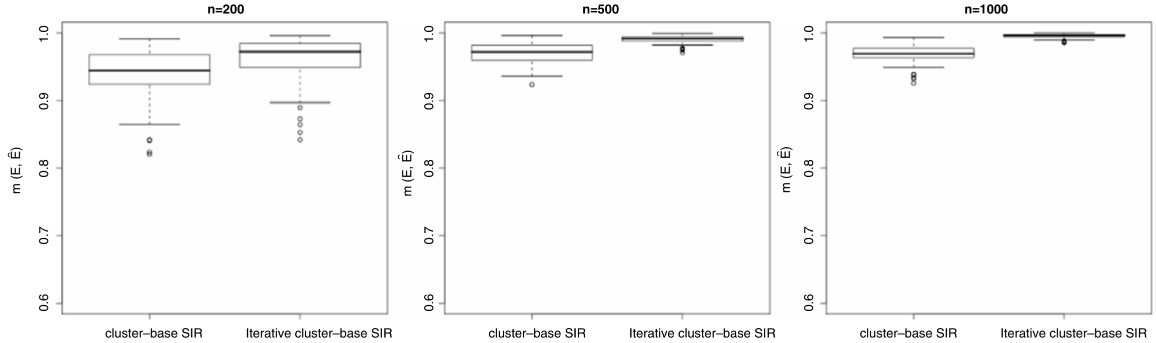

Fig. 16. Boxplots of the efficiency measures for model (7) with cluster-based SIR and iterative cluster-based SIR, for different values of n and $\theta = 0$.

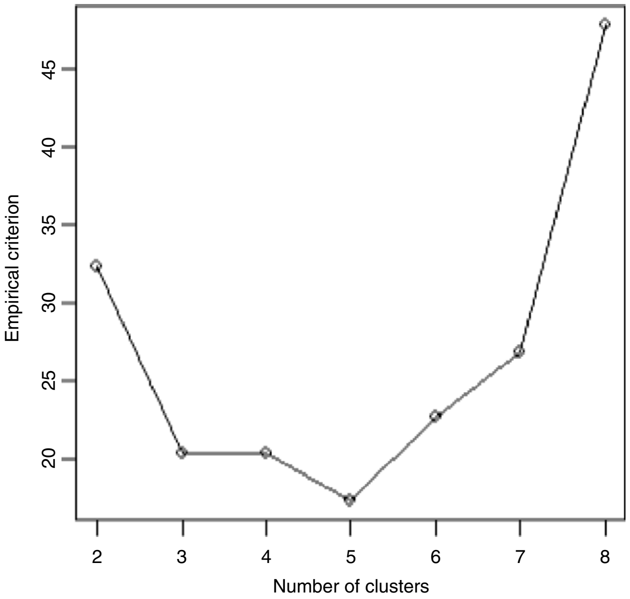

Fig. 17. Empirical criterion defined in (5) with cluster-based SIR ($c > 1$).

Fig. 17 plots empirical criterion (5) and shows that the choice $\hat{c}^* = 5$ of the number of clusters seems to be the best adapted to iterative cluster-based SIR.

For the chosen couple of parameters ($\hat{K} = 1, \hat{c}^* = 5$) we check at the top of Fig. 18 that there is a relevant structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{b}_1), i = 1, \ldots, n\}$. On the contrary for the couple ($\hat{K} = 2, \hat{c}^* = 5$), there is no structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{\hat{\tilde{b}}}_2), i = 1, \ldots, n\}$, see bottom of Fig. 18.

Then we compare the prediction reached on a test sample with SIR and iterative cluster-based SIR using the following algorithm.

Step 1. We split the data into two subsets: $S_J = \{(y_j, \mathbf{x}_j'), j \in J\}$ the training sample containing almost 80% of the total number of observations, and $S_I = \{(y_i, \mathbf{x}_i'), i \in I\}$ the test sample of the remaining observations. Let $n_I = \mathrm{card}(I)$.

Step 2. We use the training sample S_J to compute the estimated e.d.r. direction with SIR, denoted $\hat{b}_{[1]}$, and with iterative cluster-based SIR for $\hat{c}^* = 5$ clusters, denoted $\hat{b}_{[5]}$.

Step 3. We compute the kernel estimate $\hat{y}_{i,[c]}$ of $\mathbb{E}(y|\mathbf{x}_i'\hat{b}_{[c]})$ for $i \in I$ using the sample $\{(y_i, \mathbf{x}_i'\hat{b}_{[c]}), i \in J\}$. We get $\hat{y}_{i,[1]}, i \in I$ for SIR and $\hat{y}_{i,[5]}, i \in I$ for iterative cluster-based SIR.

Step 4. We compute the Mean Absolute Relative Error (MARE) for both SIR and cluster-based SIR estimates as follows:

$$\mathrm{MARE} = \frac{1}{n_I} \sum_{i \in S_I} \left| \frac{y_i - \hat{y}_{i,[c]}}{y_i} \right| .$$

The previous algorithm is repeated $N = 100$ times. Fig. 19 shows the boxplots of the MARE values obtained with SIR and iterative cluster-based SIR. Iterative cluster-based SIR is clearly more efficient than SIR. The range of the boxplot is smaller

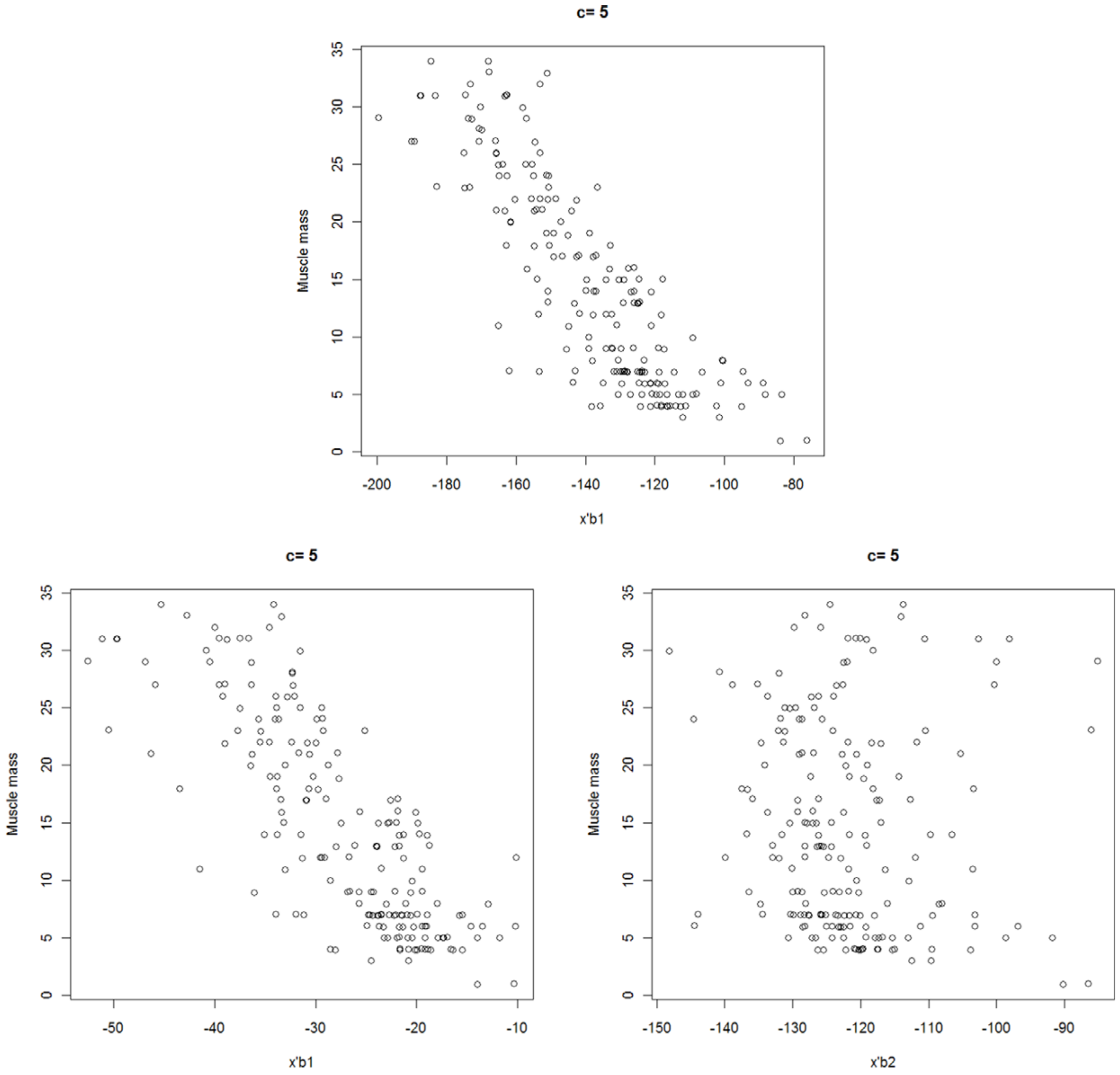

Fig. 18. Scatter plot $\{(y_i, \mathbf{x}_i'\hat{b}_1), i = 1, \ldots, n\}$ at the top and scatter plot $\{(y_i, \mathbf{x}_i'\hat{\tilde{\tilde{b}}}_1, \mathbf{x}_i'\hat{\tilde{\tilde{b}}}_2), i = 1, \ldots, n\}$ at the bottom.

with the use of clustering the predictor space. The median MARE obtained with cluster-based SIR is decreased by half (0.2 versus 0.4 with classical SIR).

6. Concluding remarks

In this article, we have proposed an extension of the well-known dimension reduction method SIR, called cluster-based SIR, which can be used when the crucial linearity condition is not verified. The idea is to partition the predictor space so that the linearity condition approximately holds in each cluster. The optimal number of clusters can be computed from a minimization criterion. Asymptotic properties of the estimator have been obtained. A simulation study has shown the good numerical behaviour of the proposed approach. A real data application has shown the better predictive performance of cluster-based SIR over SIR. Note that cluster-based SIR is less sensitive than SIR to violation of the linearity condition. Thus it opens future prospects for a broader use of SIR. The method has been implemented in R and source codes are available from the authors. As we mentioned in the introduction, the k-means clustering does not always ensure to construct approximately elliptical clusters. However from our simulation results, the use of cluster-based SIR instead of classical SIR globally provides better estimation of the e.d.r. space. The small price to pay is that the cluster-based SIR method is relative much time consuming computationally. Finally our method can be extended to SIR-II and SIR$_\alpha$ (see Gannoun & Saracco, 2003; Li, 1991) or to multivariate SIR approach (see Barreda, Gannoun, & Sarraco, 2007).

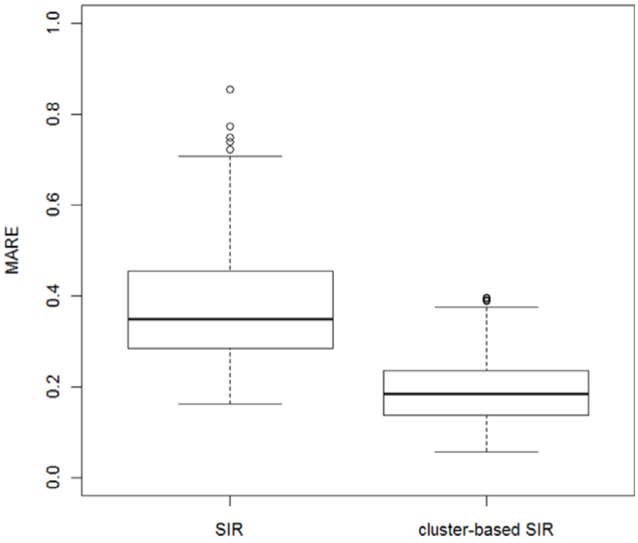

Fig. 19. Boxplots of the MARE values obtained with SIR and cluster-based SIR ($\hat{c}^* = 5$).

Appendix. Proof of Theorem 2

Let $D_1 \otimes D_2$ denote the Kronecker product of the matrices D_1 and D_2 (see for instance Harville (1999) for some useful properties of the Kronecker product). Let $D = [d_1, \ldots, d_q]$ be a $(p \times q)$ matrix, where the d_k's are p-dimensional column vectors. We note vec(D) the pq-dimensional column vector: $\text{vec}(D) = \begin{pmatrix} d_1 \\ \vdots \\ d_q \end{pmatrix}$.

The proof of Theorem 2 is divided into four steps.

Step 1: Asymptotic distribution of $\hat{b}^{(j)}$

Classical asymptotic theory of SIR gives us the following result for each cluster $j = 1, \ldots, c$:

$$\sqrt{n}(\hat{b}^{(j)} - b^{(j)}) \longrightarrow_d W^{(j)} \sim \mathcal{N}(0, \Gamma^{(j)}), \tag{9}$$

where the expression of $\Gamma^{(j)}$ can be found in Saracco (1997).

Step 2: Asymptotic distribution of $\hat{B}$

Under conditions (A2) and (A3), we have:

$$\sqrt{n}(\text{vec}(\hat{B}) - \text{vec}(B)) \longrightarrow_d \text{vec}\begin{pmatrix} W^{(1)} \\ \ldots \\ W^{(j)} \end{pmatrix} \sim \mathcal{N}(0, \Gamma), \tag{10}$$

where:

$$\Gamma = \begin{pmatrix} \Gamma^{(1)} & & 0 \\ & \ddots & \\ 0 & & \Gamma^{(c)} \end{pmatrix}. \tag{11}$$

Step 3: Asymptotic distribution of $\hat{B}\hat{B}'$

We use Delta method. For that, we have to write vec(BB') in terms of vec(B).

We have: $\text{vec}(BB') = \text{vec}(BI_cB') = (B \otimes B)\text{vec}(I_c)$. As $\text{vec}(\text{vec}(BB')) = \text{vec}(BB')$, we can write:

$$\begin{aligned} \text{vec}(BB') &= \text{vec}((B \otimes B)\text{vec}(I_c)) \\ &= (\text{vec}(I_c)' \otimes I_{p^2})\text{vec}(B \otimes B) \\ &= (\text{vec}(I_c)' \otimes I_{p^2})(I_c \otimes K_{cp} \otimes I_p)(\text{vec}(B) \otimes \text{vec}(B)), \end{aligned} \tag{12}$$

where the vec-permutation matrix K_{cp} is equal to $K_{cp} = \sum_{i=1}^{c}\sum_{i=1}^{p}(U_{ij} \otimes U'_{ij})$ with $U_{ij} = e_i u'_j$ and e_i is the ith column of I_c and u_j the jth column of I_p.

Thus we define the following function:

$$\begin{aligned} f:\ & \mathbb{R}^{pc} \to \mathbb{R}^{p^2} \\ & x \mapsto M_1 M_2 (x \otimes x), \end{aligned} \tag{13}$$

with matrices $M_1 = (\text{vec}(I_c)' \otimes I_{p^2})$ and $M_2 = (I_c \otimes K_{cp} \otimes I_p)$.

The Jacobian matrix D associated to f is then equal to:

$$\begin{aligned} D &= \frac{\partial f(x)}{\partial x'} = M_1 M_2 \frac{\partial (x \otimes x)}{\partial x'} = M_1 M_2 \frac{\partial \text{vec}(x \otimes x)}{\partial x'} \\ &= M_1 M_2 (K_{1pc} \otimes I_{pc}) \left[x \otimes \frac{\partial x}{\partial x'} + \frac{\partial x}{\partial x'} \otimes x \right] \\ &= M_1 M_2 (K_{1pc} \otimes I_{pc}) [x \otimes I_{pc} + I_{pc} \otimes x]. \end{aligned} \tag{14}$$

Then applying Delta method with function f defined in (13) and Jacobian matrix D defined in (14), we get:

$$\sqrt{n}(\text{vec}(\hat{B}\hat{B}') - \text{vec}(BB')) \longrightarrow_d V \sim \mathcal{N}(0, \Gamma_V = D\Gamma D'), \tag{15}$$

with matrices Γ and D respectively defined in (11) and (14).

Step 4: Asymptotic distribution of $\hat{b}$

Recall that $\hat{b}$ (resp. b) is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ (resp. λ) of $\hat{B}\hat{B}'$ (resp. BB'). We will note N^+ the Moore–Penrose generalized inverse of the square matrix N.

Since $\hat{B}\hat{B}' = BB' + O_p(n^{-1/2})$ and using (15), according to Lemma 1 of Saracco (1997), we get that:

$$\sqrt{n}(\hat{b} - b) \longrightarrow_d (BB' - \lambda I_p)^+ V b, \tag{16}$$

where:

$$(BB' - \lambda I_p)^+ V b \sim \mathcal{N}(0, \Gamma_U), \tag{17}$$

with:

$$\Gamma_U = [b' \otimes (BB' - \lambda I_p)^+] \Gamma_V [b \otimes (BB' - \lambda I_p)^+]. \tag{18}$$

References

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

Barreda, L., Gannoun, A., & Sarraco, J. (2007). Some extensions of multivariate Sliced Inverse Regression. *Journal of Statistical Computation and Simulation*, *77*, 1–17.

Barrios, M. P., & Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statistics & Probability Letters*, *77*, 247–255.

Brillinger, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In P. J. Bickel, K. A. Doksum, & J. L. Hodges (Eds.), *Festschrift for Erich L. Lehmann in honor of his sixty-fifth birthday* (pp. 97–114). Belmont, Calif: Wadsworth.

Camden, M. (1989). *The data bundle*. Wellington: New Zealand Statistical Association.

Chavent, M., Liquet, B., & Saracco, J. (2009). A semiparametric approach for a multivariate sample selection model. *Statistica Sinica* (in press).

Cheung, Y. M. (2003). *K*-means: A new generalized *k*-means clustering algorithm. *Pattern Recognition Letters*, *24*(15), 2883–2893.

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

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). A Hodder Arnold Publication.

Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association*, *93*(441), 132–140.

Gannoun, A., & Saracco, J. (2003). An asymptotic theory for SIR$_\alpha$ method. *Statistica Sinica*, *13*(2), 297–310.

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

Harville, D. A. (1999). *Matrix algebra from a statistician's perspective*. Springer-Verlag.

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

Li, L., Cook, R. D., & Nachtsheim, C. J. (2004). Cluster-based estimation for sufficient dimension reduction. *Computational Statistics & Data Analysis*, *47*, 175–193.

Liquet, B., & Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR$_\alpha$ method. *Communications in Statistics — Simulation and Computation*, *37*(6), 1198–1218.

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*(3), 489–513.

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

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, *9*, 725–736.