Statistics and Its Interface Volume 9 (2016) 509–520

BIG-SIR a Sliced Inverse Regression approach for massive data

Benoit Liquet* and Jerome Saracco

In a massive data setting, we focus on a semiparametric regression model involving a real dependent variable Y and a p-dimensional covariate X (with $p \geq 1$). This model includes a dimension reduction of X via an index $X'\beta$. The Effective Dimension Reduction (EDR) direction β cannot be directly estimated by the Sliced Inverse Regression (SIR) method due to the large volume of the data. To deal with the main challenges of analyzing massive data sets which are the storage and computational efficiency, we propose a new SIR estimator of the EDR direction by following the "divide and conquer" strategy. The data is divided into subsets. EDR directions are estimated in each subset which is a small data set. The recombination step is based on the optimization of a criterion which assesses the proximity between the EDR directions of each subset. Computations are run in parallel with no communication among them.

The consistency of our estimator is established and its asymptotic distribution is given. Extensions to multiple indices models, q-dimensional response variable and/or SIR_α-based methods are also discussed. A simulation study using our `edrGraphicalTools` **R** package shows that our approach enables us to reduce the computation time and conquer the memory constraint problem posed by massive data sets. A combination of `foreach` and `bigmemory` **R** packages are exploited to offer efficiency of execution in both speed and memory. Results are visualized using the bin-summarise-smooth approach through the `bigvis` **R** package. Finally, we illustrate our proposed approach on a massive airline data set.

Keywords and phrases: High performance computing, Effective Dimension Reduction (EDR), Parallel programming, R software, Sliced Inverse Regression (SIR).

1. INTRODUCTION

Regression analysis studies the relationship between a response variable Y and a covariate 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 as kernel or smoothing splines methods). However as the dimension p of the covariate X becomes large, increased difficulties in modeling are often encountered. This is the well-known curse of dimensionality. One way to overcome this problem is to use dimension reduction techniques which aim at replacing X with a projection onto a smaller dimensional subspace.

In the framework of high dimensional data, the following semiparametric dimension reduction single index model has been proposed by Duan and Li [17]:

$$Y = f(X'\beta, \epsilon), \tag{1}$$

where the univariate response variable Y is linked with the p-dimensional regressor X (with expectation $\mathbb{E}(X) = \mu$ and covariance matrix $\mathbb{V}(X) = \Sigma$) only through the single index $X'\beta$. The error term ϵ is independent of X. The link function f and the vector β are unknown. Since β is not totally identifiable in this model, we are interested in finding the linear subspace spanned by β, called the Effective Dimension Reduction (EDR) space.

Li [31] introduced Sliced Inverse Regression (SIR) which is a computationally simple and fast method to estimate the EDR space without assuming neither the functional form of f 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 $\mathbb{E}(X|Y)$; see for instance Duan and Li [17], Carroll and Li [7], Hsing and Carroll [27], Zhu and Ng [50], Kötter [29], Saracco [38, 39], Aragon and Saracco [3], Bura and Cook [6] or Gather et al. [23] among others.

In this paper we focus on massive data sets, that is, the size of the data is large and analyzing it takes a significant amount of time and computer memory. Emerson & Kane [18] consider a data set large if it exceeds 20% of the RAM (Random Access Memory) on a given machine, and massive if it exceeds 50%. While SIR is a computationally simple and fast method, the current version implemented in our `edrGraphicalTools` [15] **R** package cannot directly handle massive data sets.

To tackle the analysis of massive data sets through the SIR approach, we introduce a new SIR estimator. The main idea follows the "divide and conquer" principle (also called

*Corresponding author.

"divide and recombine" [25, 46]) by processing the data by chunks (blocks) and combining/aggregating the results (see e.g., [11, 33]). The aggregation step is based on the optimization of a criterion which assesses the proximity between the EDR directions of each block. This optimization problem is similar to the one exploited by Chavent at al. [8] in a data stream context. Our new SIR (denoted BIG-SIR) estimator can easily handle massive data sets by using parallel computing as there is no dependency between parallel tasks and only the storage of the EDR directions in each chunk is required to get the final estimator.

This BIG-SIR estimator can be easily implemented using R software through our `edrGraphicalTools` package. We exploit the latest development in R for dealing with massive data sets exceeding available computer memory ([28, 45]). As recommended by Kane et al. ([28]) a combination of the `foreach` [1] and `bigmemory` [28] packages offers efficiency of execution in both speed and memory.

Finally, the link function between the variable of interest and the common estimated index can be first nonparametrically estimated with any smooth method, and subsequently parametrically modeled if necessary. However, this task is still not straightforward in the massive data set setting. Lumley [37] developed the R package `biglm` which offers the possibility to fit a linear generalized model. More recently, Wood et al. [49] proposed the `bam` function from the `mvcv` R package to perform generalized additive models for large data sets. In this work, we use a bin-summarise-smooth approach developed by Wickham [47] and available through the R package `bigvis` [48] which enables us to visualize and display an estimation of the link function in the massive data set setting.

The remainder of the paper is structured as follows: in Section 2, after a brief review on SIR, we introduce our BIG-SIR estimator for massive data. Both population and sample versions are described. Several extensions of this approach are presented in Section 3: (i) multiple indices models; (ii) symmetric dependent models; (iii) models with a multivariate response variable. A simulation study is carried out in Section 4 in order to illustrate the behavior of our estimator and to compare it to classical SIR. Different strategies are compared using parallel computing and a memory mapping approach. R code is presented through the simulation section and results are visualized using the bin-summarise-smooth approach through the `bigvis` R package. In section 5, we illustrate the proposed BIG-SIR approach on a massive airline data set.

2. AN SIR ESTIMATOR FOR MASSIVE DATA: BIG-SIR

Let us first recall in Section 2.1 the population and sample versions of SIR based on the whole data set. Then, the population and sample versions of our BIG-SIR estimator for massive data are presented in Section 2.2.

2.1 Brief review of usual SIR

Inverse regression step The basic principle of the SIR method is to reverse the role of Y and X, that is, instead of regressing the univariate variable Y on the multivariate variable X, the covariate X is regressed on the response variable Y. The price we have to pay in order to succeed in inverting the role of X and Y is an additional assumption on the distribution of X, named the linearity condition (described hereafter).

Usual SIR estimate is based on the first moment $\mathbb{E}(X|Y)$. It has been initially introduced by Duan and Li [17] for the single index model and by Li [31] for the multiple indices model.

Recall now the geometric property on which SIR is based. Let us introduce the linearity condition (LC):

$$\forall \theta \in \mathbb{R}^p, \mathbb{E}(X'\theta|X'\beta) \text{ is linear in } X'\beta. \tag{2}$$

Note that this condition is satisfied when X is elliptically distributed (for instance normally distributed). The reader can find an interesting discussion on this linearity condition in [10].

Assuming model (1) and (LC), Li [31] showed that the centered inverse regression curve is contained in the linear subspace spanned by $\Sigma\beta$. Let T denote a monotonic transformation of Y. He considered the eigendecomposition of the Σ-symmetric matrix $\Sigma^{-1}M$ where $M = \mathbb{V}(\mathbb{E}(X|T(Y)))$. Straightforwardly the eigenvector u of $\Sigma^{-1}M$ associated with the non-null eigenvalue is an EDR direction (i.e., is collinear with β). The vector u is Σ-normalized. Let us define b the I_p-normalized version of u as $b = u/||u||$ with $||u||^2 = u'u$.

Slicing step To easily estimate the matrix M, Li [31] proposed a transformation T, called a slicing, which categorizes the response Y into a new response. The support of Y is partitioned into H non-overlapping slices $s_1, \ldots, s_h, \ldots, s_H$. With such a transformation T, the matrix of interest M can be now written as $M = \sum_{h=1}^{H} p_h(m_h - \mu)(m_h - \mu)'$ where $p_h = \mathbb{P}(Y \in s_h)$ and $m_h = \mathbb{E}(X|Y \in s_h)$.

Estimation process When a sample $\{(X_i, Y_i), i = 1, \ldots, n\}$ is available, matrices Σ and M are estimated by substituting empirical versions of the moments for their theoretical counterparts. Let

$$\widehat{M} = \sum_{h=1}^{H} \hat{p}_h(\hat{m}_h - \hat{\mu})(\hat{m}_h - \hat{\mu})', \tag{3}$$

where $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[Y_i \in s_h]$ and $\hat{m}_h = \frac{1}{n\hat{p}_h}\sum_{i=1}^{n} X_i\mathbb{I}[Y_i \in s_h]$. Therefore the estimated EDR direction $\widehat{u}$ is the eigenvector associated with the largest eigenvalue of $\widehat{\Sigma}^{-1}\widehat{M}$. Let us highlight that $\widehat{\Sigma}$ is assumed to be invertible which implies that $n > p$. The vector $\widehat{u}$ is $\widehat{\Sigma}$-normalized. Let us define $\widehat{b}$ the I_p-normalized version of $\widehat{u}$ as $\widehat{b} = \widehat{u}/||\widehat{u}||$. Convergence

at rate $\sqrt{n}$ and asymptotic normality of estimated EDR direction have been obtained, see [31, 38] for instance.

The choice of the slicing T is discussed in [30, 31, 40] but, theoretically, there is no optimal one. In practice, we fix the number of observations per slice to $\lfloor n/H \rfloor$ where $\lfloor a \rfloor$ stands for the integer part of a. If the sample size n is not proportional to the number H of slices, some slices will then contain $\lfloor n/H \rfloor + 1$ observations.

Standardized version for SIR Another way to obtain a basis of the EDR space is to consider the eigendecomposition of $\Sigma^{-1/2} M^* \Sigma^{-1/2}$, that is the eigendecomposition of $M^* = \mathbb{V}(\mathbb{E}(Z|T(Y))$ where $Z = \Sigma^{-1/2}(X - \mu)$ is the standardized version of the covariate X. Define η as the eigenvector of the I_p-symmetric matrix M^* associated with the largest eigenvalue. Transforming back to the original scale, the vector $\Sigma^{-1/2}\eta$ is an EDR direction. The estimation procedure is a straightforward replication of the previous estimation process using $\widehat{M}^* = \widehat{\Sigma}^{-1/2} \widehat{M} \widehat{\Sigma}^{-1/2}$.

Computational complexity and data storage The computational complexity of SIR is of order $p^2(n+p)$ (denoted as $O(p^2(n+p))$ hereafter). The first term (np^2) corresponds to the cost of computing the empirical covariance matrix $\widehat{\Sigma}$, the second term (p^3) is the cost for computing the matrix $\widehat{\Sigma}^{-1}\widehat{M}$ and its eigendecomposition. SIR requires the storage of the whole matrix of regressors which corresponds to $O(np)$. This is clearly problematic in a massive data set setting.

2.2 Population and sample version of BIG-SIR

Our proposed BIG-SIR estimator deals with the issue of analyzing massive data sets which exceed 50% of the RAM or cannot be loaded in a single computer. BIG-SIR is based on the divide and recombine (D&R) principle (also called divide and conquer) which consists in: (i) divide the massive data set into G chunks (blocks); (ii) apply the usual SIR estimator on each block separately; (iii) combine the EDR directions from each block to get a solution to the full data. Define $b_1, \ldots, b_g, \ldots, b_G$ the EDR directions obtained from each block. We exploit the same idea as in Chavent et al. [8] to aggregate the EDR directions by recovering the direction "most collinear" with the vectors $b_1, \ldots, b_G$. Noting that the collinearity between two unit vectors a and b is measured by $m(a,b) = cos^2(a,b) = (a'b)^2$, the following optimization problem is proposed:

$$\max_{a \in \Re^p} \sum_{g=1}^{G} w_g m(b_g, a) \quad \text{s.t. } ||a|| = 1. \tag{4}$$

where the w_g's are positive weights such that $\sum_{g=1}^{G} w_g = 1$. These weights allow the algorithm to take into account different block sizes.

Theorem 2.1. *(i) The solution $v_G \in \Re^p$ of the maximization problem (4) is the normalized principal eigenvector of*

$$M_G = \sum_{g=1}^{G} w_g b_g b_g' \tag{5}$$

associated to the largest eigenvalue $\sum_{g=1}^{G} w_g m(b_g, v_G)$.
(ii) Under the linearity condition (LC) and model (1) for each block g, v_G is an EDR direction.

The proof is similar to the one offered by Chavent et al. [8] in a data stream context (see Appendix of [8]). Note that their optimization problem takes into account the possible evolution of the parametric part of the semiparametric model in each block. We adopt a different setting by assuming the same model in each block. In this context, our matrix of interest is similar to the one proposed in the SIR approach for a stratified population developed by Chavent et al. [9].

Sample version of BIG-SIR For $g = 1, \ldots, G$, let us denote by $\hat{b}_g$ the estimator of the EDR direction calculated on each block G. The estimator $\widehat{v}_G$ of the EDR direction v_G with the BIG-SIR approach is the principal eigenvector of the $p \times p$ matrix defined as

$$\widehat{M}_G = \sum_{g=1}^{G} w_g \hat{b}_g \hat{b}_g' \tag{6}$$

where $w_g = \frac{n_g}{\sum_{g=1}^{G} n_g}$ for $g = 1, \ldots, G$. I.e., we take as weights the relative size of the block g. The most natural way is to divide the data set into blocks with equal sample size $(n_g = n/G)$. Convergence at rate $\sqrt{n}$ and the asymptotic normality of estimated EDR direction have been obtained, see Chavent et al. [8].

Computational complexity and data storage The computational complexity of BIG-SIR is given by:

$$O\left(G\left(\frac{n}{G}p^2\right) + Gp^2 + Gp^3\right) = O\left(np^2 + Gp^2 + Gp^3\right),$$

where $O(np^2)$ corresponds to the cost of computing the empirical covariance matrices $\widehat{\Sigma}$ in each block g, $O(Gp^2)$ represents the calculation of the matrix $\widehat{M}_G$ and $O(Gp^3)$ stands for the cost of the eigendecompositions. The computational complexity of BIG-SIR is greater than SIR. However, BIG-SIR can handle massive data sets as the method could be performed in a parallel computing environment. Then, each cluster requires the storage of only a subset of the matrix of regressors which corresponds to $O((n/G)p)$. The final computation of the estimated EDR direction by BIG-SIR requires only the storage of the G EDR directions computed on each cluster.

3. SOME 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 an SIR_α-based approach rather than classical SIR. The last extension investigates the case when the dependent variable Y is multivariate.

3.1 Extension to multiple indices models

This first extension is similar to the one we have proposed in a data stream setting (see Chavent et al. [8]). We present in the following, the main idea and procedure. The response variable Y is related to the p-dimensional quantitative regressor X (with $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \Sigma$) only through the K indices $X'\beta_k$:

$$Y = h(X'\beta_1, \ldots, X'\beta_K, \varepsilon). \tag{7}$$

As in the single index model, the error term ε is independent of X and the link function h is unknown. In other words, Y and X are independent conditionally on $(X'\beta_1, \ldots, X'\beta_K)$. In this multiple indices model, we search for a basis that spans the K-dimensional EDR space $E = \text{Span}(\beta_1, \ldots, \beta_K)$. As for the single index model, we will seek, using SIR, for a basis of the EDR space for each block. In order to get theoretical results, we need to adapt the linearity condition and we now assume that for any $\theta \in \mathbb{R}^p$ we have

(LC') $\mathbb{E}(X'\theta | X'\beta_1, \ldots, X'\beta_K)$ is linear in $X'\beta_1, \ldots, X'\beta_K$.

For each block g, the eigenvectors $u_{1,g}, \ldots, u_{K,g}$ associated with the K largest eigenvalues of the matrix $\Sigma^{-1}M_g$ are EDR directions, where the matrix M_g corresponds to the matrix, M, for the g-th chunk (block). Note that the number H of slices for each block must be greater than K in order to avoid artificial dimension reduction. Let us define the matrix $\mathbb{U}_g = [u_{1,g}, \ldots, u_{K,g}]$ containing these EDR directions which form a Σ-orthogonal basis of E. Then the first K eigenvectors, $b_{1,g}, \ldots, b_{K,g}$ of the matrix $U_g U_g'$ form an I_p-orthonormal basis of E. We store them in the $p \times K$ matrix $\mathbb{B}_g = [b_{1,g}, \ldots, b_{K,g}]$.

Population and sample version of BIG-SIR For $K > 1$, the optimization problem (4) requires some modifications. Direction b_g is replaced by an I_p-orthonormal basis $\mathbb{B}_g$ of the EDR space and the proximity measure between two linear subspaces spanned by $\mathbb{B}_g$ and $\mathbb{B}_l$ from the blocks g and l is defined by:

$$m(\mathbb{B}_g, \mathbb{B}_l) = \frac{\text{Trace}(P_g P_l)}{K},$$

where $P_t = \mathbb{B}_t(\mathbb{B}_t'\mathbb{B}_t)^{-1}\mathbb{B}_t'$ is the I_p-orthogonal projector onto $\text{Span}(\mathbb{B}_t)$, the EDR space obtained from block t (equal to g or l). This measure takes its values in [0,1]. Note that $m(\mathbb{B}_g, \mathbb{B}_l) = 1$ when $\text{Span}(\mathbb{B}_g)$=$\text{Span}(\mathbb{B}_l)$. The closer this measure to one, the closer is the linear subspace $\text{Span}(\mathbb{B}_g)$ is to the linear subspace $\text{Span}(\mathbb{B}_l)$.

Let $\mathbb{A}$ be a $p \times K$ matrix such that $\mathbb{A}'\mathbb{A} = I_K$. Now define $Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_G)$ to be the proximity measure between the linear subspace $\text{Span}(\mathbb{A})$ and the EDR spaces $\text{Span}(\mathbb{B}_1)$,...,$\text{Span}(\mathbb{B}_G)$ respectively obtained from the G blocks:

$$Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_G) = \sum_{g=1}^{G} w_g m(\mathbb{A}, \mathbb{B}_g),$$

where the w_g's are positive weights such that $\sum_{g=1}^{G} w_g = 1$. This measure takes its values in [0,1]. Note that $Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_G) = 1$ when $\text{Span}(\mathbb{A}) = \text{Span}(\mathbb{B}_g) = \ldots = \text{Span}(\mathbb{B}_G)$. The closer this measure to one, the closer is the linear subspace $\text{Span}(\mathbb{A})$ is to all the G linear subspaces $\text{Span}(\mathbb{B}_g)$, $g = 1, \ldots, G$.

The maximization problem for the multiple indices model is defined as

$$\max_{\mathbb{A}} Q(\mathbb{A}, \mathbb{B}_1, \ldots, \mathbb{B}_G) \quad \text{s.t. } \mathbb{A}'\mathbb{A} = I_K. \tag{8}$$

Theorem 3.1. *(i) The solution $\mathbb{V}_G = [v_{1,G}, \ldots, v_{K,G}]$ of the maximization problem (8) is an I_p-orthonormal basis of the K-dimensional eigenspace associated with the K largest eigenvalues $\lambda_1, \ldots, \lambda_K$ of the $p \times p$ matrix*

$$\mathbb{M}_G = \sum_{g=1}^{G} w_g \frac{\mathbb{B}_g \mathbb{B}_g'}{K}. \tag{9}$$

Moreover $Q(\mathbb{V}_G, \mathbb{B}_1, \ldots, \mathbb{B}_G) = \lambda_1 + \ldots + \lambda_K$.
(ii) Under the linearity condition (LC') and model (7), the column vectors of $\mathbb{V}_G$ form an I_p-orthonormal basis of the EDR space E.

The proof is similar to the one offered by Chavent et al. [8] in a data stream context (see Appendix of [8]).

Sample version of BIG-SIR For each block g, using the corresponding sample, a $\widehat{\Sigma}$-orthogonal basis of the EDR space is first estimated with SIR. The basis vectors are stored in the matrix $\widehat{\mathbb{U}}_g$. Then the first K eigenvectors of the matrix $\widehat{\mathbb{U}}_g \widehat{\mathbb{U}}_g'$ are computed and stored in the matrix $\widehat{\mathbb{B}}_g$. They form an I_p-orthogonal basis of the estimated EDR space. Finally the estimator of $\mathbb{M}_G$ is constructed as follows:

$$\widehat{\mathbb{M}}_G = \sum_{g=1}^{G} w_g \frac{\widehat{\mathbb{B}}_g \widehat{\mathbb{B}}_g'}{K}.$$

The K eigenvectors associated with the K major eigenvalues of this matrix $\widehat{\mathbb{M}}_G$, denoted by $\widehat{\mathbb{V}}_G = [\hat{v}_{1,G}, \ldots, \hat{v}_{K,G}]$ provide an I_p-basis of the estimated EDR space denoted $\widehat{E}$. The convergence at rate $\sqrt{n}$ and the asymptotic normality of estimated EDR directions can be obtained as in Chavent et al. [8].

Choice of dimension K In most applications the number of indices, K, is a priori unknown and hence must be estimated from the data. 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

[31], Schott [42] or Barrios and Velilla [5]. 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 [19] or Liquet and Saracco [35] for a graphical bootstrap based approach.
In the massive data set context, under assumption (LC'), the dimension K is common to all the blocks since it is assumed that the underlying model in each block relies on the same EDR space E. From the theoretical point of view, it can thus be estimated from any block or from any combinations of blocks. From the practical point of view, we recommend choosing the dimension K using classical SIR in one block chosen randomly.

3.2 Extension to SIR$_\alpha$

The proposed method described in Section 2.1 is based on SIR, also named SIR-I, which relies on a geometric property of the conditional expectation (first moment) of X given $T(Y)$. Cook and Weisberg [12] 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 [31] introduced the SIR-II approach relying on a property of $\mathbb{V}(X|T(Y))$, and Cook and Weisberg [12] developed the sliced average variance estimator (SAVE) approach, see also Cook [13]. In order to conjugate information from SIR-I and SIR-II approaches and for increasing the chance of discovering all the EDR directions, Li [31] 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. For a multiple indices model, this assumption is written as follows:

$$\text{(CV)} \qquad \mathbb{V}(X|X'\beta_1, \ldots, X'\beta_K) \text{ is non-random.}$$

Note that the (LC') and (CV) conditions are satisfied when X has a multivariate normal distribution.

Let us give now a brief overview of the SIR-II and SIR$_\alpha$ approaches. The SIR-II matrix of interest is defined by $M_{II} = \mathbb{E}\left\{(V - \mathbb{E}(V))\,\Sigma^{-1}\,(V - \mathbb{E}(V))'\right\}$ where $V = \mathbb{V}(X|T(Y))$. Under model (7) and the (LC') and (CV) assumptions, it can be shown that the eigenvectors associated with the largest K eigenvalues of $\Sigma^{-1}M_{II}$ are some EDR directions. In SIR$_\alpha$ approach, the eigendecomposition of the matrix $\Sigma^{-1}M_\alpha$ where $\alpha \in [0,1]$ and $M_\alpha = (1-\alpha)M_I\Sigma^{-1}M_I + \alpha M_{II}$. It can also be proved that the eigenvectors associated with the largest K eigenvalues of $\Sigma^{-1}M_\alpha$ are some EDR directions, see Li [31]. Let us remark that, when $\alpha = 0$ (resp. $\alpha = 1$), SIR$_\alpha$ is equivalent to SIR-I (resp. SIR-II).

When transformation T is a slicing which partitions the support of Y into $H > K$ non-overlapping slices s_h, the matrix M_{II} is now written as $M_{II} = \sum_{h=1}^{H} p_h \left(V_h - \overline{V}\right)\Sigma^{-1}\left(V_h - \overline{V}\right)$, where $V_h = \mathbb{V}(X|Y \in s_h)$ and $\overline{V} = \sum_{h=1}^{H} p_h V_h$. It is straightforward to estimate the matrices M_{II} and M_α 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 $\sqrt{n}$ rate when the corresponding eigenvalues are assumed to be distinct, see for instance Li [31] or Saracco [40]. Asymptotic normality of the SIR$_\alpha$ estimates has been studied by Gannoun and Saracco [21].

An extension of the proposed approach is to replace the SIR-I estimators of the EDR directions by the corresponding SIR$_\alpha$ 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.

Choice of α The practical choice of α can be based on hypothesis test approach (see Saracco [40]) or on cross-validation criterion (see Gannoun and Saracco [22]). A graphical bootstrap based approach has also been developed by Liquet and Saracco [35] in order to select simultaneously the couple (α, K).

3.3 Extension to a multivariate dependent variable Y

Several authors (see for instance Aragon, [2], Hsing [26], Li et al., [32], Lue, [36]) extended 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 $\mathbb{R}^q$-valued. A few methods based on SIR-I approach have been developed in this multivariate context. Saracco [41] and Barreda et al. [4] focused on some extensions of the existing multivariate SIR approaches relying on the SIR$_\alpha$ method. 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 block g. As in Liquet and Saracco [34], we suggest to use the PMS$_\alpha$ approach which is a Pooled Marginal Slicing method based on SIR$_\alpha$; see Saracco [41] for details. An alternative is to use the recent multivariate SIR (MSIR) approach proposed by Coudret et al. [14] for estimating the K-dimensional EDR space which is common to the q components of the multivariate response variable. Similarly to the pooled marginal slicing (PMS), MSIR relies on the univariate version of SIR, applied to each component of Y. An advantage of MSIR is to offer a way to cluster components of Y related to the same EDR space.

4. SIMULATION STUDIES

In this section we use **R** to carry out simulation studies in order to illustrate the numerical behaviour of the new proposed approach. Some of the code is presented through

this section for the purpose of highlighting the simplicity of the implementation for analyzing massive data sets with a semiparametric model through R software [44]. The experiments have been conducted using a laptop with a 2.53 GHz processor and 8 GB of memory.

First we introduce in section 4.1 two simulated models: a single index model ($K = 1$) and a multiple indices model ($K = 2$) with some symmetric dependence. Numerical results are presented in section 4.2 by comparing and investigating the quality and the running time of our new BIG-SIR approaches compared to traditional SIR approaches performed without dividing the whole data set. Different computational strategies for performing our BIG-SIR estimator are presented and compared to speed-up the computational task.

4.1 Simulated models

In this simulation study, two semiparametric regression models are considered:

$$Y = \frac{4}{10}(X'\beta)^3 + \epsilon, \tag{10}$$

and

$$Y = (X'\beta_1)^2 + (X'\beta_2)^2 + \epsilon, \tag{11}$$

where X follows the p-dimensional normal distribution $\mathcal{N}_p(0_p, \Sigma)$ with a covariance matrix Σ arbitrarily chosen as follows: a matrix A is randomly filled using the uniform distribution on $[-1, 1]$, then $\Sigma = AA' + I_p$ in order to avoid possible problems of inversion of Σ. The error term ϵ follows the normal distribution $\mathcal{N}(0, \sigma^2)$ and is independent of X. Model (10) is a single index model while model (11) is a multiple indices model presenting a symmetric dependence. We set $p = 10$, $\beta = \beta_1 = (1, -1, 2, -2, 0, \ldots, 0)'/\sqrt{10}$, $\beta_2 = (0, \ldots, 0, 1, -1, 2, -2)'/\sqrt{10}$ and $\sigma = \sqrt{2}$. Figure 1 presents the link functions and the scatterplots of the true index (indices) versus Y for models (10) and (11) using a small data set of $n = 500$ observations.

The EDR direction of model (10) is estimated through SIR-I approach while the EDR space of model (11) is estimated with SIR-II approach as the model includes a symmetric dependence. Our `edrGraphicalTools` R package can handle the estimation of both models providing estimation of the EDR space in the first step and nonparametric estimation of the link functions in a second step. As an illustration, synthetic R code is presented in the following for estimating model (11):

```
R> model.11 <- edr(Y,X,H=8,K=2,method="SIR-II")
R> plot(model.11)
```

The proximity measure m defined in section 3.1 evaluates the quality of the estimated EDR space. Recall that this measure belongs to $[0, 1]$. The closer this value is to one, the better is the estimation. When $K = 1$ (single index model),

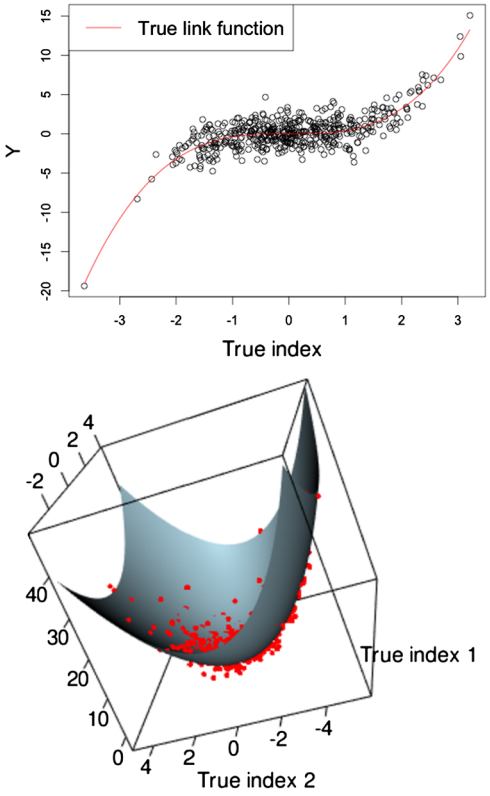

Figure 1. Link functions and scatterplots of the true index $X'\beta$ (respectively true indices $X'\beta_1$ and $X'\beta_2$) versus Y for model (10), on the top, and (11), at the bottom, using a small data set of $n = 500$ observations.

this measure is the squared cosine between the true EDR direction and the estimated one. As an example, we obtain for a sample size of $n = 500$ a quality measure equals to 0.93 for SIR-I approach on model (10) and a quality measure equals to 0.76 for SIR-II approach on model (11).

4.2 Numerical results

Let us now consider a massive data set framework by simulating large sample size data sets consuming a lot of memory in the R statistical programming environment. First, for each model, numerical results obtained with traditional SIR and BIG-SIR approaches are compared using the quality measure m. Then, we focus on different strategies to speed-up the computation of BIG-SIR. The running time of these strategies is investigated and compared to a traditional SIR analysis when the whole data set can still be loaded into memory in the R statistical programming environment. Finally, the estimation and the visualisation of the link function is discussed.

4.2.1 Comparison of SIR and BIG-SIR approaches

SIR approaches for the whole data set are compared to our "divide and conquer" BIG-SIR approaches through the

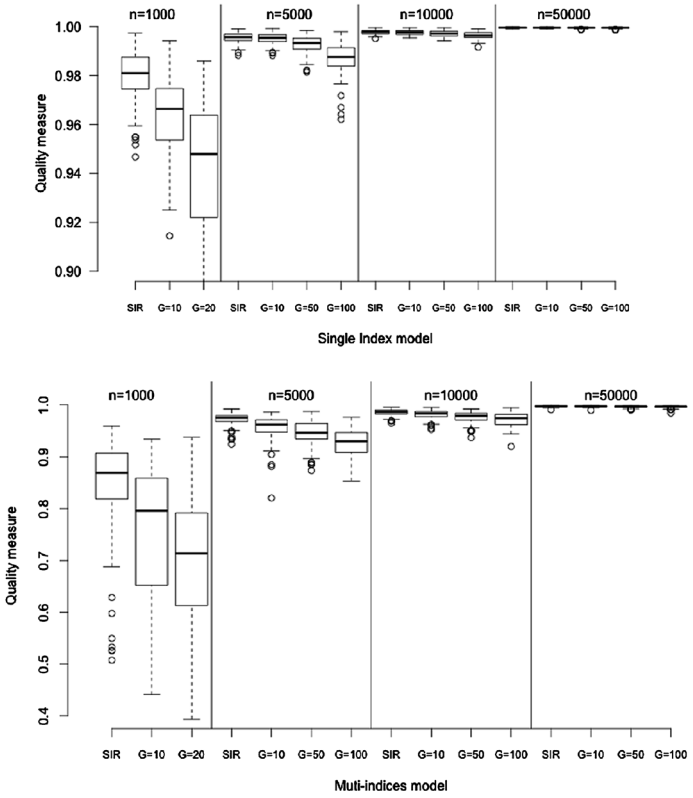

Figure 2. Boxplots of the quality measures between the true EDR direction and the EDR directions estimated with SIR and BIG-SIR for different chunk sizes for various sample size. Results for single index model on top panel and results for multiple indices model on bottom panel.

quality measure m for various sample sizes $n \in \{10^3, 5 \times 10^3, 10^4, 5 \times 10^4, 10^5, 10^6, 10^7\}$ corresponding respectively to storage requirements of approximately 86 GB, 430 GB, 860 GB, 4 MB, 8 MB, 84 MB, 840 MB, for the full data set. In this simulation, we use different number of chunks (blocks) $G \in \{10, 20, 50, 100\}$. For each model, $M = 500$ data replications of data sets are generated as previously. Figure 2 shows some boxplots of the quality measure of the corresponding estimated EDR directions for different sample sizes.

Not surprisingly, the quality measure increases with the sample size. While, as expected, the SIR approaches appear to provide slightly better results than BIG-SIR, it is evident that as the sample size increases, BIG-SIR performs effectively just as well. But, as mentioned previously, the disadvantage of the SIR approaches is in the requirement of the storage of the big data set which is problematic for massive data sets. BIG-SIR methods overcome this issue by keeping only the estimated EDR directions of each chunk (block) in memory, which is an interesting gain in storage. The price to pay is a small loss of quality in the estimation of the EDR directions. However, this loss is generally insignificant for data sets with a sample size of more than 50,000 observations. Note that for higher sample sizes the quality measures in both cases are very close to 1, hence we do not present those results.

4.2.2 Computing strategies and running time

Here we focus on different strategies for computing in the most efficient way our BIG-SIR estimators. The following strategies are investigated:

1. Using loops: (i) load the whole data set into **R**; (ii) split the data set into G subsets; (iii) apply the SIR approach on each subset in a sequential way using a loop; (iv) combine the G EDR directions to get the final estimator.
2. `foreach`: (i) load the whole data set into **R**; (ii) split the data set into G subsets; (iii) apply SIR on each subset using the `foreach` package for running in parallel the G EDR estimations. (iv) combine the G EDR directions to get the final estimator.
3. Combining `bigmemory` and `foreach`: (i) use memory-mapped files (called "filebacking") through the `bigmemory` package to allow matrices to exceed the RAM size. The creation of the filebacking avoids significant memory overhead. A `big.matrix` is created which supports the use of shared memory for efficiency in parallel computing. (ii) combine it using the `foreach` strategy to implement the parallel computation of the G EDR directions. (iii) combine the G EDR directions to get the final estimator.

The running time (in seconds) of these different strategies are now compared against the SIR approaches on the whole data set. From model (10), $M = 50$ data sets are generated for various values of the sample size ($n \in \{5 \times 10^5, 10^6, 5 \times 10^6, 10^7\}$), the dimension $p \in \{5, 10, 20, 30, 50, 80\}$ of the covariate X and the number G of chunks ($G \in \{10, 50, 100, 150, 500\}$) required for BIG-SIR estimator. Figure 3 presents the computational times measured for SIR and all BIG-SIR strategies. The parallel strategies utilize four processor cores.

As expected, we can clearly observe in Figure 3 that the BIG-SIR approaches outperforms SIR for large and massive data sets (>200 MB). The combination of `bigmemory` and `foreach` for computing BIG-SIR appears to be the best strategies for large and massive data sets irrespective of the number of chunks. This has a small effect on the running time for massive data set. Moreover, only this strategy can scale past the size of available RAM by using `bigmemory` to manage big matrix data which allows parallel workers to receive descriptors of a big matrix data and then to attach to a `big.matrix` object, rather than transmitting the data for an entire big matrix (see [28]). The **R** code to implement BIG-SIR for analyzing model (10) through the combination of `bigmemory` and `foreach` is presented in the Appendix.

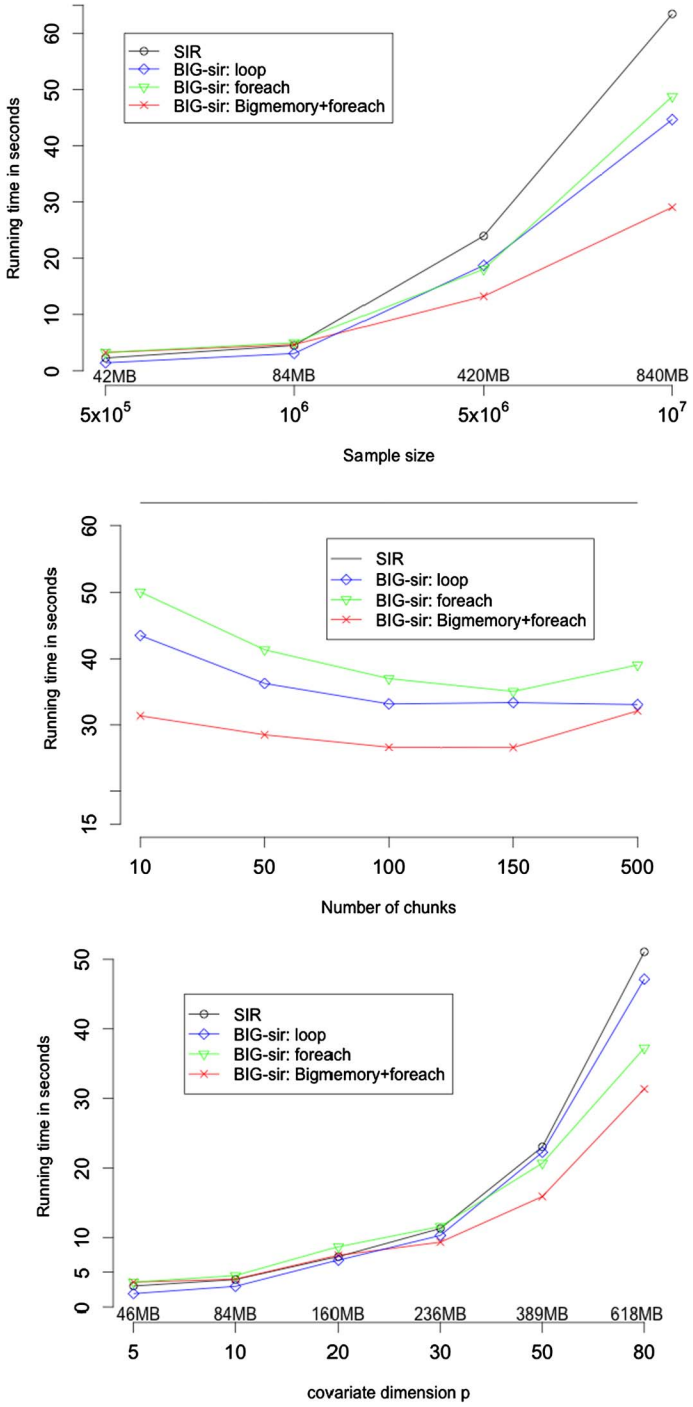

Figure 3. Mean of the running time (over 50 replications) for SIR and 3 strategies to compute BIG-sir ("loop", "foreach", "Bigmemory+foreach") for various values of n, G and p. Top panel: $n \in \{5 \times 10^5, 10^6, 5 \times 10^6, 10^7\}$ for fixed $G = 10$ chunks and $p = 10$. Middle panel: $G \in \{10, 50, 100, 150, 500\}$ for fixed $n = 10^7$ and $p = 10$. Bottom panel: $p \in \{5, 10, 20, 30, 50, 80\}$ for fixed $n = 10^6$ and $G = 10$.

4.2.3 Visualisation of BIG-SIR outputs

In a first step, BIG-SIR approaches provide an estimation of the true EDR space which are used to construct the K estimated indices. In a second step, the link function could be estimated non-parametrically through any smooth methods [24] based on the K-estimated indices and the response variable Y. While our BIG-SIR approach overcomes the big-data challenge associated with estimation of the EDR space, the problem of estimating the link function in a big-data setting now needs to be tackled. Non-parametric (or parametric) approaches should need to deal with a $n \times (K + 1)$ big matrix including the response variable and the K indices. The function `bigglm` from the **R** package `biglm` [37] offer the possibility to fit generalized linear model on a `bigmatrix` object. A smooth approach [49] has been recently implemented with the `bam` function from the `mcv` **R** package.

We adopt here the bin-summarise-smooth approach developed by Wickham [46] which can also been combine with `foreach` and `bigmemory` packages for computationally efficiency. The `bigvis` [48] **R** package condenses the large raw data to a summary on the same order of size as pixels on the screen by binning and summarising steps. Then, the smoothing step of the condensed data is fast and loses little statistical strength. As an illustration, the following synthetic **R** code presents a visualisation of the BIG-SIR outputs for data with $n = 10^7$ observations (occupying around 800 MB) from model (11):

```
R> result <-condense(bin(BIGsir[,2],0.1),
+     bin(BIGsir[,3],0.1), z = BIGsir[,1],
+     summary="mean")
R> autoplot(result)
```

where `BIGsir` is a `bigmatrix` object containing the response variable and the two estimated indices obtained from BIG-SIR approach. Figure 4 (middle) presents the average on the response variable (colour) as a function of the two indices. From this condensed object, we present in Figure 4 (top) a smooth estimation of the link function through a local quadratic regression method (loess) which can be compared to the true simulated model represented in Figure 1 (bottom). Note that we obtain a quality measure equal to 0.99 for this example. Regarding the model (10) with $n = 10^7$ observations the results of the estimation of the link function of model (10) by a kernel weighted local regression are presented in the bottom of Figure 4. Combination of the BIG-SIR approach and bin-summarise-smooth approach enables us to unravel the relationship between the response variable and the covariates (see true relation presented in the top of Figure 1).

5. REAL DATA ANALYSIS

We apply the proposed BIG-SIR approach on the airline on-time performance data from the 2009 ASA Data Expo (http://stat-computing.org/dataexpo). This data set

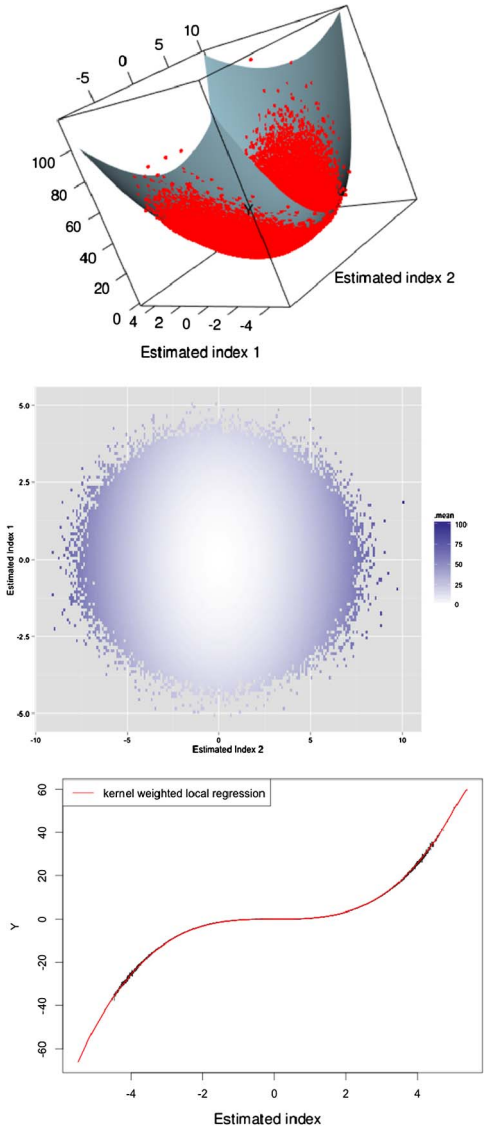

Figure 4. Model (11): Smooth estimate of the link function and scatterplot (on the top) of the estimated indices versus the mean of Y after condensing the results of BIG-SIR by using binning and summarizing approach (in the middle). Model (10): Estimation of the link function with a kernel weighted local regression and scatterplot of the estimated index versus the mean of Y after condensing the results of BIG-SIR by using binning and summarizing approach (at the bottom).

is used as an example to illustrate the estimation of the considered semiparametric regression model when dealing with a massive data set that exceeds the RAM of a single computer. The data are publicly available and have been used for demonstration with massive data by Kane et al. (2013) [28]. It consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. About 12 million flights were recorded with 29 variables. A compressed version of the pre-processed data set from the bigmemory project (http://data.jstatsoft.org/v55/i14/Airline.tar.bz2) is approximately 1.7 GB, and it takes 12 GB when uncompressed. This data set has been presented as a case study in [45] for performing a logistic regression to explain the late arrival given 4 explanatory variables: departure time (binary coding with 1 for night departure); week end (1 if departure occurred on weekends); departure hour and distance from origin to destination.

We propose to analyse using the model (1) the magnitude of delays (response variable) given the following 10 covariates: delay at departure, departure time (binary coding with 1 for night departure); week end (1 if departure occurred on weekends); departure hour (in minutes); arrival hour (in minutes); distance from origin to destination (miles); air time (in minutes); Taxi in time (minutes); Taxi out time (minutes) and age of the plane. A full description on the variables can be found in http://www.transtats.bts.gov/Fields.asp?Table_ID=236. After pre-processing the data set (using `bigmemory` and `foreach` packages) by removing missing values, negative values for air time variable and create the new variable "Age of the plane", the data set contains 84,183,043 observations for the response variable "Delay at arrival" and the 10 covariates. The dimension K of the EDR space E has been determined by considering a full SIR approach on a random sample of $n = 50{,}000$ observations. Figure 5 shows the scree plot of the eigenvalues associated to the estimated indices from the subset of observations. This plot suggests to choose a single index model ($K = 1$). In this illustration, we used $G = 45$ groups to compute our BIG-SIR estimator. The estimated EDR direction $\widehat{v}_G$ (corresponding to the coefficients of the linear combination of the covariates in the estimated EDR index) is equal to:

```
DepDelay     Night    Weekend     DepMin
-0.419       -0.532     0.426    -0.00206
ArrMin      Distance   AirTime    TaxiIn
-0.00111   0.0212      -0.169     -0.348
TaxiOut        Age
 -0.457     -0.000346
```

We used the same strategy as presented in section 4.2.3 to visualize and estimate the link function between "delay at arrival" and the estimated index (see Figure 6). This plot suggests a linear relationship between the response variable and the estimated index $X'\widehat{v}_G$. One can also visualize the scatterplot of "delay at arrival" versus the estimated index for a smaller random sample of size $n = 60{,}000$ observations for example. Since we observe a linear decreasing link function between "delay at arrival" and the estimated index, it is relevant to interpret the coefficients of the estimated EDR direction $\widehat{v}_G$ using their signs. For instance, as the "delay at departure" has a negative coefficient (-0.419), it means that an increase in "delay at departure" implies a decrease of

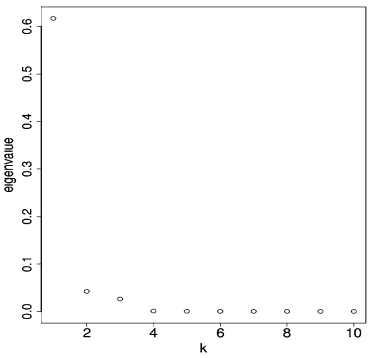

Figure 5. Eigenvalue scree plot of SIR model estimated on a random sample of $n = 50{,}000$ observations.

the estimated index and this then implies (not surprisingly) an increase of "delay at arrival". Similarly, "night departure" (negative coefficient) increases the "delay at arrival" while "week end departure" (positive coefficient) decreases the "delay at arrival". All other covariates (negative coefficients), except "Distance" (positive coefficient), impact negatively the "delay at arrival".

6. CONCLUDING REMARKS

In this paper, a new SIR estimator which we call, BIG-SIR, has been introduced for analyzing massive data sets through a semiparametric model. Exploiting the "divide and conquer" principle, BIG-SIR can handle the analysis of big data since SIR approach is a computationally very fast method in each chunk. The BIG-SIR estimator exhibits good theoretical properties. Its performance has been highlighted through a simulation study in various situations. Extensions to a multivariate indices model, multivariate dependent variable Y and SIR_α-based approach (for dependent model) have been described. The method has been implemented in **R** and the full code reproducing the results of this paper is available from the author. The implementation of BIG-SIR uses a combination of parallel strategy and shared-memory structures providing a solution for analyzing massive data sets which exceed the size of the available RAM. Moreover, the "divide and conquer" strategy of BIG-SIR offers the possibility to compute our estimator on several clusters using MapReduce programming [16] such as proposed in Hadoop software [20]. Note that **R** users can used `Rhipe` package [25] which offers the possibility to communicate directly with `Hadoop` from **R**.

In this article, BIG-SIR aims to tackle massive data sets with a very large number of observations. However the current BIG-SIR approach is not suitable for ultrahigh dimensional regression when the dimension p of the covariate is much larger than a very large sample size n. In this context, an open problem is to define a SIR estimator after dividing the data into subsets of covariates as the split-and-merge (SAM) method proposed in [43].

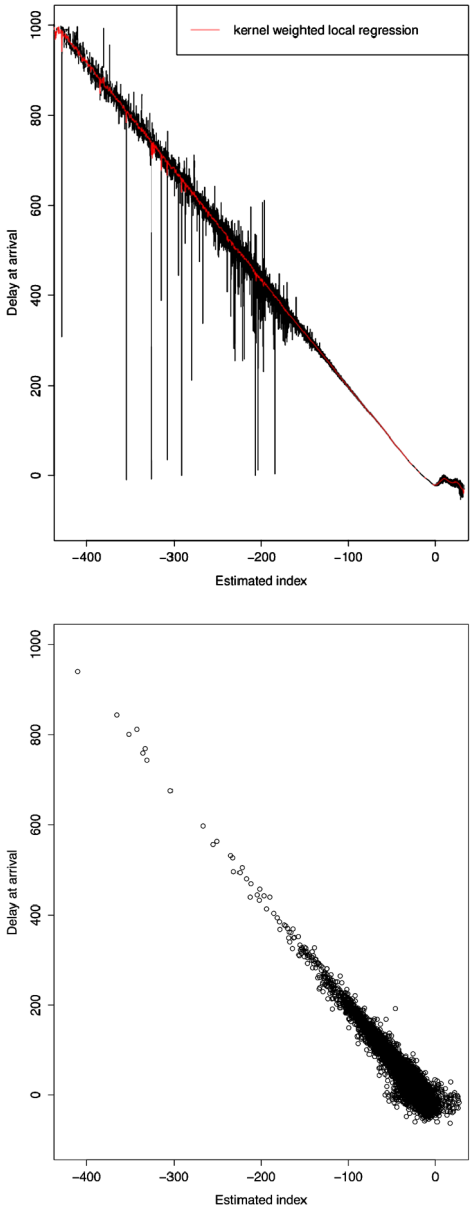

Figure 6. Model (10): Smooth estimate of the link function and scatterplot of the estimated index versus the mean of `delay at arrival` after condensing the results of BIG-SIR by using binning and summarizing approach (on the top). Visualization of the `delay at arrival` versus the index for a random sample of size $n = 60{,}000$ observations (at the bottom).

APPENDIX

In the following, we present the **R** code to compute the BIG-SIR estimator using a combination of `bigmemory` and `foreach` **R** packages.

```
R>  library("bigmemory")
R>  library("edrGraphicalTools")
```

```
R>  library("foreach")
R>  library("doSNOW")

## create a filebacking from data stored
## in a csv file

R>  dataYX <- read.big.matrix("myBIGdata.csv",
+    header=TRUE, backingfile="test-sir.bin",
+    descriptorfile="test-sir.desc", type="double")

R>  BIGmatdes <- describe(dataYX)


R>  cl <- makeCluster(4)
R>  registerDoSNOW(cl)

R> ng <- 10 # number of chunks

R>  x <- attach.big.matrix(BIGmatdes)

R>  matEDR.block <- function(x){
+    bhat <- matrix(x/sqrt((sum(x**2))),ncol=1)
+    bhat%*%t(bhat)}

R>  Scalable.sir <- function(g,data,size.chunk){
+    rows <- ((g-1)*size.chunk+1):(g*size.chunk)
+    matEDR.block(edr(data[rows,1],data[rows,-1],
+    H=8,K=1,method="SIR-I")$matEDR[,1])}

R>  size.chunk <- nrow(x)/ng

R> BIGsir <- foreach(g=1:ng, .combine="+")%dopar%{
+    require("edrGraphicalTools")
+    require("bigmemory")
+    x <- attach.big.matrix(BIGmatdes)
+    Scalable.sir(g,x,size.chunk)}

R> stopCluster(cl)
R> BIGsir.estimator <- eigen(BIGsir)$vectors[,1]
```

Received 26 April 2015

REFERENCES

[1] Analytics, R. and Weston, S. (2014). foreach: foreach looping construct for R, R package version 1.4.2.

[2] Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. *Computational Statistics* **12** 355–372. MR1477270

[3] Aragon, Y. and Saracco, J. (1997). Sliced Inverse Regression (SIR): an appraisal of small sample alternatives to slicing. *Computational Statistics* **12** 109–130. MR1435812

[4] Barreda, L., Gannoun, A. and Saracco, J. (2007). Some extensions of multivariate SIR. *Journal of Statistical Computation and Simulation* **77** 1–17. MR2343408

[5] Barrios, M. P. and Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statistics & Probability Letters* **77** 247–255. MR2339028

[6] Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **63** 393–410. MR1841422

[7] Carroll, R. J. and Li, K. C. (1992). Measurement error regression with unknown link: dimension reduction and data visualization. *Journal of the American Statistical Association* **87** 1040–1050. MR1209565

[8] Chavent, M., Girard, S., Kuentz-Simonet, V., Liquet, B., Nguyen, T. and Saracco, J. (2014). A sliced inverse regression approach for data stream. *Computational Statistics* **29** 1129–1152. MR3266051

[9] Chavent, M., Kuentz, V., Liquet, B. and Saracco, J. (2011). A sliced inverse regression approach for a stratified population. *Communications in Statistics – Theory and Methods* **40** 3857–3878. MR2864124

[10] Chen, C. H. and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica* **8** 289–316. MR1624402

[11] Chen, X. and Xie, M. G. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **24** 1655–1684. MR3308656

[12] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86** 328–332. MR1137117

[13] Cook, R. D. (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in statistics – Theory and Methods* **29** 2109–2121.

[14] Coudret, R., Girard, S. and Saracco, J. (2014). A new sliced inverse regression method for multivariate response. *Computational Statistics & Data Analysis* **77** 285–299. MR3210063

[15] Coudret, R., Liquet, B. and Saracco, J. (2013). edrGraphicalTools: provides tools for dimension reduction methods, R package version 2.1.

[16] Dean, J. and Ghemawat, S. N. (2008). MapReduce: simplified data processing on large cluster. *Commun. ACM* 107–113.

[17] Duan, N. and Li, K. C. (1991). Slicing regression: a link-free regression method. *Annals of Statistics* **19** 505–530. MR1105834

[18] Emerson, J. W. and Kane, M. J. (2012). Don't drown in the data. *Significance* **9** 38–39.

[19] Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association* **93** 132–140. MR1614604

[20] Apache Software Foundation (2015). Hadoop. https://hadoop.apache.org.

[21] Gannoun, A. and Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica* **13** 297–310. MR1977727

[22] Gannoun, A. and Saracco, J. (2003). Two cross validation criteria for SIR_α and $PSIR_\alpha$ methods in view of prediction. *Computational Statistics* **4** 297–310. MR1977727

[23] Gather, U., Hilker, T. and Becker, C. (2002). A note on outlier sensitivity of sliced inverse regression. *Statistics* **36** 271–281. MR1923466

[24] Girard, S. and Saracco, J. (2014). An introduction to dimension reduction in nonparametric kernel regression. *EAS Publications Series* **66** 167–196.

[25] Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J., Xi, B. and Cleveland, W. S. (2012). Large complex data: divide and recombine (D&R) with RHIPE. *Stat* **1** 53–67.

[26] Hsing, T. (1999). Nearest neighbor inverse regression. *Annals of Statistics* **27** 697–731. MR1714711

[27] Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Annals of Statistics* **20** 1040–1061. MR1165605

[28] Kane, M. J., Emerson, J. and Weston, S. (2013). Scalable strategies for computing with massive data. *Journal of Statistical Software* **55** 1–19.

[29] Kötter, T. T. (1996). An asymptotic result for Sliced Inverse Regression. *Computational Statistics* **11** 113–136. MR1394544

[30] Kötter, T. T. (2000). *Smoothing and regression. Approaches, computation and application.* Sliced Inverse Regression, 497–512. Wiley, Chichester. MR1795148

[31] Li, K. C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association* **86** 316–342. MR1137117

[32] Li, K. C., Aragon, Y., Shedden, K. and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association* **98** 99–109. MR1965677

[33] Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and Its Interface* **4** 73–83. MR2775250

[34] Liquet, B. and Saracco, J. (2007). Pooled marginal slicing approach via SIR_α with discrete covariables. *Computational Statistics* **4** 599–617. MR2358429

[35] Liquet, B. and Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics – Simulation and Computation* **37** 1198–1218. MR2528270

[36] Lue, H. H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference* **139** 2656–2664. MR2523656

[37] Lumley, T. (2013). biglm: bounded memory linear and generalized linear models, R package version 0.9-1.

[38] Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics – Theory and Methods* **26** 2141–2171. MR1484246

[39] Saracco, J. (1999). Sliced inverse regression under linear constraints. *Communications in Statistics – Theory and Methods* **28** 2367–2393. MR1720535

[40] Saracco, J. (2001). Pooled slicing methods versus slicing methods. *Communications in Statistics – Simulation and Computation* **30** 489–511. MR1869125

[41] Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis* **96** 117–135. MR2202403

[42] Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association* **89** 141–148. MR1266291

[43] Song, Q. and Liang, F. (2014). A Split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 947–972. MR3414135

[44] R Core Team (2013). R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

[45] Wang, C., Chen, M.-H., Schifano, E., Wu, J. and Yan, J. (2015). A survey of statistical methods and computing for big data. *arXiv:1502.07989*.

[46] Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software* **40** 1–29.

[47] Wickham, H. (2013). Bin-summarise-smooth: a framework for visualising large data, Technical Report, had.co.nz.

[48] Wickham, H. and Hue, Y. bigvis: tools for visualisation of big data sets, R package version 0.1.

[49] Wood, S. N., Goude, Y. and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64** 139–155. MR3293922

[50] Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica* **5** 727–736. MR1347616

Benoit Liquet
Laboratoire de Mathématiques et de leurs Applications
Université de Pau et des Pays de l'Adour
UMR CNRS 5142
Pau
France
E-mail address: benoit.liquet@univ-pau.fr

ARC Centre of Excellence for Mathematical
and Statistical Frontierss
Queensland University of Technology (QUT)
Brisbane
Australia

Jerome Saracco
Bordeaux INP
IMB, UMR CNRS 5251
Inria Bordeaux Sud Ouest, CQFD team
351, cours de la Liberation
Talence
France
E-mail address: jerome.saracco@math.u-bordeaux1.fr