A sliced inverse regression approach for a stratified population

Marie Chavent[1,2], Vanessa Kuentz[1,2], Benoît Liquet[3] and Jérôme Saracco[1,2,4]

[1] Institut de Mathématiques de Bordeaux, UMR CNRS 5251
Université Bordeaux 1 / Université Victor Segalen Bordeaux 2,
351 cours de la libération, 33405 Talence Cedex, France
e-mail: {marie.chavent,vanessa.kuentz,jerome.saracco}@math.u-bordeaux1.fr

[2] INRIA Bordeaux Sud-Ouest, CQFD team, France

[3] INSERM U897, ISPED
Université Victor Segalen Bordeaux 2
146 rue Leo Saignat, 33076 Bordeaux Cedex, France
e-mail: benoit.liquet@isped.u-bordeaux2.fr

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

ABSTRACT

In this paper, we consider a semiparametric single index regression model involving a real dependent variable Y, a p-dimensional quantitative covariable X and a categorical predictor Z which defines a stratification of the population. This model includes a dimension reduction of X via an index $X'\beta$. We propose an approach based on sliced inverse regression in order to estimate the space spanned by the common dimension reduction direction β. We establish $\sqrt{n}$-consistency of the proposed estimator and its asymptotic normality. Simulation study shows good numerical performance of the proposed estimator in homoscedastic and heteroscedastic cases. Extensions to multiple indices models, q-dimensional response variable and/or SIR$_\alpha$-based methods are also discussed. The case of unbalanced subpopulations is treated. Finally a practical method to investigate if there is or not a common direction β is proposed.

Keywords: dimension reduction, Sliced Inverse Regression (SIR), categorical covariate, eigen decomposition.

1 Introduction

Regression analysis studies the relationship between a response variable Y and a covariable X. In parametric regression, the link function is a simple algebraic function of X, and least squares or maximum likelihood methods (among others) can be applied in order to find the best global fit. In nonparametric regression, the class of fitted functions is enlarged in order to obtain greater flexibility via sophisticated smoothing procedures (such kernel or smoothing splines methods). However as the dimension p of the covariable X becomes large, increased difficulties in modeling are often encountered. This is the well-known curse of dimensionality.

In this framework of high dimensional regression, Duan and Li (1991) proposed the following semiparametric dimension reduction single index model:

$$Y = g(X'\beta, \varepsilon), \tag{1}$$

where the univariate response variable Y is associated with the p-dimensional regressor X (with expectation $\mathbf{E}(X) = \mu$ and covariance matrix $\mathbf{V}(X) = \Sigma$) only through the reduced one dimensional variable $X'\beta$. The error term ε is independent of X. The link function g and the vector β are unknown. We are interested in finding the linear subspace spanned by the unknown vector β, called the Effective Dimension Reduction (EDR) space.

Li (1991) introduced Sliced Inverse Regression (SIR) which is a computationally simple and fast method to estimate the EDR space without assuming netiher the functional form of g nor the distribution of ε. This method is based on some properties of the conditional distribution of X given Y and exploits a property of the first inverse moment $\mathbf{E}(X|Y)$; see for instance Duan and Li (1991), Carroll and Li (1992), Hsing and Carroll (1992), Zhu and Ng (1995), Kötter (1996), Saracco (1997, 1999), Aragon and Saracco (1997), Bura and Cook (2001a, 2001b) or Gather et al. (2002) among others.

Since a very large number of high-dimensional data sets do contain quantitative and categorical variables, the introduction of discrete predictors in dimension reduction models appears to be very useful. An extension of model (1) is then to incorporate a categorical predictor Z in addition to the quantitative covariable X. Many covariates (often called factors) are qualitative in the nature such as gender, treatment, type of population, ... Generally, the categorical predictor Z can be viewed as a stratification variable with L "levels" which identifies a number of subpopulations. To introduce this qualitative predictor, we assume that Y and (X, Z) are independent conditionally on $(X'\beta, Z)$. In terms of dimension reduction model, we assume that the relation $Y = f(X'\beta, Z, \varepsilon)$ holds. Thus, when $Z = l$ (for $l = 1, \dots, L$) it follows that

$$Y = g^{(l)}(X'\beta, \varepsilon), \tag{2}$$

where $g^{(l)}(X'\beta, \varepsilon) = f(X'\beta, l, \varepsilon)$. For each subpopulation l, Y is related to the p-dimensional quantitative regressor X only through the index $X'\beta$. The quantitative predictor $X \in \Re^p$ is the covariable with respect to which we will perform dimension reduction, while the discrete predictor Z is an additional categorical covariable that is not included in the reduction of the dimension. This covariable may represent one or more discrete covariables that identify L subpopulations. The categorical variable Z is not assumed to be independent of X. It affects the conditional distribution of X given Z as follows: $\mathbf{E}(X|Z = l) = \mu^{(l)}$ and $\mathbf{V}(X|Z = l) = \Sigma^{(l)}$ for $l = 1, \dots, L$. It also influences the dependency between Y and the index $X'\beta$ via the different link function $g^{(l)}$ associated with each subpopulation l.

As in the standard SIR approaches, a design condition is required for the consistency of the method. In our context, let us assume that X is elliptically symmetric for each subpopulation. Note that we then get the following linear condition for each subpopulation:

$$\text{(LC)} \quad \text{For each } l = 1, \dots, L,\ \mathbf{E}(X'v|X'\beta, Z = l) \ \text{ is linear in } X'\beta \ \text{ for any } v \in \Re^p. \tag{3}$$

In a similar dimension reduction model context with binary regressor, Carroll and Li (1995) presented a new look at treatment comparisons. They considered the covariable Z as the treatment indicator with the following proposed model $Y = g(X'\beta + \theta Z, \varepsilon)$. Estimates of β and θ are obtained without assuming any functional form for g. Their method is based on the use of SIR in order to estimate the direction of β (EDR directions estimated from all subpopulations are combined in order to obtain a final EDR direction), followed by a partial-inverse mean matching method to estimate the treatment effect θ.

When the number L of levels of Z is greater than two, Chiaromonte et al. (2002) considered a similar context to (2) and they presented a *partial* dimension reduction of X, for the regression of Y on (X, Z). They mentioned that this approach does not need to coincide with *marginal* dimension reduction for the regression of Y on X, nor with *conditional* dimension reduction for the regression of Y on X within the subpopulations identified by Z. Assuming the simplifying hypothesis that the predictors' covariance structure is the same across subpopulations:

$$\Sigma^{(l)} = \Sigma^{\star}, \quad l = 1, \ldots, L, \tag{4}$$

Chiaromonte et al. (2002) introduced a corresponding estimation method of the EDR space, based on SIR technique and named Partial SIR. Hereafter, this common covariance assumption will be refered as the "homoscedastic case" in constrast to the more general "heteroscedastic case".

Several authors also worked on dimension reduction approaches in the presence of categorical predictor: see for instance Li et al. (2003a), Yin (2005), Yin and Cook (2005), Liquet and Saracco (2007) or Wang and Yin (2008).

In this paper, we propose a new method to estimate the EDR space which runs smoothly in the general case (that is the heteroscedastic one, including the homoscedastic case). In Section 2, we introduce the population version and the sample version of the corresponding estimator which is obtained from the eigen-decomposition of a symmetric matrix without any pathological problem contrary to the estimator of Liquet and Saracco (2007), see section 5.2 for details. Asymptotics results (consistency and asymptotic normality) are given in Section 3. Possible extensions are described in Section 4: multiple indices model, multivariate response Y, and SIR$_\alpha$-based approach. Section 5 provides a simulation study in order to show the numerical behaviour of the proposed estimator and to compare it with the estimators introduced by Liquet and Saracco (2007). The case of unbalanced subpopulations is also considered. Finally a practical method to investigate if there is or not a common direction β is proposed. Finally, concluding remarks are given in Section 6.

2 The proposed estimator

The idea of the approach is to compute the EDR direction with SIR for each subpopulation and then combine these directions to find the EDR direction of model (2) taking into account the whole population. This approach works in both homoscedastic and heteroscedastic situations. First we describe the population version of the method, and then we give its sample version.

2.1 Population version

Let us consider the L subpopulations defined by the categorical variable Z and let us assume the linearity condition (LC) defined in (3). For each subpopulation l, let us define the covariance matrix of interest, denoted $M_I^{(l)}$, used in the usual SIR approach. For a monotonic transformation $T^{(l)}$ of the dependent variable Y given $Z = l$, we have $M_I^{(l)} = \mathbf{V}(\mathbf{E}(X|T^{(l)}(Y), Z = l))$. In order to easily estimate this matrix $M_I^{(l)}$, Li (1991) proposed a transformation $T^{(l)}$, called a slicing, which categorizes the response variable into a new response with $H^{(l)} > 1$ levels. The support of Y given $Z = l$ is partitioned into $H^{(l)}$ non-overlapping fixed slices $s_1^{(l)}, \ldots, s_h^{(l)}, \ldots, s_{H^{(l)}}^{(l)}$. Then in each subpopulation l, the matrix $M_I^{(l)}$ can be written as $M_I^{(l)} = \sum_{h=1}^{H^{(l)}} p_h^{(l)}(m_h^{(l)} - \mu^{(l)})(m_h^{(l)} - \mu^{(l)})'$, where $p_h^{(l)} = P(Y \in s_h^{(l)}|Z = l)$, $m_h^{(l)} = \mathbf{E}(X^{(l)}|Y \in s_h^{(l)}, Z = l)$ and $\mu^{(l)} = \mathbf{E}(X|Z = l)$. Let $\Sigma^{(l)} = \mathbf{V}(X|Z = l)$. Under the linearity condition (LC), the eigenvector $b^{(l)}$

associated with the largest eigenvalue of the matrix $(\Sigma^{(l)})^{-1}M_I^{(l)}$ is an EDR direction. We can now define the matrix $B = [b^{(1)}, \ldots, b^{(L)}]$ which contains all the EDR directions obtained from all the L subpopulations. We note b the eigenvector associated with the largest eigenvalue BB'. Then Theorem 1 guarantees that this vector is an EDR direction.

Theorem 1 *Assuming the linearity condition (LC) and model* (2)*, the major eigenvector b of the matrix BB' is colinear with β.*

PROOF of Theorem 1. For each subpopulation $l = 1, \ldots, L$, $b^{(l)}$ is colinear with β, i.e. $b^{(l)} = \alpha_l \beta$, where α_l is a nonnull real. As $B = [\alpha_1\beta, \ldots, \alpha_L\beta]$, we obtain $BB' = \sum_{l=1}^{L} \alpha_l^2 \beta\beta' = \|\alpha\|^2 \beta\beta'$, where $\alpha = (\alpha_1, \ldots, \alpha_L)'$ and $\|.\|$ is the norm associated to usual scalar product. Therefore the eigenvector b associated with the strictly positive eigenvalue of BB' is colinear with β. □

2.2 Sample version

We assume that an independent and identically distributed (i.i.d.) sample $\{(X_i, Y_i, Z_i),\ i = 1, \ldots, n\}$ is available from model (2). In order to get an estimator of the matrices $M_I^{(l)}$, the usual idea of the SIR approach is to substitute empirical versions of all the moments for their theoretical counterparts.

Let $\mathcal{S}^{(l)} = \{(Y_i, X_i), i = 1, \ldots, n^{(l)}$ such that $Z_i = l\}$ be the subsample corresponding to the subpopulation l, where $n^{(l)}$ is the size of the subsample $\mathcal{S}^{(l)}$. Let us denote $\mathcal{I}^{(l)}$ the set of indices of the $n^{(l)}$ observations in the subsample $\mathcal{S}^{(l)}$. In each subpopulation, the empirical mean and covariance matrix of the X_i's are respectively given by $\overline{X}^{(l)} = \frac{1}{n^{(l)}} \sum_{i \in \mathcal{I}^{(l)}} X_i$ and $\widehat{\Sigma}^{(l)} = \frac{1}{n^{(l)}} \sum_{i \in \mathcal{I}^{(l)}} (X_i - \overline{X}^{(l)})(X_i - \overline{X}^{(l)})'$. The matrix $M_I^{(l)}$ is estimated by $\widehat{M_I}^{(l)} = \sum_{h=1}^{H^{(l)}} \hat{p}_h^{(l)} (\hat{m}_h^{(l)} - \overline{X}^{(l)})(\hat{m}_h^{(l)} - \overline{X}^{(l)})'$ with $\hat{p}_h^{(l)} = \frac{1}{n^{(l)}} \sum_{i \in \mathcal{I}^{(l)}} \mathbb{I}_{[y_i \in s_h^{(l)}]}$ and $\hat{m}_h^{(l)} = \frac{1}{n^{(l)} \hat{p}_h^{(l)}} \sum_{i \in \mathcal{I}^{(l)}}^{n^{(l)}} X_i \mathbb{I}_{[y_i \in s_h^{(l)}]}$, where the notation $\mathbb{I}$ designates the indicator function. Then the eigenvector $\hat{b}^{(l)}$ associated with the largest eigenvalue of $(\widehat{\Sigma}^{(l)})^{-1} \widehat{M_I}^{(l)}$ is the estimated EDR direction in the subpopulation l. We construct the matrix $\widehat{B} = [\hat{b}^{(1)}, \ldots, \hat{b}^{(L)}]$. The major eigenvector $\hat{b}$ of the matrix $\widehat{B}\widehat{B}'$ is then the EDR estimated direction in model (2).

Remarks. In the usual SIR approach, the practical choice of the slicing function T is discussed in Li (1991), Kötter (2000) and Saracco (2001). The user has to fix the slicing strategy and the number H of slices. In the simulation study in Section 5, for each subpopulation l, the number $H^{(l)}$ of slices is fixed to 10 and each slice contains nearly the same number of observations. Note that in order to avoid the choice of a slicing, kernel-based estimate of SIR 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.

The link functions between the variables of interest and the common estimated index can be first nonparametrically estimated with a kernel method for instance, and subsequently parametrically modelled if necessary, see the simulation study in Section 5 for an illustration.

3 Asymptotic results

In the sequel, the notation $Z_n \rightarrow_d Z$ means that Z_n converges in distribution to Z as $n \rightarrow \infty$. The number of observations in the hth slice for the subpopulation l is denoted $n_h^{(l)}$. The assumptions that are necessary to state our results are gathered below for easy reference.

(A1) Each sample $\mathcal{S}^{(l)}$, $l = 1, \ldots, L$, is a sample of independent observations from the corresponding single index model (2).

(A2) For each subpopulation l, the support of Y is partitioned into a fixed number $H^{(l)}$ of slices such that $p_h^{(l)} \neq 0, h = 1, \ldots, H^{(l)}$.

(A3) For $l = 1, \ldots, L$ and $h = 1, \ldots, H^{(l)}$, $n_h^{(l)} \rightarrow \infty$ (and therefore $n^{(l)} \rightarrow \infty$) as $n \rightarrow \infty$.

We obtain in Theorem 2 the convergence in probability of the estimator $\hat{b}$ and we give its asymptotic distribution in Theorem 3.

Theorem 2 *Under linearity condition (LC) and **(A1)-(A3)**, we have $\hat{b} = b + O_p(n^{-1/2})$.*

PROOF of Theorem 2. For each subpopulation l and under the assumptions of the theorem, we have from SIR theory of Li (1991) that each estimated EDR direction $\hat{b}^{(l)}$ converges to an EDR direction $b^{(l)}$ at root n rate: that is, for each $\mathcal{S}^{(l)}, l = 1, \ldots, L$, $\hat{b}^{(l)} = b^{(l)} + O_p(n^{-1/2})$. Then we get $\widehat{B} = B + O_p(n^{-1/2})$ and $\widehat{B}\widehat{B}' = BB' + O_p(n^{-1/2})$. Therefore the major eigenvector of $\widehat{B}\widehat{B}'$ 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, then the estimated EDR direction $\hat{b}$ converges to an EDR direction at root n rate. □

Theorem 3 *Under linearity condition (LC) and **(A1)-(A3)**, we have $\sqrt{n}(\hat{b} - b) \longrightarrow_d W \sim \mathcal{N}(0, \Gamma_W)$, where the expression of Γ_W is given in (7).*

PROOF of Theorem 3. Let $C_1 \otimes C_2$ denote the Kronecker product of the matrices C_1 and C_2 (see for instance Harville, 1997, for some useful properties of the Kronecker product). Let $C = [c_1, \ldots, c_q]$ be a $(p \times q)$ matrix, where the c_k's are p-dimensional column vectors. We note $\text{vec}(C)$ the pq-dimensional column vector: $\text{vec}(C) = \left(c_1', \ldots, c_q'\right)'$. We will note N^+ the Moore-Penrose generalized inverse of the square matrix N. The proof splits into three steps.

Step 1: Asymptotic distribution of vec($\hat{B}$). Under **(A1)-(A3)**, asymptotic theory of SIR gives us the following result for each subpopulation $l = 1, \ldots, L$: $\sqrt{n}(\hat{b}^{(l)} - b^{(l)}) \longrightarrow_d U^{(l)} \sim \mathcal{N}(0, \Gamma^{(l)})$, where the expression of $\Gamma^{(l)}$ can be found in Saracco (1997) for instance. Then, we have

$$\sqrt{n}(\text{vec}(\hat{B}) - \text{vec}(B)) \longrightarrow_d \text{vec} \begin{pmatrix} U^{(1)} \\ \vdots \\ U^{(L)} \end{pmatrix} \sim \mathcal{N}\left(0, \Gamma\right) \text{ where } \Gamma = \begin{pmatrix} \Gamma^{(1)} & & 0 \\ & \ddots & \\ 0 & & \Gamma^{(L)} \end{pmatrix}. \qquad (5)$$

Step 2: Asymptotic distribution of vec($\hat{B}\hat{B}'$). Remark that

$$\begin{aligned}
\mathrm{vec}(BB') &= \mathrm{vec}(\mathrm{vec}(BB')) = \mathrm{vec}(\mathrm{vec}(BI_LB')) \\
&= \mathrm{vec}((B \otimes B)\mathrm{vec}(I_L)) = (\mathrm{vec}(I_L)' \otimes I_{p^2})\mathrm{vec}(B \otimes B) \\
&= (\mathrm{vec}(I_L)' \otimes I_{p^2})(I_L \otimes K_{L,p} \otimes I_p)(\mathrm{vec}(B) \otimes \mathrm{vec}(B)),
\end{aligned}$$

where the vec-permutation matrix $K_{L,p}$ is equal to $K_{L,p} = \sum_{i=1}^{L}\sum_{j=1}^{p}(E_{ij} \otimes E'_{ij})$ with $E_{ij} = e_{i,L}e'_{j,p}$ and $e_{i,L}$ is the ith column of I_L and $e_{j,p}$ the jth column of I_p. Thus we define the function

$$\begin{aligned}
f: \quad \mathbb{R}^{pL} &\quad\rightarrow\quad \mathbb{R}^{p^2} \\
x &\quad\mapsto\quad A(x \otimes x),
\end{aligned}$$

where $A = (\mathrm{vec}(I_L)' \otimes I_{p^2})(I_L \otimes K_{L,p} \otimes I_p)$. The Jacobian matrix J associated to f is then equal to

$$\begin{aligned}
J = \frac{\partial f(x)}{\partial x'} &= A(K_{1,pL} \otimes I_{pL})[x \otimes \frac{\partial x}{\partial x'} + \frac{\partial x}{\partial x'} \otimes x] \\
&= A(K_{1,pL} \otimes I_{pL})[x \otimes I_{pL} + I_{pL} \otimes x].
\end{aligned}$$

Then using (5) and applying Delta-method, we obtain

$$\sqrt{n}(\mathrm{vec}(\hat{B}\hat{B}') - \mathrm{vec}(BB')) \longrightarrow_d V \sim \mathcal{N}(0, \Gamma_V = J\Gamma J'). \tag{6}$$

Step 3: Asymptotic distribution of $\hat{b}$. The vector $\hat{b}$ (resp. b) is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ (resp. λ) of $\hat{B}\hat{B}'$ (resp. BB'). Since $\hat{B}\hat{B}' = BB' + O_p(n^{-1/2})$ and using (6), according to Lemma 1 of Saracco (1997), we obtain

$$\sqrt{n}(\hat{b} - b) \longrightarrow_d W = (BB' - \lambda I_p)^{+}Vb \sim \mathcal{N}(0, \Gamma_W)$$

with

$$\Gamma_W = [b' \otimes (BB' - \lambda I_p)^{+}]\Gamma_V[b \otimes (BB' - \lambda I_p)^{+}]. \tag{7}$$

□

4 Various possible extensions of the proposed approach

We suggest some possible extensions of the proposed approach. The first one concerns the case of a multiple indices model. In the second one, we suggest to use SIR$_\alpha$-based approach rather than classical SIR. The last extension investigates the case when the dependent variable Y is multivariate.

4.1 Extension to multiple indices model

We can extend the proposed approach to multiple indices model. For each subpopulation $l = 1, \dots, L$, the response variable Y is related to the p-dimensional quantitative regressor X only through the K indices:

$$Y = g^{(l)}(X'\beta_1, \dots, X'\beta_K, \varepsilon) \quad \text{when} \quad Z = l. \tag{8}$$

As in the single index model, the categorical variable Z is not independent of X: the conditional distribution of X given Z is such that $\mathbf{E}(X|Z = l) = \mu^{(l)}$ and $\mathbf{V}(X|Z = l) = \Sigma^{(l)}$ for $l = 1, \dots, L$. Moreover it also affects the dependency between Y and the indices $X'\beta_k$ via specific link functions $g^{(l)}$ for each subpopulation l. In other words, Y and (X, Z) are independent conditionally on $(X'\beta_1, \ldots, X'\beta_K, Z)$.

In this multiple indices model, we search for a basis that spans the EDR space $E = \text{Span}(\beta_1, \ldots, \beta_K)$. As for the single index model, we seek with SIR for a basis of the EDR space for each subpopulation. In order to get theoretical results, we need to adapt the linearity condition and we now assume:

(LC) For each $l = 1, \ldots, L$, $\mathbf{E}(X'v|X'\beta_1, \ldots, X'\beta_K, Z = l)$ is linear in $X'\beta_1, \ldots, X'\beta_K$ for any $v \in \Re^p$.

The eigenvectors $b_1^{(l)}, \ldots, b_K^{(l)}$ associated with the largest K eigenvalues of the matrix $(\Sigma^{(l)})^{-1}M_I^{(l)}$ are EDR directions, where the matrix $M_I^{(l)}$ has been defined in Section 2. Note that the number $H^{(l)}$ of slices for each subpopulation must be greater than K in order to avoid artificial dimension reduction. We define the matrix $B^{(l)} = [b_1^{(l)}, \ldots, b_K^{(l)}]$ containing these EDR directions which form a $\Sigma^{(l)}$-orthogonal basis of E. Then the first K eigenvectors of the matrix $B^{(l)}B^{(l)'}$ form an I_p-orthonormal basis of E. We store them in the $p \times L$ matrix $\tilde{B}^{(l)}$. We can now pool the matrices $\tilde{B}^{(l)}$ in the $p \times KL$ matrix $\mathbb{B}^{(L)} = [\tilde{B}^{(1)}, \ldots, \tilde{B}^{(L)}]$. The K eigenvectors associated with the largest K eigenvalues of $\mathbb{B}^{(L)}\mathbb{B}^{(L)'}$ are denoted by $\tilde{b}_1, \ldots, \tilde{b}_K$.

Theorem 4 *Assuming the linearity condition (LC) and model* (8)*, the vectors* $\tilde{b}_1, \ldots, \tilde{b}_K$ *form an* I_p*-orthogonal basis of the EDR space* E*.*

PROOF of Theorem 4. Since the column vectors of $\tilde{B}^{(l)}$ form an I_p-orthonormal basis of E, we have $\text{Span}(\mathbb{B}^{(L)}) = E$. Then the eigenvectors associated with the K largest eigenvalues of $\mathbb{B}^{(L)}\mathbb{B}^{(L)'}$ form an I_p-orthonormal basis of E. □

Let us now briefly describe the corresponding sample version. We estimate in each subpopulation sample a $\widehat{\Sigma}^{(l)}$-orthogonal basis of the EDR space via SIR: the first K eigenvectors of the matrix $(\widehat{\Sigma}^{(l)})^{-1}\widehat{M_I}^{(l)}$ defined in Section 2. We store them in the matrix $\widehat{B}^{(l)} = [\hat{b}_1^{(l)}, \ldots, \hat{b}_K^{(l)}]$. Then we consider the first K eigenvectors of the matrix $\widehat{B}^{(l)}\widehat{B}^{(l)'}$ which form an I_p-orthogonal basis of the estimated EDR space and we store them in the matrix $\widehat{\tilde{B}}^{(l)}$. Finally the first K eigenvectors of the matrix $\widehat{\mathbb{B}}^{(L)}\widehat{\mathbb{B}}^{(L)'}$, denoted by $\hat{\tilde{b}}_1, \ldots, \hat{\tilde{b}}_K$ provide an I_p-basis of the estimated EDR space, where $\widehat{\mathbb{B}}^{(L)} = [\widehat{\tilde{B}}^{(1)}, \ldots, \widehat{\tilde{B}}^{(L)}]$.

Asymptotics. Under the linearity condition (LC) and the assumptions **(A1)-(A3)**, as for single index model, we can show that the estimated EDR basis converges to an EDR basis at root n rate. Indeed, SIR theory provides $\widehat{B}^{(l)} = B^{(l)} + O_p(n^{-1/2})$ for each subpopulation. Then the eigenvectors associated with the K largest eigenvalues of the matrix $\widehat{B}^{(l)}\widehat{B}^{(l)'}$ converge at same rate to the corresponding ones of $B^{(l)}B^{(l)'}$. Analogously $\widehat{\mathbb{B}}^{(L)} = \mathbb{B}^{(L)} + O_p(n^{-1/2})$ and $\widehat{\mathbb{B}}^{(L)}\widehat{\mathbb{B}}^{(L)\prime} = \mathbb{B}^{(L)}\mathbb{B}^{(L)\prime} + O_p(n^{-1/2})$. Finally $\hat{\tilde{b}}_k = \tilde{b}_k + O_p(n^{-1/2})$, $k = 1, \ldots, K$. Moreover, as for the single index model, using Delta-method, asymptotic results of Tyler (1981) and Saracco (1997), the asymptotic normality of the eigenprojector onto the estimated EDR space can be obtained, as well as the asymptotic distribution of the estimated EDR directions, associated with eigenvalues assumed to be different.

Choice of dimension K. In most applications the number K of indices 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 in each subpopulation (in order to confirm that the true dimension of the whole EDR space is K). Several approaches have been proposed in the literature for SIR. Some approaches are based on hypothesis tests 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 EDR space and its estimate, see for instance Ferré (1998) or Liquet and Saracco (2008) for a graphical bootstrap based approach.

4.2 Extension to SIR$_\alpha$-based approach

The proposed method described in Section 2 is based on SIR, also named SIR-I, which relies on a geometric property of the conditional expectation (first moment) of X given $T(Y)$. Unfortunately, Cook and Weisberg (1991) exhibited a pathological case for SIR-I; they showed that SIR-I is "blind" for "symmetric dependencies". Then, several methods based on higher inverse conditional moment have been proposed in the literature. For instance, Li (1991) introduced the SIR-II approach relying on a property of $\mathbf{V}(X|T(Y))$, and Cook and Weisberg (1991) developed the sliced average variance estimator (SAVE) approach, see also Cook (2000). In order to conjugate information from SIR-I and SIR-II approaches and for increasing the chance of discovering all the EDR directions, Li (1991) proposed the SIR$_\alpha$ method which is a suitable combination of the matrices of interest of these methods. Note that SAVE can be viewed as a particular case of SIR$_\alpha$ when $\alpha = 0.5$.

An additional condition (called the constant variance assumption) is necessary for the consistency of the SIR-II, SAVE and SIR$_\alpha$ methods. In our framework with a categorical predictor and for a multiple indices model, this assumption is written this way:

(CV) $\forall l = 1, \ldots, L,$ the conditional variance $\mathbf{V}(X|X'\beta_1, \ldots, X'\beta_K, Z = l)$ is assumed to be non-random.

Alternatively, we can make the assumption that, for each subpopulation l, X has a multivariate normal distribution which implies that (LC) and (CV) conditions are satisfied.

Let us give now a brief overview of the SIR-II and SIR$_\alpha$ approaches. For the subpopulation l, the SIR-II matrix of interest is defined by $M_{II}^{(l)} = \mathbf{E}\left\{\left(V_T^{(l)} - \mathbf{E}(V_T^{(l)})\right)(\Sigma^{(l)})^{-1}\left(V_T^{(l)} - \mathbf{E}(V_T^{(l)})\right)'\right\}$ where $V_T^{(l)} = \mathbf{V}(X|T(Y), Z = l)$. Under model (8) and (LC) and (CV) assumptions, it can be shown that the eigenvectors associated with the largest K eigenvalues of $(\Sigma^{(l)})^{-1}M_{II}^{(l)}$ are some EDR directions. In SIR$_\alpha$ approach, we consider, for the subpopulation l, the eigen-decomposition of the matrix $(\Sigma^{(l)})^{-1}M_\alpha^{(l)}$ where $\alpha \in [0, 1]$ and $M_\alpha^{(l)} = (1-\alpha)M_I^{(l)}\Sigma^{-1}M_I^{(l)} + \alpha M_{II}^{(l)}$. It can also be proved that the eigenvectors associated with the largest K eigenvalues of $(\Sigma^{(l)})^{-1}M_\alpha^{(l)}$ are some EDR directions, see Li (1991). Let us remark that, when $\alpha = 0$ (resp. $\alpha = 1$), SIR$_\alpha$ is equivalent to SIR-I (resp. SIR-II).

For each subpopulation such that $Z = l$, when transformation T is a slicing which partitions the support of Y is partitioned into $H^{(l)} > K$ non-overlapping slices $s_h^{(l)}$, the matrix $M_{II}^{(l)}$ is now written as $M_{II}^{(l)} = \sum_{h=1}^{H^{(l)}} p_h^{(l)}\left(V_h^{(l)} - \overline{V}^{(l)}\right)(\Sigma^{(l)})^{-1}\left(V_h^{(l)} - \overline{V}^{(l)}\right)$, where $V_h^{(l)} = \mathbf{V}(X|Y \in s_h^{(l)}, Z = l)$ and $\overline{V}^{(l)} = \sum\limits_{h=1}^{H^{(l)}} p_h^{(l)}V_h^{(l)}$. It is straightforward to estimate the matrices $M_{II}^{(l)}$ and $M_\alpha^{(l)}$ by substituting empirical versions of the moments for their theoretical counterparts, and therefore to obtain the estimation of the EDR directions. Each estimated EDR direction converges to an EDR direction at root n rate when the corresponding eigenvalues are assumed to be distinct, see for instance Li (1991) or Saracco (2001). Asymptotic normality of the SIR$_\alpha$ estimates has been studied by Gannoun and Saracco (2003a).

The practical choice of α can be based on test approach (see Saracco, 2001) or on cross-validation criterion (see Gannoun and Saracco, 2003b). A graphical bootstrap based approach has also been developped by Liquet and Saracco (2008) in order to select simultaneously the couple (α, K).

Hence, an extension of the proposed approach is to replace the SIR-I estimators of the EDR directions by the corresponding SIR_α ones in the population and sample versions. Then, the corresponding version will be insensitive to symmetric dependence in the model for a good choice of α.

4.3 Extension to a multivariate dependent variable Y

Several authors (see for instance Aragon, 1997, Hsing, 1999, Li et al., 2003b, Lue, 2009) extented the univariate model (1) to a multivariate response variable: Y is assumed to be q-dimensional with $q > 1$, the corresponding link function is then $\Re^q$-valued. A few methods based on SIR-I approach have been developed in this multivariate context. Saracco (2005) and Barreda et al. (2007) focused on some extensions of the existing multivariate SIR approaches relying on SIR_α method.

In our framework, we consider a multivariate extension of model (8) in which $Y \in \Re^q$. Let Y^j be the jth component of Y. We can now introduce the multivariate semiparametric regression model: for $l = 1, \ldots, L$,

$$Y = \begin{cases} Y^1 = g_1^{(l)}(X'\beta_1, \ldots, X'\beta_K, \varepsilon_1^{(l)}) & \text{when } Z = l, \\ \vdots & \\ Y^q = g_q^{(l)}(X'\beta_1, \ldots, X'\beta_K, \varepsilon_q^{(l)}) & \text{when } Z = l, \end{cases} \tag{9}$$

where the error terms ε_j are assumed independent of X and the link functions g_j's are unknown real-valued functions. As in model (1), only the EDR space is identifiable. Straightforwardly, we can extend our proposed method to this multivariate framework. The idea is to use a multivariate SIR method rather than SIR-I in order to get an EDR basis for each subpopulation. As in Liquet and Saracco (2007), we suggest to use the PMS_α approach which is a Pooled Marginal Slicing method based on SIR_α; see Saracco (2005) for details.

5 Simulation studies

In this section we perform with R simulation studies to illustrate the numerical behaviour of the new proposed approach and to compare it to the homoscedastic and heteroscedastic approaches developped by Liquet and Saracco (2007). We also compare these approaches to a naive SIR-I approach which consists in estimating the EDR directions without using the information of the categorical variable Z and perform a simple SIR-I method on the global sample. All the source codes are available from the authors by E-mail.

First we introduce the efficiency measure which will be used to compare the performances of these methods. Then we briefly recall the homoscedastic and heteroscedastic approaches of Liquet and Saracco (2007). In the following, we consider two single index models ($K = 1$): the first one with a categorical variable with $L = 2$ levels, and the other one with a categorical variable with 3 levels ($L = 3$). Then we present some results for a two indices model ($K = 2$) with $L = 2$. The case of unbalanced subpopulations is also treated. Finally a practical method to investigate if there is or not a common direction β is proposed.

5.1 Efficiency measure

Let $\hat{b}_1, \ldots, \hat{b}_K$ be the K estimated EDR directions. We note $\hat{B} = [\hat{b}_1, \ldots, \hat{b}_K]$ and $\hat{E} = \text{Span}(\hat{B})$ the linear subspace spanned by the $\hat{b}_k$'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. $P_{\hat{E}}$) be the I_p-orthogonal projector onto E (resp. $\hat{E}$) defined as follows: $P_E = B(B'B)^{-1}B'$ and $P_{\hat{E}} = \hat{B}(\hat{B}'\hat{B})^{-1}\hat{B}'$. The quality of the estimate $\hat{E}$ of E is measured by:

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

This measure belongs to $[0, 1]$ with $m(E, \hat{E}) = 0$ if $\hat{E}$ and E are I_p-orthogonal and $m(E, \hat{E}) = 1$ if $\hat{E} = E$. Then 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 $\hat{b}$.

5.2 The estimator of Liquet and Saracco (2007)

Let us briefly recall the principle of the estimator based on PMS$_\alpha$ approach introduced by Liquet and Saracco (2007). For sake of simplicity, we only consider the case $K = 1$. The idea is to pool the "marginal" SIR$_\alpha$ matrices of interest obtained for each component Y^j of Y and for each level l of Z. Then, the population version of the pooled matrix of interest is defined as follows:

$$\mathcal{M}^P_{q,L} = \sum_{j=1}^{q} \tilde{w}_q^{(j)} \left\{ \sum_{l=1}^{L} w_L^{(l)} \left(\Sigma^{(l)}\right)^{-1} M^{(j,l)}_{\alpha^{(j,l)}} \right\}, \tag{10}$$

where the matrices $M^{(j,l)}_{\alpha^{(j,l)}}$ are the M_α matrices corresponding to the subpopulation l for the component Y^j of Y, that is the matrix of interest of SIR$_\alpha$ defined for the pairs (X, Y^j) given $Z = l$. The weights $\{w_L^{(l)},\ l = 1, \ldots, L\}$ are the probability of the events $Z = l$, and the weights $\{\tilde{w}_q^{(j)},\ j = 1, \ldots, q\}$ are some positive weights such that $\sum_{j=1}^q \tilde{w}_q^{(j)} = 1$. The parameter α of each M_α matrix can be individually adapted and is denoted by $\alpha^{(j,l)}$.

In the homoscedastic case (4), the expression in (10) can be written this way $(\Sigma^\star)^{-1} M^P_{q,L}$, where $M^P_{q,L} = \sum_{j=1}^q \tilde{w}_q^{(j)} \left\{ \sum_{l=1}^L w_L^{(l)} M^{(j,l)}_{\alpha^{(j,l)}} \right\}$. Under some classical assumptions in dimension reduction framework, Liquet and Saracco (2007) showed that the eigenvector associated with the largest eigenvalue of $(\Sigma^\star)^{-1} M^P_{q,L}$ is an EDR direction. The sample version can be easily obtained as for usual SIR approach by substituting empirical versions of all the moments for their theoretical counterparts. The estimated EDR direction $\hat{b}$ (principal eigenvector of $(\hat{\Sigma}^\star)^{-1} \widehat{M^P_{q,L}}$) converges to the EDR direction b (principal eigenvector of $(\Sigma^\star)^{-1} M^P_{q,L}$) at root n rate. In the sequel of the simulation study, the values of the $\alpha^{(j,l)}$'s are fixed to zero, the corresponding method will be called "homo".

In the heteroscedastic case, we consider the eigenvalue decomposition of the matrix $\mathcal{M}^P_{q,L}$ defined in (10) under the design condition. Contrary to the homoscedastic case, this matrix has no reason to be symmetric (with respect to a specific inner product) or positive definite. However, for each matrix $\left(\Sigma^{(l)}\right)^{-1} M^{(j,l)}_{\alpha^{(j,l)}}$, the eigenvector associated with the largest eigenvalue is an EDR direction. Then, since the matrix $\mathcal{M}^P_{q,L}$ is a convex combination of these matrices, there exists an eigenvector b which is an EDR direction associated with a real positive eigenvalue λ. From an algebraic point of view, there is no guarantee that this corresponding eigenvalue is the largest one (in descending order of modulus, since it is possible to obtain complex eigen-elements). Geometrically speaking, one can find pathological cases in which the largest eigenvalue is not associated with an EDR direction, see for instance the example given in the Appendix of Liquet and Saracco (2007). Unfortunately, to the best of our knowledge, there is no characterization of these pathological cases, nor necessary conditions allowing us to avoid these cases. The matrix $\mathcal{M}^P_{q,L}$ is estimated by substituting empirical versions of the moments for their theoretical counterparts. As in the homoscedastic case, the root n consistency of the estimated EDR direction to the corresponding EDR direction has been established. In the sequel of the simulation study, this method is named "hetero" with the values of the $\alpha^{(j,l)}$'s fixed at zero.

5.3 Single index model ($K = 1$)

In the simulation study, we first consider a single index model with categorical predictor with $L = 2$ levels, then we will consider an underlying model which contains a categorical predictor with $L = 3$ levels. For each simulated sample, we estimate the EDR direction with the four methods specified below: the proposed new method (named "new" hereafter), the "homo" and "hetero" methods of Liquet and Saracco (2007), and the naive SIR-I method (named "SIR" hereafter). The quality measure of each estimated direction is then evaluated. In the following, we only present results obtained in the heteroscedastic setup which is the most important one in practice. Note that simulations have been made in the homoscedastic setup and have provided good values of the quality measure for the three methods "homo", "hetero" and "new". None method appears to be uniformly better than the others, all the methods provide the same high level of performance. This is not surprising since the homoscedastic setup is included in the heteroscedastic one. In the heteroscedastic case, only the "new" and "hetero" methods should provide the best performance, the "homo" (which relies on an homoscedastic assumption) and "SIR" (which does not use the information from the categorical covariate) approaches should have difficulties to retrieve the EDR direction.

5.3.1 Categorical predictor with $L = 2$ levels

In this simulation, we generate simulated data from the following semiparametric regression model:

$$\begin{cases} Y = (\frac{1}{8}\beta' X)^3 + \epsilon_1 & \text{for} \quad Z = 1, \\ Y = -(\beta' X)/2 + \epsilon_2 & \text{for} \quad Z = 2, \end{cases} \tag{11}$$

where $X|Z = l$ (for $l = 1, 2$) follows a 5-dimensional normal distribution with means $\mu^{(l)} = \mathbf{0}_5$ and randomly generated covariance matrices $\Sigma^{(l)}$. In this paper, the way of generating the matrices $\Sigma^{(l)}$ is the following. A matrix $\Lambda^{(l)}$ is randomly filled with numbers belonging to $[-2, 2]$. Then $\Sigma^{(l)} = \Lambda^{(l)}\Lambda^{(l)\prime} + 0.5 I_p$ in order to avoid problem of inversion of $\Sigma^{(l)}$. Each random error term ϵ_l is independent of X given $Z = l$ and $\epsilon_l \sim \mathcal{N}(0, 0.7^2)$ for $l = 1, 2$. We take $\beta = (1, 2, -1, -2, 0)'$.

An illustrated example. A sample of size $n = 200$ was generated from model (11) with the same number of observations in each subsample defined by the levels of Z: $n^{(1)} = n^{(2)} = 100$. The EDR direction was estimated with the different methods ("SIR", "hetero", "homo" and "new"). As "SIR" does not give good estimation, we only give the eigenvalues for the three other methods: $\lambda_{\text{homo}} = (1.70, 0.83, 0.29, 0.08, 0.01)$, $\lambda_{\text{hetero}} = (0.88, 0.06, 0.06, 0.01, 0.01)$ and $\lambda_{\text{new}} = (0.99, 0.01, 0, 0, 0)$. Clearly, there is a visible jump between the first and the second eigenvalues, then we retain only one EDR direction. The two methods, "hetero" and "new", give excellent estimations with $m(E, \hat{E}) \simeq 0.99$. The "homo" method provides $m(E, \hat{E}) \simeq 0.84$. Note that the naive approach “SIR” fails : $m(E,\hat{E}) \simeq 0.40$. The corresponding estimated EDR directions for the “hetero”, “homo” and “new” approaches, are respectively $\hat{b}_{\text{homo}} = (0.43, -0.47, 0.48, 0.22, 0.56)$, $\hat{b}_{\text{hetero}} = (-0.34, -0.64, 0.39, 0.56, 0.08)$, and $\hat{b}_{\text{new}} = (-0.35, -0.64, 0.39, 0.56, 0.08)$. The plots, for each subpopulation l, of the response variable Y versus the true (resp. estimated) common index $X'\beta$ (resp. $X'\hat{b}_{\text{new}}$) are represented on the left (resp. right) handside of Figure 1. In the right handside of this figure, we exhibit the smoothing spline estimates of the link functions $g^{(l)}$ between the variables of interest Y and the common estimated index when $Z = l$. On the left hand side, the true link function $g^{(l)}$ has been plotted.

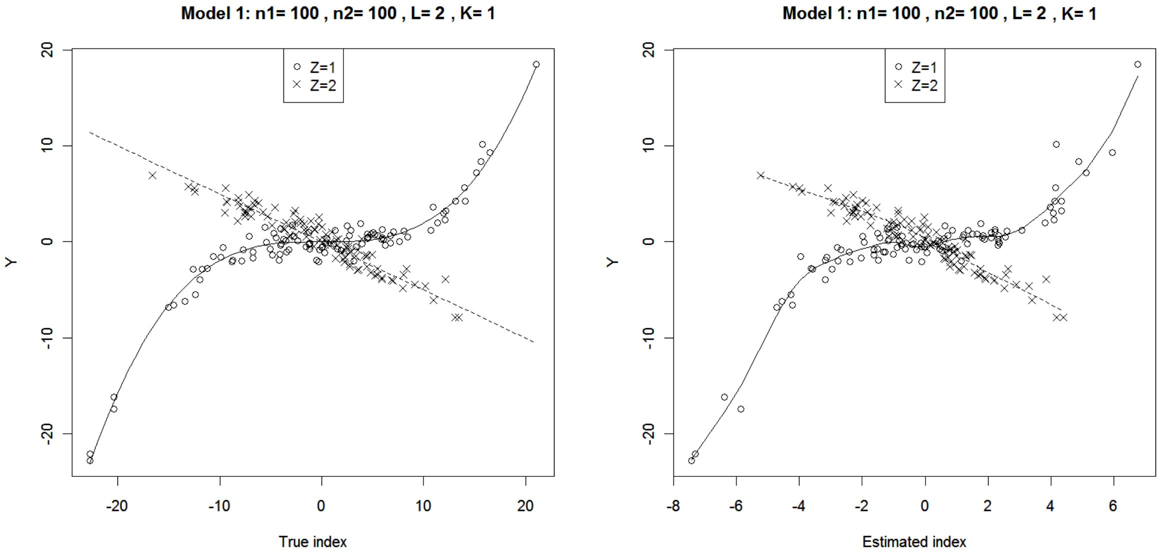

Figure 1: On the left: plots of the true index versus $(X'\beta)$ Y. On the right: plots of the estimated index $(X'\hat{b})$ versus Y with plots of the estimated link functions for $Z = 1$ (solid line) and for $Z = 2$ (dotted line).

Results of a simulation study. From model (11) , $N = 500$ samples of size $n = 200$ (with $n^{(1)} = n^{(2)} = 100$) were generated. For each simulated sample, the EDR direction was estimated with the four methods: “homo”, “hetero”, “new” and “SIR”. Then, in order to compare the different estimates, we calculated, for each estimation, the corresponding efficiency measures. We represent on the left handside of Figure 2 the boxplots of the $N = 500$ squared cosines obtained with the four methods. Moreover we observe that the naive “SIR” approach can not recover the EDR direction. The “new” method gives as best results as the “hetero” method, but it has the advantage not to suffer from pathological cases. On the contrary with the “hetero” method, there is no guarantee that the eigenvalue corresponding to the EDR direction is the largest one (in descending order of modulus). The “homo” method does not give good measures of quality, which highlights the fact that taking the heteroscedastic setup into account improves the estimation of the EDR direction. On the right handside of Figure 2 we plot the squared cosines obtained with “hetero” versus the ones provided by “new”. This graphic shows that the “new method” seems to be slightly better than the “hetero” one.

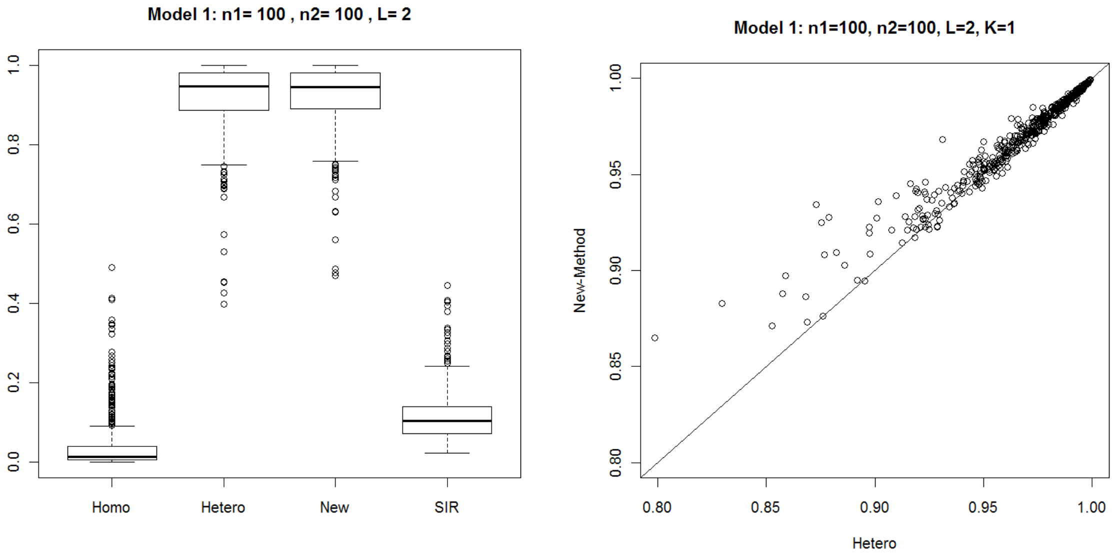

Figure 2: On the left: boxplots of the squared cosines for the four different methods for $n^{(1)} = n^{(2)} = 100$. On the right: comparison of the squared cosines for the "hetero" and "new" methods.

5.3.2 Categorical predictor with $L = 3$ levels

We generate here simulated data from the following semiparametric regression model with a categorical covariate having $L = 3$ levels:

$$\begin{cases} Y = \beta' X + \epsilon_1 & \text{for} \quad Z = 1, \\ Y = (\beta' X/5)^3 + 0.1\epsilon_2 & \text{for} \quad Z = 2, \\ Y = -(\beta' X)/2 + \epsilon_3 & \text{for} \quad Z = 3, \end{cases} \tag{12}$$

where $X|Z = l$ (for $l = 1, \dots, 3$) follows a 5-dimensional normal distribution with mean $\mu^{(l)} = \mathbf{0}_5$ and randomly generated covariance matrices $\Sigma^{(l)}$ as in the previous subsection. Each random error term $\epsilon_l, l = 1, 2, 3$ is independent of X given $Z = l$ and $\epsilon_l \sim \mathcal{N}(0, 0.9^2)$ for $l = 1, 2$. We take $\beta = (1, 1, -1, -1, 0)'$.

An illustrated example. A sample of size $n = 450$ was generated from model (12) with the same number of observations in each subsample defined by Z ($n^{(1)} = n^{(2)} = n^{(3)} = 150$). The EDR direction was estimated with the four different methods ("homo", "hetero", "new" and naive "SIR"). Not surprisingly, both methods, "hetero" and "new", give excellent estimations with $m(E, \hat{E}) \simeq 0.99$. The "homo" method gives $m(E, \hat{E}) \simeq 0.80$. Note that the naive "SIR" approach fails with $m(E, \hat{E}) \simeq 0.56$. The plots, for each subpopulation l, of the response variable Y versus the estimated index $X'\hat{b}_{\text{new}}$ are represented on the right handside of Figure 3. In this graphic, we add the smoothing spline estimates of the link functions $g^{(l)}$ between the variable of interest Y and the common estimated index when $Z = l$. On the left handside of Figure 3, we represent the scatterplot of the $(X_i'\beta, Y_i)$'s and the true link functions $g^{(l)}$ for $l = 1, 2, 3$.

Results of a simulation study We represent on the left of Figure 4 the boxplots of the $N = 500$ squared cosines obtained with the four methods with sample sizes $n^{(1)} = n^{(2)} = n^{(3)} = 150$. For this simulation study, one can observe that neither the naive "SIR" approach nor the "homo" method recover the EDR direction. The "new" and "hetero" methods achieve the best performances.

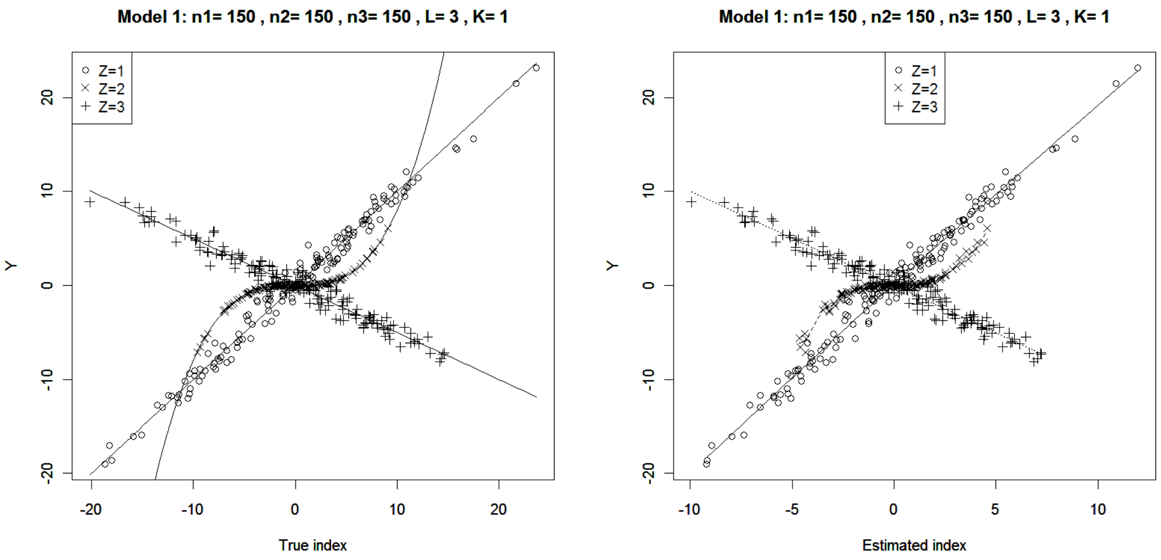

Figure 3: On the left: plot of the true index $(X'\beta)$ versus Y. On the right: plots of the estimated index $(X'\hat{b})$ versus Y with plots of the estimated link functions for $Z = l$ $(l = 1, 2, 3)$.

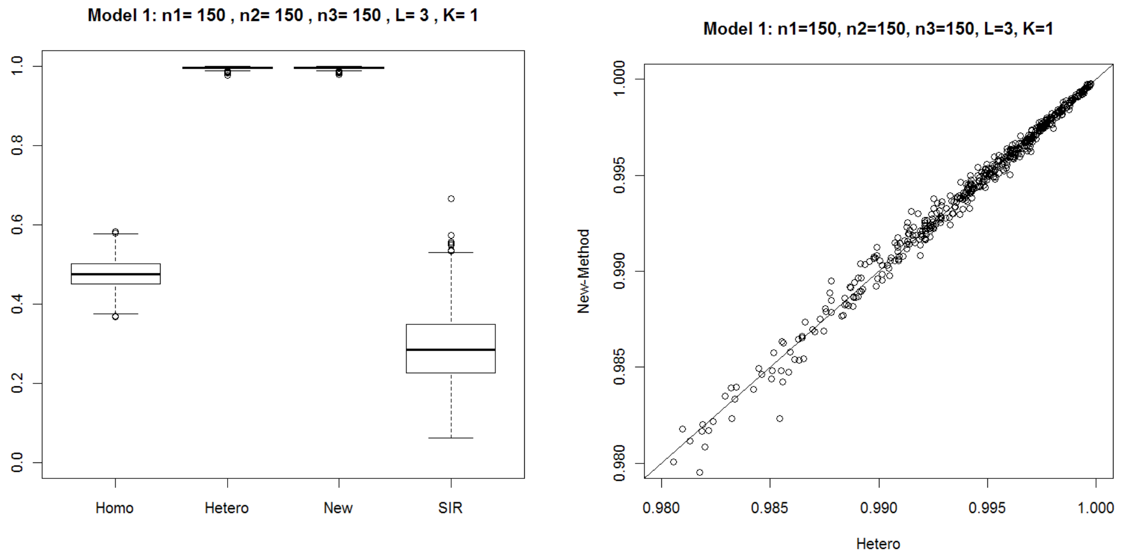

Figure 4: On the left: boxplots of the squared cosines for the four different methods for $n^{(1)} = n^{(2)} = n^{(3)} = 150$. On the right: comparison of the squared cosines for the "hetero" and "new" methods.

5.4 A multiple indices model ($K = 2$)

In this simulation, we generate simulated data from the following semiparametric regression model with two indices $X'\beta_1$ and $X'\beta_2$ and in presence of a categorical predictor with $L = 2$ levels:

$$\begin{cases} Y = (X'\beta_1)\exp(X'\beta_2) + \epsilon_1 & \text{for} \quad Z = 1, \\ Y = -(X'\beta_1)\exp(X'\beta_2) + \epsilon_2 & \text{for} \quad Z = 2, \end{cases} \tag{13}$$

where $X|Z = l$ (for $l = 1, 2$) follows a 5-dimensional normal distribution with mean $\mu^{(l)} = \mathbf{0}_5$ and randomly generated covariance matrices $\Sigma^{(l)}$ as in the previous subsection. Each random error term ϵ_l is standard normally distributed and is independent of X given $Z = l$. We take $\beta_1 = (1, 1, 0, 0, 0)'$ and $\beta_2 = (0, 0, 0, 1, 1)'$.

Results of a simulation study. From model (13), $N = 500$ samples of size $n = 400$ (with $n^{(1)} = n^{(2)} = 200$) were generated. For each simulated sample, the EDR direction was estimated with the four methods: "homo", "hetero", "new" and "SIR". We represent on Figure 5 the boxplots of the $N = 500$ efficiency measures obtained with each method. Not surprisingly, one can again observe that the naive "SIR" and "homo" approaches fail to recover the EDR direction. Both "new" and "hetero" methods outperform the previous ones since they are devoted to heteroscedastic setup.

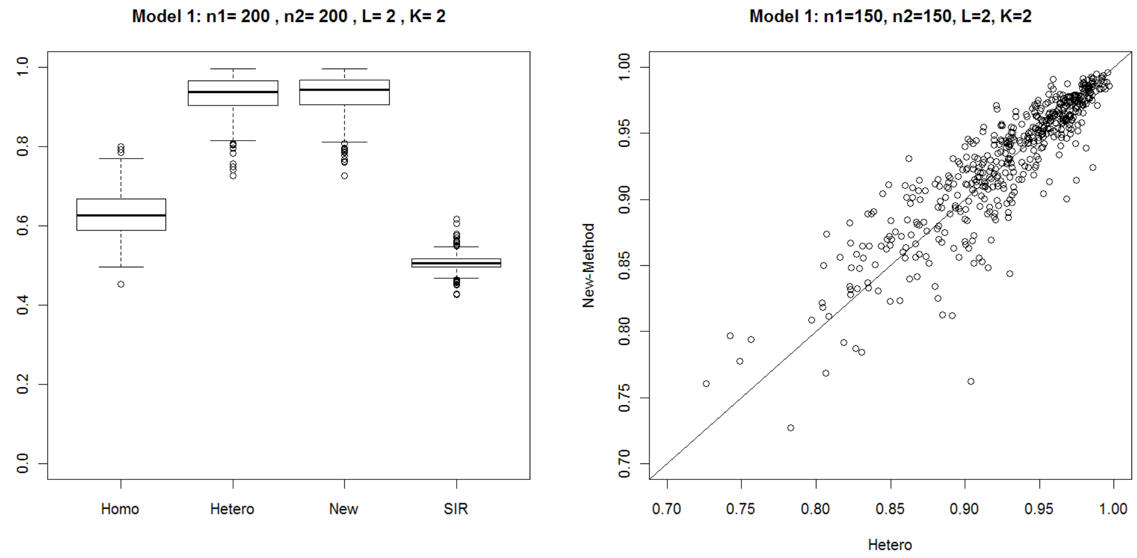

Figure 5: On the left: boxplots of the efficiency measures for the four different methods with $n^{(1)} = n^{(2)} = 400$. On the right: comparison of the efficiency measures for the "hetero" and "new" methods.

5.5 Case of unbalanced subpopulations

We consider here the case where the L subpopulations do not have the same weigths, that is the probability of the events $Z = l$, $p^{(l)} := \mathbb{P}(Z = l)$, can strongly differ. Then the associated subsample sizes $n^{(l)}$ can also be different. Let us assume that the sampling scheme is such that $\hat{p}^{(l)} := n^{(l)}/n = p^{(l)} + O_p(n^{-1/2})$ where $n = \sum_{l=1}^{L} n^{(l)}$ is the global sample size. We can notice that in the sample version of the pooled matrix of interest (10) of the "hetero" method, the weights $\hat{p}^{(l)}$ are used. It appears interesting to also take the weights $p^{(l)}$ or $\hat{p}^{(l)}$ into account in the population and sample versions of the "new" method. For that, we define the diagonal matrix D which contains the probabilities $p^{(l)}$. Then it can be shown that the eigenvector associated with the strictly positive eigenvalue of BDB' is an EDR direction. Let us introduce the diagonal matrix $\widehat{D}$ of the $\hat{p}^{(l)}$'s. The sample version of BDB' is obtained by substituting the empirical matrices by their theoretical counterparts. From a theoretical point of view, $\hat{B} = B + O_p(n^{-1/2})$ and $\hat{D} = D + O_p(n^{-1/2})$, then the estimated EDR direction converges at root n rate to the true one. The asymptotic distribution of the estimator can aslo be obtained with a similar proof given in Section 3 for the balanced subpopulations case. In this sequel this adaptation of the "new" method will be named "weighted new" method. We illustrate in the following the good numerical behaviour of the "weighted new" method with a simulation study.

We consider the single index model given in (11). We generate $N = 500$ data replications of sample of size $n = 200$ with $n^{(1)} = 70$ and $n^{(2)} = 130$. Figure 6 shows that both methods "homo" and "SIR" again fail to recover the EDR direction. On the contrary the "hetero" and "new" methods provide high efficiency measures. Furthermore we can see that taking into account the proportion of observations in each subsample increases the quality measure. We also observed on other simulations (not exhibited here) that the "weighted new" method outperforms the "new" one when the ratio $\hat{p}^{(1)}/\hat{p}^{(2)}$ hardly differs from one.

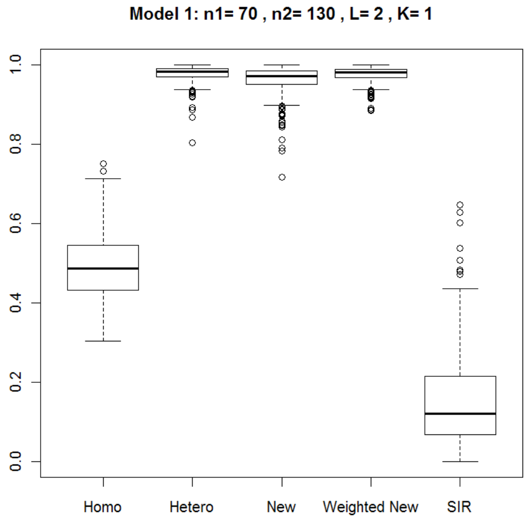

Figure 6: Boxplots of the efficiency measures for "homo", "hetero", "new", "weighted new" and "SIR" methods with unbalanced samples ($n^{(1)} = 70$ and $n^{(2)} = 130$).

5.6 A practical way to investigate a common EDR direction

Let us consider here a modified version of the semiparametric regression model (12):

$$\begin{cases} Y = (\beta + \theta_1\gamma_1)'X + \epsilon_1 & \text{for } Z = 1, \\ Y = ((\beta + \theta_2\gamma_2)'X/5)^3 + 0.1\epsilon_2 & \text{for } Z = 2, \\ Y = -(\beta'X)/2 + \epsilon_3 & \text{for } Z = 3. \end{cases} \quad (14)$$

In this model, the dimension reduction direction in each subpopulation can be different contrary to model (12). We take $\beta = (1, 1, -1, -1, 0)'$, $\gamma_1 = (1, 1, 0, 1, 1)'$ and $\gamma_2 = (-1, 0, 1, 0, -1)'$. The parameters θ_1 and θ_2 which belong to $[0; +\infty[$ allow to control the presence of a common dimension reduction direction in the model.

- When $\theta_1 = \theta_2 = 0$, there is a common dimension reduction direction β in each subpopulation. Hence, in practice, we can apply our approach on these global sample (with all subpopulations).

- When $\theta_1 > 0$ and $\theta_2 = 0$, there is only a common dimension reduction direction β in subpopulations such that $Z = 2$ or 3. From a practical point of view, in order to recover the two dimension reduction directions ($\beta + \theta_1\gamma_1$ for the first subpopulation, and β for the two other ones), we can apply SIR on the first subpopulation (such that $Z = 1$) and our approach on the two remaining subpopulations.

- When $\theta_1 > 0$ and $\theta_2 > 0$, there is no common dimension reduction direction in these 3 subpopulations. Hence, we have to apply SIR on each subpopulation in order to estimate the three directions, $\beta + \theta_1\gamma_1$ (subpopulation 1), $\beta + \theta_2\gamma_2$ (subpopulation 2) and β (subpopulation 3).

In this subsection, we propose a practical approach to investigate if there is a common dimension reduction direction for various subpopulations. For simplicity's sake, we limit the presentation to single index model (that is when $K = 1$). The idea is to use the following measure which will indicate if the direction estimated with our approach (based on some L subpopulations) is common for these L subpopulations (i.e. is in the intersection of the individual dimension reduction spaces of each subpopulation). Let us denote by $\widehat{P}_l$ the I_p-orthogonal projector onto $\widehat{E}_l$, the estimated EDR space of the subpopulation such that $Z = l$. Let us recall that $\widehat{b}$ is the direction estimated by our approach based on the L subpopulations. We consider (here for L subpopulations) the proximity measure

$$T_n = \left|\left|\prod_{l=1}^{L} \widehat{P}_l\widehat{b}\right|\right|$$

which belongs to [0,1] since $||\widehat{b}|| = 1$. If T_n is close to 1, $\widehat{b}$ can be considered as a common dimension reduction direction, either there is no common dimension reduction direction between the considered subpopulations (with our approach based on SIR). In order to view the variability of T_n, we consider $\mathcal{B} = 1000$ bootstrap replications of the observations, by resampling with replacement from the original data set. We calculate for the ath bootstrap sample $\widehat{P}_l^{(a)}$, $\widehat{b}^{(a)}$ and then $T_n(a) = ||\prod_{l=1}^{L} \widehat{P}_l^{(a)}\widehat{b}^{(a)}||$. Finally, we provide a histogram of the $T_n(a)$'s values.

To illustrate our proposed practical criterion, let us illustrate its numerical behaviour on three examples (for various values of θ_1 and θ_2). For each example, a sample of size $n = 300$ is generated from model (14). For each sample, we first evaluate $T_n =: T_n^{(1,2,3)} = ||\prod_{l=1}^{L=3} \widehat{P}_l\widehat{b}||$ and its bootstrap distribution. If a common direction for the $L = 3$ subpopulations is detected (with a value of $T_n^{(1,2,3)}$ close to one), the procedure is finished. If a common direction for the $L = 3$ subpopulations is not suspected, then we evaluate, for the couples (i, j) in the set $\{(1, 2), (1, 3), (2, 3)\}$, the following statistic $T_n^{(i,j)}$ and its bootstrap distribution in order to detect a possible common direction between only two subpopulations: $T_n^{(i,j)} = ||\widehat{P}_i\widehat{P}_j\widehat{b}^{(i,j)}||$ where $\widehat{b}^{(i,j)}$ is the estimator $\widehat{b}$ calculated using the data from the subpopulations such that $Z = i$ and $Z = j$.

Example 1: $\theta_1 = \theta_2 = 0$. We clearly observe in Figure 7 that $T_n^{(1,2,3)}$ is close to one. Then we can conclude that the common dimension reduction direction $\widehat{b}$ has to be retained for the $L = 3$ subpopulations.

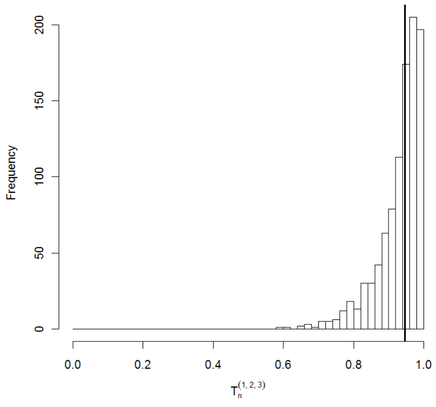

Figure 7: Evaluation of $T_n^{(1,2,3)}$ (bold vertical line) and its bootstrap distribution (histogram) for a sample from model (14) with $\theta_1 = \theta_2 = 0$.

Example 2: $\theta_1 > 0$ **and** $\theta_2 = 0$. We plainly observe in Figure 8 that $T_n^{(1,2,3)}$, $T_n^{(1,2)}$ and $T_n^{(1,3)}$ have values clearly lower than one contrary to $T_n^{(2,3)}$ which has a value close to one. Considering the first three graphics, it seems that there is no common dimension reduction direction between the 3 subpopulations, as well as between the subpopulations such that $Z = 1$ and $Z = 2$, and between the subpopulations such that $Z = 1$ and $Z = 3$. The last graphic allows us to conclude that a common dimension reduction direction $\widehat{b}^{(2,3)}$ has to be retained for the subpopulations such that $Z = 2$ and $Z = 3$. For the subpopulation $Z = 1$, the dimension reduction direction can be estimated via SIR using the subsample such that $Z = 1$.

Example 3: $\theta_1 > 0$ **and** $\theta_2 > 0$. In this example, there is no common dimension reduction direction in the model. This is confirmed from a graphical point of view in Figure 9 with the values of $T_n^{(1,2,3)}$, $T_n^{(1,2)}$, $T_n^{(1,3)}$ and $T_n^{(2,3)}$ clearly lower than one. Then we have to apply SIR individually on each subsample.

6 Concluding remarks

In this paper we propose a new estimator of the EDR space in a semiparametric regression model in presence of a categorical covariate which defines a stratification of the population. The main advantage of the proposed method is that the estimator can be used in homoscedastic and heteroscedastic setups. Moreover it has good theoretical properties and by this way it does not suffer from pathological cases. Root n consistency and asymptotic normality have been obtained. Extensions to multivariate indices model, multivariate dependent variable Y and SIR$_\alpha$-based approach have been described. We have also adapted our estimate to the case of unbalanced subpopulations. Simulation studies showed a good behavior of the proposed estimator in various situations. Moreover we have proposed a practical procedure in order to determine if the underlying model with a common EDR space for all the L subpopulations is adequate for the available data. The method has been implemented in R and the corresponding codes are available from the authors.

Acknowledgment

The authors are very grateful to the Editor and the referees for their valuable comments and constructive suggestions.

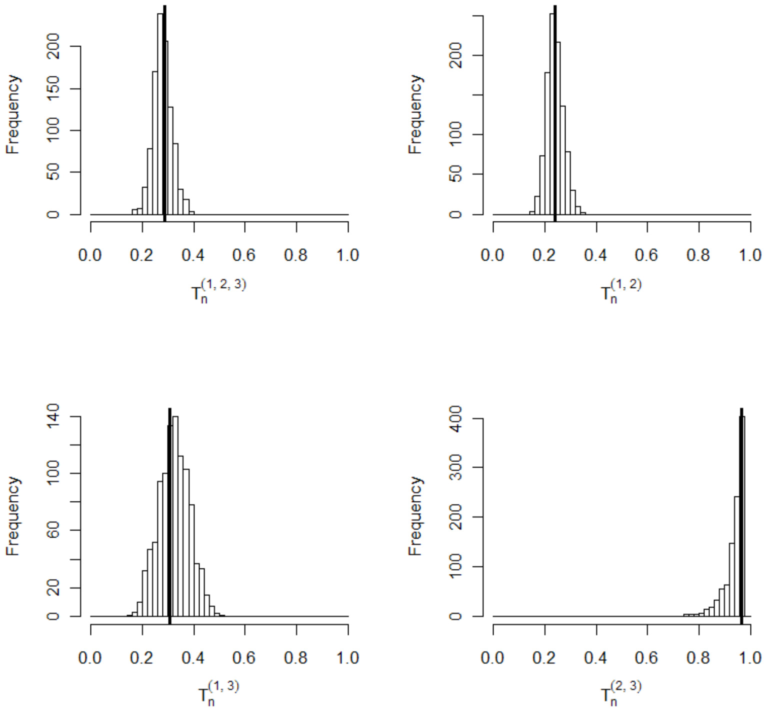

Figure 8: Evaluation of $T_n^{(1,2,3)}$, $T_n^{(1,2)}$, $T_n^{(1,3)}$ and $T_n^{(2,3)}$ (bold vertical lines) and their bootstrap distributions (histograms) for a sample from model (14) with $\theta_1 > 0$ and $\theta_2 = 0$.

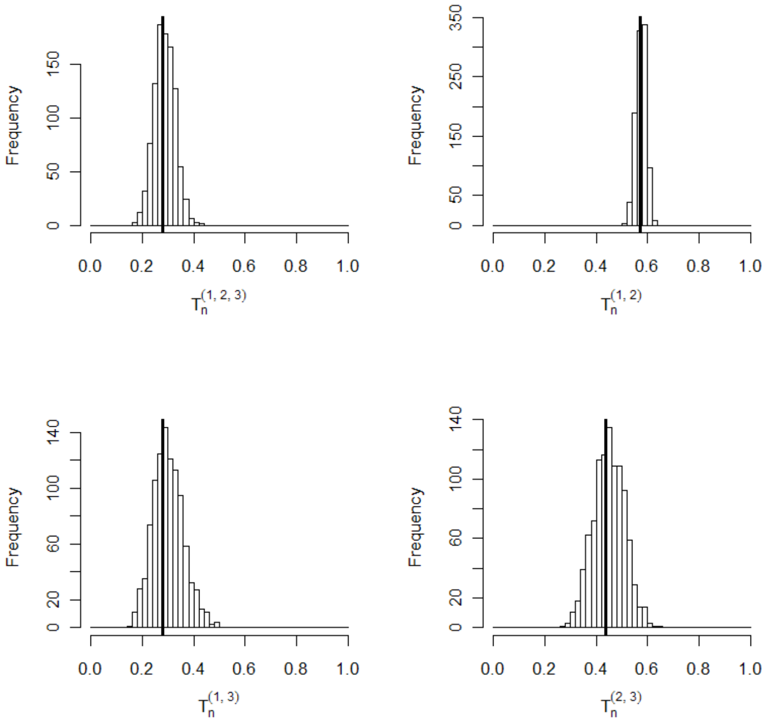

Figure 9: Evaluation of $T_n^{(1,2,3)}$, $T_n^{(1,2)}$, $T_n^{(1,3)}$ and $T_n^{(2,3)}$ (bold vertical lines) and their bootstrap distributions (histograms) for a sample from model (14) with $\theta_1 > 0$ and $\theta_2 > 0$.

References

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

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., & Saracco, 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.

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

Bura, E. & Cook, R.D. (2001b). Extending sliced inverse regression: the weighted chi-squared test. *Journal of the American Statistical Association*, **96**, 996-1003.

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

Carroll, R.J. & Li, K.C. (1995). Binary regressors in dimension reduction models: A new look at treatment comparisons. *Statistica Sinica*, **5**, 667-688.

Chiaromonte, F., Cook, R.D. & Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics*, **30**, 475-497.

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

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

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

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

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

Gannoun, A. & Saracco, J. (2003b). Two Cross Validation Criteria for SIR$_\alpha$ and PSIR$_\alpha$ methods in view of prediction. To appear in *Computational Statistics*, **4**.

Gather, U., Hilker, T. & Becker, C. (2002). A note on outlier sensitivity of sliced inverse regression. *Statistics*, **36**, 271-281.

Harville, D.A. (1997). *Matrix algebra from a statistician's perspective*, Springer-Verlag, New York.

Hsing, T. & 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, B., Cook, R.D. Chiaromonte, F. (2003a). Dimension reduction for the conditional mean in regressions with categorical predictors. *The Annals of Statistics*, **31**, 1636-1668.

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

Liquet, B. & Saracco, J. (2007). Pooled marginal slicing approach via SIR$_\alpha$ with discrete covariables. *Computational Statistics*, **4**, 599-617.

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.

Lue, H.-H. (2009). Sliced inverse regression for multivariate response regression.*Journal of statistical planning and inference*, **139**(8), 2656-266.

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

Saracco, J. (1999). Sliced Inverse Regression under linear constraints. *Communications in Statistics - Theory and methods*, **28**(10), 2367-2393.

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 SIR$_\alpha$. *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.

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

Wang, Q., Yin, X. (2008). Sufficient dimension reduction and variable selection for regression mean function with two types of predictors. *Statistics and Probability Letters*, **78**, 2798-2803.

Yin, X. (2005). Non-parametric estimation of direction in single-index models with categorical predictors. *Australian & New Zealand Journal of Statistics*, **47**(2), 147-161.

Yin, X. & Cook, R.D. (2005). Direction estimation in single-index regressions. *Biometrika*, **92**(2), 371-384.

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

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