Collaborative Sliced Inverse Regression

Alessandro Chiancone, Stephane Girard, Jocelyn Chanussot

Collaborative Sliced Inverse Regression

Alessandro Chiancone[1,2,3], Stéphane Girard[1], Jocelyn Chanussot[2]

[1] Inria Grenoble Rhône-Alpes & LJK, team Mistis, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.
[2] GIPSA-Lab, Grenoble INP, Saint Martin d'Hères, France.
[3] Institute of Statistics, Graz University of Technology, Kopernikusgasse 24/III, A-8010 Graz, Austria.

Abstract

Sliced Inverse Regression (SIR) is an effective method for dimensionality reduction in high-dimensional regression problems. However, the method has requirements on the distribution of the predictors that are hard to check since they depend on unobserved variables. It has been shown that, if the distribution of the predictors is elliptical, then these requirements are satisfied. In case of mixture models, the ellipticity is violated and in addition there is no assurance of a single underlying regression model among the different components. Our approach clusterizes the predictors space to force the condition to hold on each cluster and includes a merging technique to look for different underlying models in the data. A study on simulated data as well as two real applications are provided. It appears that SIR, unsurprisingly, is not capable of dealing with a mixture of Gaussians involving different underlying models whereas our approach is able to correctly investigate the mixture.

Keywords: Mixture models, inverse regression, sufficient dimension regression.

1 Introduction

In multidimensional data analysis, one has to deal with a dataset made of n points in dimension p. When p is large, classical statistical analysis methods

and models fail. Supervised and unsupervised dimensionality reduction (d.r.) techniques are widely used to preprocess high dimensional data retaining the information useful to solve the original problem. Recently, more and more investigations aim at developing non-linear unsupervised techniques to better adapt to the complexity of our, often non-linear, world. A review is provided in [28] concluding that even if the variety of non-linear methods is huge, Principal Component Analysis (PCA) [23], despite its intrinsic limitations, is still one of the best choices. PCA is not the best in specific cases (i.e. when additional information on the structure of the data are available) but, as expected, is rather general and can be easily controlled and applied. What about the case of supervised d.r.? In unsupervised d.r. one is interested in preserving all the information getting rid of the redundancies in the data. In other words, the goal is to catch the intrinsic dimensionality of the data, which is the minimum number of parameters needed to describe it [1, 15]. In supervised d.r., a response variable Y is given and the analysis aims at providing a prediction (classification, when Y is categorical, or regression, when Y is continuous). There is some additional information encoded in Y of what we want to select in the data. Estimating the intrinsic dimensionality is no more our goal since we are oriented by the information present in Y.

The regression framework is characterized by the assumption of a link function between X and Y i.e. $Y = f(X, \epsilon)$, where ϵ is a random noise. In this environment, it can be assumed that only a portion of X is needed to correctly explain Y. This is a reasonable assumption since data nowadays are rarely tailored on the application and filled by too many details. If Y depends on the multivariate predictor through an unknown number of linear projections $Y = f(X^T\beta_1, \ldots, X^T\beta_k, \epsilon)$, the effective dimension reduction (e.d.r.) space is what we are looking for [19]. It is defined as the smallest linear space containing the information needed to correctly regress the function f. Under the previous assumption, the e.d.r. space is spanned by $\beta_1, \ldots, \beta_k$. Sliced Inverse Regression (SIR) [19] has proven to achieve good results retrieving a basis of the e.d.r. space. Recently, many papers focused on the complex structure of real data showing that often the data is organized in subspaces, see [6, 18] or [27] for a detailed discussion and references.

Here, our hypothesis is that the e.d.r. space is not unique all over the data but varies through the components. We introduce a novel technique to identify the number of e.d.r. spaces based on a weighted distance. With this paper we try to give an answer to the question: Can SIR be as popular as multiple linear regression? [9]. In Section 2, we briefly describe SIR and

provide a discussion on the limitations of the method. The following Section 3 is the core of our paper where our contribution, Collaborative SIR, is introduced. The motivations and the main problem are described, some asymptotic results are established under mild conditions. The simulation study, Section 4, is where the performances of Collaborative SIR are shown and analyzed on simulated specific test cases. The stability of the results is detailed and commented. In Section 5, two real data applications are reported showing the interest of this technique. A discussion and a conclusion are finally drawn encouraging the community to improve our idea.

2 Sliced Inverse Regression (SIR)

2.1 Method

Back in 1991, SIR is introduced in [19] as a *data-analytic tool*: even if the performance of computers and the capability to explore huge dataset increased tremendously, SIR remains a useful tool for d.r. in the framework of regression. The visualization of high dimensional datasets is nowadays of extreme importance because human beings are still, unfortunately, limited by a perception which only allows us to display three dimensions at a time while the capability to gather data is amazingly increasing. When p is large, a possible approach is to suppose that *interesting features of high-dimensional data are retrievable from low-dimensional projections*. In other words, the model proposed by Li is:

$$Y = f(X^T\beta_1, \ldots, X^T\beta_k, \epsilon) \tag{1}$$

where $Y \in \mathbb{R}$ is the response variable, X is a p-dimensional random vector with finite expectation, $\mu = \mathbb{E}(X)$, and finite covariance matrix, $\Sigma = \mathrm{Cov}(X)$. The noise ϵ is a random variable supposed to be independent of X. If $k \ll p$ the function f depends on k linear combinations of the original predictors and the d.r. is achieved. The goal of SIR is to retrieve a basis of the e.d.r. space. Under the Linearity Design Condition:

(LDC) $\mathbb{E}(X^T b|X^T\beta_1, \ldots, X^T\beta_k)$ is linear in $X^T\beta_1, \ldots, X^T\beta_k$ for any $b \in \mathbb{R}^p$,

it has been shown [14] that the centered inverse regression curve is contained in the k-dimensional linear subspace of $\mathbb{R}^p$ spanned by $\Sigma\beta_1, \ldots, \Sigma\beta_k$. If we consider a monotone transformation T of Y, the matrix $\Sigma^{-1}\Gamma$ is degenerated in any direction orthogonal to $\beta_1, \ldots, \beta_k$, where $\Gamma = \mathrm{Cov}(\mathbb{E}(X|T(Y)))$.

Therefore, the k eigenvectors corresponding to the k non zero eigenvalues form a basis of the e.d.r. space. To estimate Γ, [19] used a slicing procedure as candidate for T. Dividing the range of Y in $H > 1$ non-overlapping slices, $s^1, \ldots, s^H$, Γ can then be written as:

$$\Gamma = \sum_{h=1}^{H} p^h (m^h - \mu)(m^h - \mu)^T,$$

where $p^h = \mathbb{P}(Y \in s^h)$ and $m^h = \mathbb{E}(X|Y \in s^h)$. The estimator $\hat{\Gamma}$ can then be defined by substituting p^h, m^h with the corresponding sample versions. The range of Y can be divided either by fixing the width or the proportion of samples p^h in each slice. Through the paper we adopted the second slicing strategy [9]. The k eigenvectors corresponding to the largest eigenvalues of $\hat{\Sigma}^{-1}\hat{\Gamma}$ are then an estimation of a basis of the e.d.r. space, where $\hat{\Sigma}$ is the usual sample covariance matrix.

2.2 Limitations

SIR theory is well established and comes fully equipped by asymptotic results [17, 24]. However, two main limitations affect the practical use of the method:

- The inversion of the estimated covariance matrix $\hat{\Sigma}$;
- The impossibility to check if the (LDC) holds.

When $n \leq p$, the sample covariance matrix is singular, and when the variables are highly correlated (e.g. in hyperspectral images) the covariance matrix is ill-conditioned [4]. To compute the e.d.r. directions, the inversion of $\hat{\Sigma}$ must nevertheless be achieved. Recently, many papers faced this problem and provided solutions [10, 21, 25, 26, 29]. An homogeneous framework to perform regularized SIR has also been proposed in [5] where, depending on the choice of the prior covariance matrix, the above mentioned techniques can be obtained and extended.

The (LDC), less studied in the literature, is yet the central assumption of the theory and it depends on the unobserved e.d.r. directions, therefore it cannot be directly checked [30]. It can be yet proved that, if X is elliptical distributed, the condition holds. This ellipticity condition is much stronger than (LDC) but easier to verify in practice since it does not depend on

$\beta_1, \ldots, \beta_k$. Good hope comes from a result of [16] that shows that, when the dimension p tends to infinity, the measure of the set of directions for which the (LDC) does not hold tends to zero. The condition becomes weaker and weaker as soon as the dimension increases. The intuition comes from [13] where the authors show that high dimensional dataset are nearly Gaussian in most low dimensional projections. Since the (LDC) condition holds for elliptical distributions, it is desirable to work in the direction that allows us to use this property. Unfortunately, when X follows a mixture of elliptical distributions, this property is not globally verified: [18], starting from an idea of [20], proposed to clusterize the space to look locally for ellipticity rather than globally while [8] introduced categorical predictors to distinguish different populations. This is our very start, assuming X distributed from a mixture model, we focus on decomposing the mixture and we extend the basic model to improve SIR capability to explore complex datasets.

3 Collaborative SIR

First, we motivate and introduce in Subsection 3.1 the population version of Collaborative SIR. Second, the different steps of the sample version are detailed and an algorithm is outlined in Subsections 3.2–3.5. For sake of simplicity, we focus on the case when $k = 1$ i.e. the effective dimension reduction space is one-dimensional. Some extensions to $k > 1$ are discussed in Subsection 5.3.

3.1 Population version

Recall that the underlying model of SIR through the whole predictors space is $Y = f(\beta^T X, \epsilon)$. When dealing with complex data, one could allow the underlying model to change depending on the predictor space. Mixture models provide a good framework to deal with such an hypothesis, considering the data as realizations from a weighted sum of distributions with different parameters. As mentioned before, in such a case, there is no straightforward way to check if the (LDC) holds. Let X be a p-dimensional random vector from a mixture model and be Z an unobserved latent random variable $Z \in \{1, \ldots, c\}$, where c is the number of components. Given $Z = i$, we have the following model:

$$Y = f_{F(i)}(\beta_{F(i)}^T X) + \epsilon_i, \tag{2}$$

where Y is the random variable to predict, F is an unknown deterministic function $F : \{1, \ldots, c\} \to \{1, \ldots, D\}$, $D \in \mathbb{N}\backslash\{0\}$. The functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, D$ are unknown link functions between $\beta_j^T X$ and Y. Finally, ϵ_i are random errors, $i = 1, \ldots, c$, i.e. each component is allowed to have a different related error.

Under model (2), D is the number of different e.d.r. spaces. The goal is to find a basis of the D one-dimensional spaces spanned by $\beta_1, \ldots, \beta_D$. The number D ($D \leq c$) of e.d.r. spaces is unknown and the link function may change depending on the component. Function F selects the underlying model for the specific component. It is assumed that the (LDC) holds in each component:

(LDC) $\forall i = 1, \ldots, c,\ \ \mathbb{E}(X^T b | X^T \beta_{F(i)}, Z = i)$ is linear in $X^T \beta_{F(i)}$ for any b.

Given $Z = i$, we define the mean $\mu_i = \mathbb{E}(X|Z = i)$, the covariance matrices $\Sigma_i = Cov(X|Z = i)$ and $\Gamma_i = Cov(\mathbb{E}(X|Y, Z = i))$. Hence, the eigenvector b_i corresponding to the highest eigenvalue of $\Sigma_i^{-1}\Gamma_i$, is a basis of the e.d.r. space: $Span\{b_i\} = Span\{\beta_{F(i)}\}$ from SIR theory [19]. If $F : \{1, \ldots, c\} \to \{1, \ldots, D\}$ is known, then the inverse image of the elements $j \in \{1, \ldots, D\}$ can be defined:

$$F^{-1}(j) = \{i \in \{1, \ldots, c\} \text{ such that } F(i) = j\}.$$

Since F is not supposed to be injective, an e.d.r. direction β_i may be associated with several components. Suppose that the set $\{b_i, i \in F^{-1}(j)\}$ is observed. Given the proximity criteria between two unit vectors a and b:

$$m(a, b) := \cos^2(a, b) = (a^T b)^2, \tag{3}$$

the "most collinear vector" to the set of directions $\{b_i, i \in F^{-1}(j)\}$ is the solution of the following problem:

$$\begin{aligned}
\max_{v \in \mathbb{R}^p, \|v\|=1} \sum_{i \in F^{-1}(j)} m(v, b_i) &= \max_{v \in \mathbb{R}^p, \|v\|=1} \sum_{i \in F^{-1}(j)} (v^T b_i)^2 \\
&= \max_{v \in \mathbb{R}^p, \|v\|=1} v^T \Big(\sum_{i \in F^{-1}(j)} b_i b_i^T \Big) v \\
&= \max_{v \in \mathbb{R}^p, \|v\|=1} v^T (B_j B_j^T) v,
\end{aligned}$$

where B_j is the $p \times |F^{-1}(j)|$ matrix defined by $B_j := [b_{i, i \in F^{-1}(j)}]$. Using Lagrange multipliers, it is easily shown that v is the eigenvector of the matrix

$B_j B_j^T$ associated with the largest eigenvalue. The following lemma motivates this argument.

Lemma 1. *Assuming the (LCD) and model* (2)*, for all* $j = 1, \dots, D$*, the eigenvector* $\tilde{\beta}_j$ *associated to the unique non-zero eigenvalue of the matrix* $B_j B_j^T$ *is collinear with* β_j*.*

Proof. For each $j = 1, \dots, D$ and $i \in F^{-1}(j)$, b_i is collinear with β_j, that is $b_i = \alpha_i \beta_{F(i)}$, $\alpha_i \in \mathbb{R} \setminus \{0\}$. Since $B_j = [\alpha_i \beta_i, i \in F^{-1}(j)]$, we have:

$$B_j B_j^T = \sum_{i \in F^{-1}(j)} \alpha_i^2 \beta_j \beta_j^T = \|\alpha\|^2 \beta_j \beta_j^T$$

which concludes the proof. □

This lemma shows that $\tilde{\beta}_j$ is an e.d.r. direction for each $j = 1, \dots, D$ and the previous argument gives a strategy to estimate the directions β_j based on the proximity criteria (3).

Remark. If $D = 1$ then $F^{-1}(1) = \{1, \dots, c\}$, the e.d.r. directions and the link functions do not vary through all the mixture. This particular case is addressed in [18].

3.2 Sample version: Z is observed, F and D known

Let $\{Y_1, \dots, Y_n\}$ be a sample from Y, $\{X_1, \dots, X_n\}$ a sample from X, $\{Z_1, \dots, Z_n\}$ a sample from Z. We suppose Z observed at this stage. Let $\mathcal{C}_i = \{t \text{ such that } Z_t = i\}$ and $n_i = card(\mathcal{C}_i)$, $i = 1, \dots, c$. We can now estimate for each $\mathcal{C}_i$ the mean and covariance matrix:

$$\bar{X}_i = \frac{1}{n_i} \sum_{t \in \mathcal{C}_i} X_t, \quad \hat{\Sigma}_i = \frac{1}{n_i} \sum_{t \in \mathcal{C}_i} (X_t - \bar{X}_i)(X_t - \bar{X}_i)^T,$$

for each $i = 1, \dots, c$. To obtain an estimator for Γ_i, we introduce as in classical SIR a slicing. For each $\mathcal{C}_i$, we can define the slicing T_i of Y_i into $H_i \in \mathbb{N} \setminus \{0\}$ slices, $i = 1, \dots, c$. Let $s_i^1, \dots, s_i^{H_i}$ be the slicing associated to $\mathcal{C}_i$, $\Gamma_i = Cov(\mathbb{E}(X|Y, Z = i))$ can be written as:

$$\Gamma_i = \sum_{h=1}^{H_i} p_i^h (m_i^h - \mu_i)(m_i^h - \mu_i)^T,$$

where $p_i^h = \mathbb{P}(Y \in s_i^h | Z = i)$, $m_i^h = \mathbb{E}(X|Z = i, Y \in s_i^h)$. Let us recall that $\mu_i = \mathbb{E}(X|Z = i)$ and $\Sigma_i = Cov(X|Z = i)$, as defined in Subsection 3.1. Replacing p_i^h, m_i^h with the corresponding sample versions, it is possible to estimate Γ_i as follows

$$\begin{aligned}
\hat{\Gamma}_i &= \sum_{h=1}^{H_j} \hat{p}_i^h (\hat{m}_i^h - \bar{X}_i)(\hat{m}_i^h - \bar{X}_i)^T, \\
\hat{m}_i^h &= \frac{1}{n_{h,i}} \sum_{t \in \mathcal{C}_i} X_t \mathbb{I}[Y_t \in s_t^h], \\
\hat{p}_i^h &= \frac{n_{h,i}}{n_i}, \\
n_{h,i} &= \sum_{t \in \mathcal{C}_i} \mathbb{I}[Y_t \in s_t^h],
\end{aligned}$$

where $\mathbb{I}$ is the indicator function. The estimated e.d.r. directions are then $\hat{b}_1, \ldots, \hat{b}_c$ where $\hat{b}_i$ is the major eigenvector of the matrix $\hat{\Sigma}_i^{-1}\hat{\Gamma}_i$. This allows us to estimate B_j and β_j as $\hat{B}_j = [\hat{b}_{i, i \in F^{-1}(j)}]$ and with $\hat{\beta}_j$ the major eigenvector of $\hat{B}_j \hat{B}_j^T$, $j = 1, \ldots, D$.

Asymptotic results can be established similarly to [7]. We fix $j \in \{1, \ldots, D\}$ and consider the sample $\{X_t,\, t \in \bigcup_{i \in F^{-1}(j)} \mathcal{C}_i\}$ of size $n^j = \sum_{i \in F^{-1}(j)} n_i$. The following three assumptions are introduced:

- (A1) $\{X_t,\ t \in \bigcup_{i \in F^{-1}(j)} \mathcal{C}_i\}$ is a sample of independent observations from the single index model (2).
- (A2) For each $i = 1, \ldots, c$, the support of $\{Y_t, t \in \mathcal{C}_i\}$ is partitioned into a fixed number H_i of slices such that $p_i^h > 0, h = 1, \ldots, H_i$.
- (A3) For each $i = 1, \ldots, c$ and $h = 1, \ldots, H_i$, $n_{h,i} \to \infty$ as $n \to \infty$.

Let us highlight that (A3) implies $n_i \to \infty$ for all $i = 1, \ldots, c$ and therefore $n^j \to \infty$.

Theorem 1. *Under model (2), linearity condition (LDC) and assumptions (A1)-(A3), we have:*

(i) $\widehat{\beta}_j = \beta_j + O_p(1/\sqrt{\underline{n}^j})$, *where* $\underline{n}^j = \min_{i \in F^{-1}(j)} n_i$;

(*ii*) *If, in addition* $n_i = \theta_{ij} n^j$, $\theta_{ij} \in (0,1)$ *for each* $i \in F^{-1}(j)$, *then* $\sqrt{n^j}(\widehat{\beta}_j - \beta_j)$ *converges to a centered Gaussian distribution.*

Proof. (*i*) For each $i \in F^{-1}(j)$ and under the assumptions (LC), (A1)-(A3), from the SIR theory [19], each estimated EDR direction $\widehat{b}_i$ converges to β_j at root $\underline{n}^j$ rate: that is, for $i \in F^{-1}(j)$, $\widehat{b}_i = \beta_j + O_p(1/\sqrt{\underline{n}^j})$. We then have $\widehat{B}_j \widehat{B}_j^T = B_j B_j^T + O_p(1/\sqrt{\underline{n}^j})$. Therefore, the principal eigenvector of $\widehat{B}_j \widehat{B}_j^T$ converges to the corresponding one of $B_j B_j^T$ at the same rate: $\hat{\beta}_j = \beta_j + O_p(1/\sqrt{\underline{n}^j})$. The estimated e.d.r. direction $\hat{\beta}_j$ thus converges to an e.d.r. direction at root $\underline{n}^j$ rate.

(*ii*) The proof is similar to the one of [7], Theorem 2. □

In the following paragraphs, a merging algorithm is introduced to infer the number D based on the collinearity of the vectors b_i and a procedure is given to estimate the function F.

3.3 Sample version: D unknown, Z is observed and F known

We assumed, so far, D known. To estimate D, a hierarchical merging procedure is introduced based on the proximity measure (3) between the estimated e.d.r. directions $\hat{b}_1, \ldots, \hat{b}_c$. Let us note that a similar procedure has been used in [12], Subsection 3.6 to cluster the components of the multivariate response variable Y related to the same e.d.r. spaces.

Let $V = \{v_1, \ldots, v_{|V|}\}$ be a set of vectors in dimension p with associated set of weights $\Omega = \{w_1, \ldots, w_{|V|}\}$. We define the quantity $\lambda(V)$ as

$$\lambda(V) = \max_{v \in \mathbb{R}^p} \frac{1}{w_V} \sum_{i=1}^{|V|} w_i m(v_i, v) \text{ such that } \|v\| = 1,$$

where $w_V = \sum_{i=1}^{|V|} w_i$ is a normalizing constant. From an intuitive point of view, the vector v maximizing $\lambda(V)$ is the most collinear vector to our set of vectors V given the proximity criteria (3) and the set of weights Ω. It is easily seen that $\lambda(V)$ is the largest eigenvalue of the matrix $\frac{1}{w_V} \sum_{i=1}^{|V|} w_i v_i v_i^T$.

In practice, to build the hierarchy, we consider the following iterative algorithm initialized with the set $A = \{\{\hat{b}_1\}, \ldots, \{\hat{b}_c\}\}$:

`While` $card(A) \neq 1$,

- `Let` $a, b \in A$ `such that` $\lambda(a \cup b) > \lambda(c \cup d)$ `for all` $c, d \in A$,
- `Set` $A := (A \setminus \{a, b\}) \bigcup a \cup b$,

`end.`

In our applications, the weights are set equal to the number of samples in each components, i.e. $w_i = n_i$, $i = 1, \ldots, c$. At each step the cardinality of the set A decreases merging the most collinear sets of directions, see Figure 1 for an illustration. The bottom up greedy algorithm proceeds as follows:

- First, the two most similar elements of A are merged considering all the $|A| \times (|A| - 1) = c \times (c - 1)$ pairs. For instance, $\hat{b}_1$ and $\hat{b}_2$ are selected to be merged in Figure 1.
- In the following steps, the two most similar sets of vectors are merged, considering all $|A| \times (|A| - 1)$ pairs in A. E.g., in the second step, we have $A = \{\{\hat{b}_1, \hat{b}_2\}, \{\hat{b}_3\}, \ldots, \{b_{12}\}\}$ in Figure 1.

Therefore, it is possible to infer the number D of underlying e.d.r. spaces analyzing the values of λ in the hierarchy (see Figure 1 for an example) looking for a discontinuity that will occur when two sets with different underlying β_j (i.e. non collinear) are merged. We automatically estimate D with the following procedure:

- Draw a line from the first value of the graph $(1, \lambda_1)$ to the last (c, λ_c).
- Compute the distance between points in the graph and the line.
- Select the merging point maximizing that distance. $\hat{D} := c-$number of merge selected.

Once achieved an estimation of D, denoted by $\hat{D}$, the function F can be estimated. Even if we used an automatic procedure, a visual selection of $\hat{D}$ depending on the task and prior knowledge is strongly recommended.

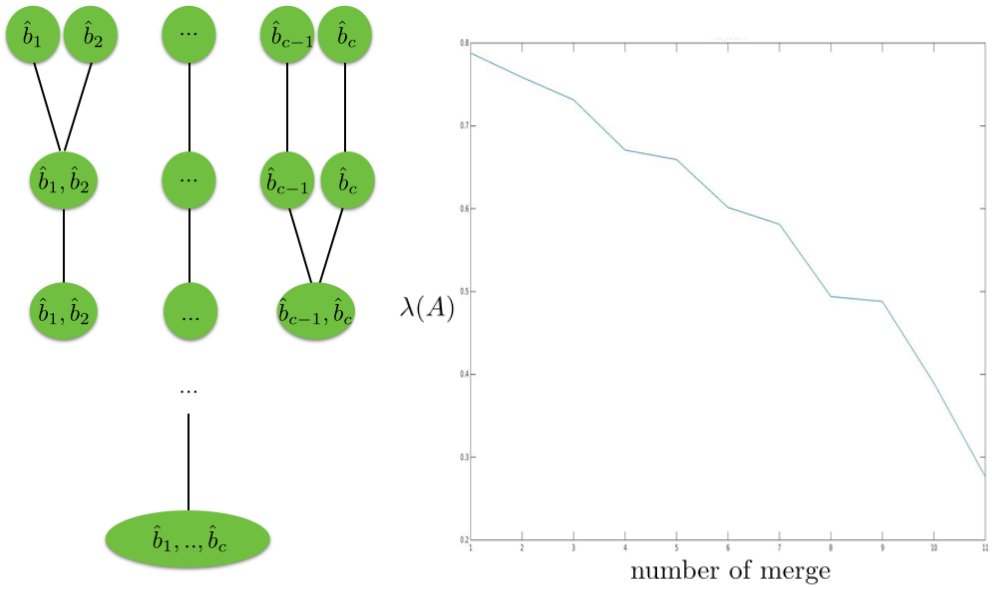

Figure 1: Left: Hierarchy built following the proximity criteria (3). Right: Cost function $\lambda(A)$, the number D of unknown e.d.r. directions decreases at each step by one. $\hat{D} := c-$number of merge selected. In the above example $c = 12$. The algorithm selects merge step 9 which corresponds to the correct estimation of the parameter: $\hat{D} = 3$.

3.4 Sample version: F unknown

For each node of the tree at level $\hat{D}$, the "most collinear direction", using (3), is computed. Solving the related $\hat{D}$ diagonalization problems gives $\hat{\beta}_1, \ldots, \hat{\beta}_{\hat{D}}$. In this paragraph, a procedure for the estimation $\hat{F}$ of the function F is detailed.

Once the candidates $\hat{\beta}_1, \ldots, \hat{\beta}_{\hat{D}}$ are estimated, the whole dataset is considered to estimate F. Starting from $i \in \{1, \ldots, \hat{c}\}$, the goal is to find $j \in \{1, \ldots, \hat{D}\}$ such that $F(i) = j$, under certain conditions. The $\hat{D}$ covariance matrices of the pairs $(X_t^T \hat{\beta}_j, Y_t)$, $t \in \mathcal{C}_i$, $j \in \{1, \ldots, \hat{D}\}$ are considered. The idea is to select the direction that best explains Y_t, $t \in \mathcal{C}_i$ among the estimated directions $\hat{\beta}_1, \ldots, \hat{\beta}_{\hat{D}}$. Let us assume the f_j functions to be "locally" linear

- (A4): For all $j = 1, \ldots, D$, f_j can be approximated by a piecewise linear function so that $f_j(X_t^T \beta_j) = k_i X_t^T \beta_j$, $\forall t \in \mathcal{C}_i$, $i \in F^{-1}(j)$.

Lemma 2. *Let $j \in \{1, \ldots, D\}$. Under assumption (A4), the e.d.r. direction β_j is the vector minimizing the second eigenvalue of the covariance matrix of the pairs $(X_t^T \beta_s, Y_t)$, $s = 1, \ldots, D$, $t \in \mathcal{C}_i$, $i \in F^{-1}(j)$.*

Proof. We have

$$\begin{aligned} cov(X_t^T \beta_s, Y_t) &= cov(X_t^T \beta_s, k_i X_t^T \beta_j) = \begin{pmatrix} \beta_s^T \Sigma \beta_s & k_i \beta_s^T \Sigma \beta_j \\ k_i \beta_s^T \Sigma \beta_j & k_i^2 \beta_j^T \Sigma \beta_j \end{pmatrix} \\ &= \begin{pmatrix} \langle \beta_s, \beta_s \rangle & k_i \langle \beta_s, \beta_j \rangle \\ k_i \langle \beta_s, \beta_j \rangle & k_i^2 \langle \beta_j, \beta_j \rangle \end{pmatrix} = \begin{pmatrix} \|\beta_s\|^2 & k_i \langle \beta_s, \beta_j \rangle \\ k_i \langle \beta_s, \beta_j \rangle & k_i^2 \|\beta_j\|^2 \end{pmatrix} \end{aligned}$$

where the scalar product and the norm are induced by Σ. The characteristic polynomial is given by

$$p(\lambda) = \lambda^2 - \lambda(\|\beta_s\|^2 + k_j^2 \|\beta_j\|^2) + k_j^2 (\|\beta_s\|^2 \|\beta_j\|^2 - \langle \beta_s, \beta_j \rangle^2),$$

and has two real roots $\lambda_1 \geq \lambda_2$. Cauchy-Schwarz inequality yields $\|\beta_s\|^2 \|\beta_j\|^2 - \langle \beta_s, \beta_j \rangle^2 \geq 0$ and thus $\lambda_1 \geq \lambda_2 \geq 0$. Moreover, $\lambda_2 = 0$ if and only if the equality holds. Since the β_s, $s = 1, \ldots, D$ are linearly independent, it follows that $\lambda_2 = 0$ if and only if $s = j$. □

In practice, for all fixed $i \in \{1, \ldots, \hat{c}\}$, the vectors $\hat{\beta}_j$, $j = 1, \ldots, \hat{D}$ are the candidates for (X_t, Y_t), $t \in \mathcal{C}_i$. Lemma 2 is stating that, under assumption

(A4), the vector $\hat{\beta}_j$ minimizing the second eigenvalue of the covariance matrix of the pairs $(X_t^T\hat{\beta}_s, Y_t)$, $s = 1, \ldots, D$, $t \in \mathcal{C}_i$ is such that $j = F(i)$. Here, the link functions are supposed to be locally linear. If the functions are approximately linear the estimation will work, but, in case of dramatic non linearities the method may lead to unreasonable results. A possibility is to resize the interval where we want to regress the functions and zoom until we find a reasonable local behavior of the functions. It must be noted that, in case where D is overestimated $\hat{D} > D$ (e.g. due to instabilities in the estimation of the direction in some components), we observed on simulated data that the estimation of F mitigates this error often avoiding to select the aberrant directions β_j, $j > D$.

3.5 Estimation of Z via clustering

To estimate the latent variable Z the explanatory space X is partitioned using a k-means algorithm. It is worth noticing that we decided to use k-means for simplicity and also to compare our results with [18], see Section 5. Several other clustering procedures can be adopted, for instance High Dimensional Data Clustering (HDDC) [3], a clustering method designed for high dimension. Twenty initial random centroids are chosen as initialization of k-means, the one minimizing the sum of squares is retained.

4 Simulation study

We performed a study on simulated data, in order to test in a controlled setting and evidence the weaknesses and strengths of the method. Two aspects are of interest:

(A) Study the sensitivity to clustering (estimation of Z).

(B) Analyze the quality of the estimation of the e.d.r. space compared to SIR performed independently in each cluster.

The first experiment is performed on a same dataset to investigate the effect of different initializations of k-means and how the quality of clustering affects the result. In the second experiment, different simulated datasets are analyzed to test the method under a variety of conditions.

4.1 Test case A

To study the sensitivity to clustering, $n = 2500$ samples are drawn from a Gaussian mixture model with uniform mixing proportions and $c = 10$ components. Each component follows a Gaussian distribution $\mathcal{N}(\mu_i, \Sigma_i)$, $\Sigma_i = Q_i \Delta_i Q_i^t$ where Q_i is a matrix drawn from the uniform distribution on the set of orthogonal matrices and $\Delta_i = diag\{(p+1-k)/p)^{\theta_i},\ k = 1, \ldots, p\}$, $i = 1, \ldots, c$. The parameter θ_i is randomly drawn from the standard uniform distribution. To prevent too close centroids, each entry of μ_i is the result of adding two samples from the standard uniform distribution. The projection on the two first principal components of the considered mixture is reported on Figure 2, where different colors represent different components. Data appear mixed and the clustering non-trivial. Clustering centroids are randomly initialized 100 times, the iterations of k-means are limited to five to prevent the clustering to converge. The number of clusters is supposed to be known. The response variable Y is simulated as follows:

- For each $i \in \{1, \ldots, c\}$, one of the two possible directions $\beta_j \in \{\beta_1, \beta_2\}$ is randomly selected with probability $1/2$.
- $Y_t = \sinh(X_t^T \beta_j) + \epsilon$, for all $t \in \mathcal{C}_i$, $i \in F^{-1}(j)$ where $\epsilon \sim \mathcal{N}(0, 0.1^2)$ is an error independent of X_t.

The two e.d.r. spaces are randomly generated and orthogonalized: $\beta_1^T \beta_2 = 0$. We are interested in the case where we insert in the same cluster samples from different components. This is the case when we estimate Z by $\hat{Z}$ such that for some (t_1, t_2) we have $\hat{Z}_{t_1} = \hat{Z}_{t_2}$ but $Z_{t_1} \neq Z_{t_2}$.

For each of the 100 runs of k-means, the set of estimated directions by Collaborative SIR $\{\hat{\beta}_{\hat{F}(1)}, \ldots, \hat{\beta}_{\hat{F}(c)}\}$ is considered. The number of samples in each slice is set to 250 resulting in $H = 10$ uniform slices. The average of the squared cosines (3) between the estimated and real directions $\{\beta_{F(1)}, \ldots, \beta_{F(c)}\}$ is computed according to Table 1.

SIR	Collaborative SIR
$\frac{1}{c}\sum_{i=1}^{c} \cos^2(\hat{b}_i, \beta_{F(i)})$	$\frac{1}{c}\sum_{i=1}^{c} \cos^2(\hat{\beta}_{\hat{F}(i)}, \beta_{F(i)})$

Table 1: Quality measures

The 100 results are then averaged. The histogram of the percentage of badly clustered samples is depiceted on Figure 3. In the cases where the clustering has no error, the average of the quality measure is 0.90. Averaging only on the runs of k-means with more than 10 percent of errors, the quality measure decreases to 0.83. This shows that, even if an error on the estimation of Z affects the solution, the influence is, empirically proved, not to be severe. It must be noted that we obtain the worst results when we insert in the same clusters samples with different underlying models: $\hat{Z}_{t_1} = \hat{Z}_{t_2}$ but $Z_{t_1} \neq Z_{t_2}$ and there is no j such that $Z_{t_1}, Z_{t_2} \in F^{-1}(j)$. This is indeed the reason why we extended SIR methodology.

4.2 Test case B

To investigate the strengths and limitations of the method, 100 different mixtures of Gaussian models are generated. Different values of sample size n, dimension p, number of components c and number of e.d.r. spaces D were investigated. Only the case where $n = 2500$, $p = 200$, $D = 2$, $c = 10$ and $\beta_1^T \beta_2 = 0$ is displayed here. The response variable Y is generated as in test case A for each of the 100 datasets. The same slicing strategy with $H = 10$ is applied. We selected such dimension p to mimic the dimensionality of hyperspectral satellite sensors that are of interest in future works. The number of clusters is supposed to be known. Not surprisingly, as soon as the dimension decreases, the performance of the algorithm are more and more stable. Here, at dimension $p = 50$, the performance are still stable and accurate. Analyzing on Figure 4 the histograms of the differences of the average of the squared cosines (Table 1) between Collaborative SIR and SIR, it is evident that Collaborative SIR is always improving the quality of the estimation leading to a significant difference. The average and standard deviation of the 100 quality measures are 0.50 ± 0.05 for SIR and 0.80 ± 0.07 for Collaborative SIR. Since the quality measure is bounded to one, a relevant improvement is found using Collaborative SIR. The estimation $\hat{D}$ of the number of e.d.r. spaces is displayed on Figure 5. The estimation is concentrated around the true value, $D = 2$.

4.3 Comments on simulation results

In the simulations, the sensitivity to clustering and the effective gain in using Collaborative SIR have been analyzed. Several tests changing the dimension

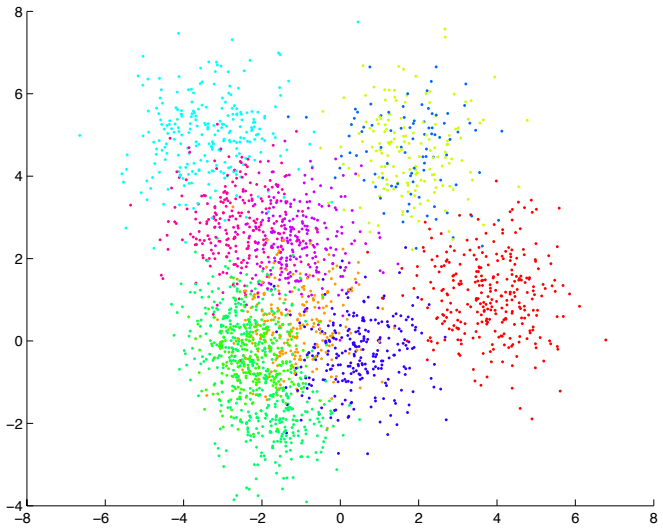

Figure 2: Projection on the two first principal components of the considered mixture, different colors represent different components.

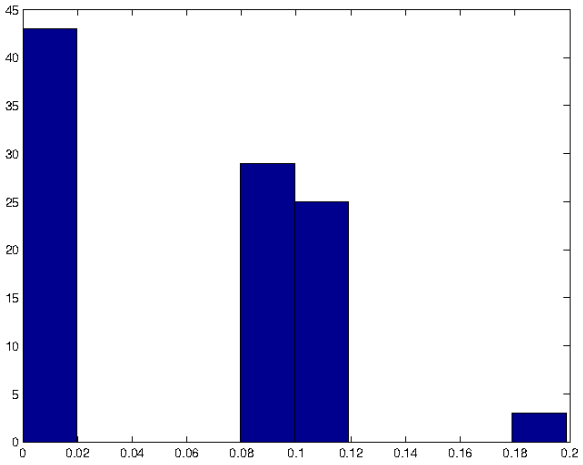

Figure 3: Histogram of the percentage of badly clustered samples over 100 runs of k-means.

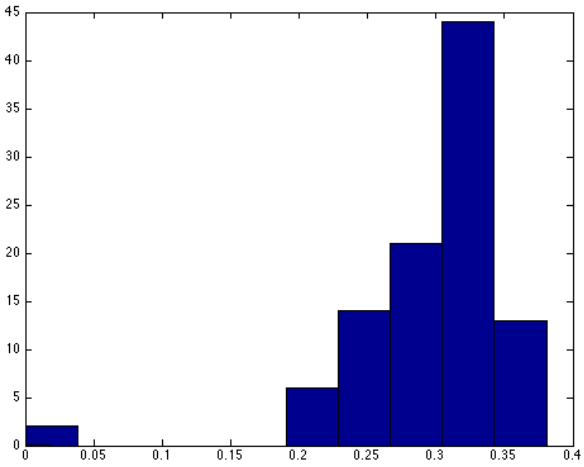

Figure 4: Histogram of the differences between the quality measures (see Table 1) of Collaborative SIR and SIR obtained over 100 different datasets.

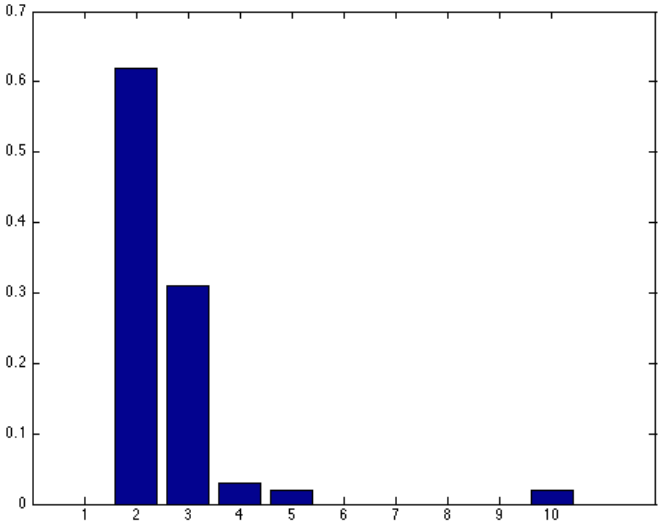

Figure 5: Bar plot of frequencies of the number of estimated e.d.r. spaces $\hat{D}$ over 100 different datasets. Here, the true value is $D = 2$.

p and the collinearity of the β_j's were carried out. Non orthogonal e.d.r. directions and multiple e.d.r. spaces ($D = 2, 3$) have been analyzed reporting good results in case of orthogonality and non orthogonality of the β_j's. As soon as the directions get collinear, our model is no more identifiable, despite that, the results are not affected. In such a case, the different e.d.r. spaces simply reduce to one.

Simulations are interesting but cannot cover the complexity of the real applications. In the following, two real datasets where Collaborative SIR shows its capabilities are discussed and analyzed.

5 Real data application

We show, in the following, two real applications where the number D of different effective dimension spaces differs from one. Nevertheless, it must be underlined that, for many different datasets, $D = 1$ was found. This is extremely satisfying because it means that in those cases a single underlying model, $Y = f(\beta^T X, \epsilon)$, is the best choice for the considered dataset. First, the Horse-mussel dataset, that can be found in [18], is considered. Second, a dataset composed of different parameters measured on galaxies is investigated. Finally, a discussion on possible improvements, strengths and limitations is drawn.

5.1 Horse-mussel dataset

The horse-mussel dataset X is composed of $n = 192$ samples of $p = 4$ numerical measures of the shell: length, width, height and weight. A detailed description can be found in [11]. The response variable Y to predict is the weight of the edible portion of the mussel. To compare to [18], the discrete response variable is transformed into a continuous variable $Y := Y + \epsilon$, $\epsilon \sim N(0, 0.01^2)$. The clustering obtained by [18] into $c = 5$ clusters is adopted and the number of slices is set to four: $H_i = 4$ for all $i \in \{1, \ldots, 5\}$. The following algorithm is used to analyze and compare SIR, cluster SIR and Collaborative SIR:

(1) Randomly select 80% of X for training T and 20% for validation, V.

(2) Apply SIR, cluster SIR and collaborative SIR on the training.

(3) Project and regress the functions using the training samples. Here, a polynomial of degree 2 was used.

(4) Compute the Mean Absolute Relative Error (MARE) on the test: MARE= $\frac{1}{|V|}\sum_{Y\in V}\frac{|Y-\hat{Y}|}{Y}$, where $\hat{Y}$ is the prediction of Y.

We computed 100 different training and validation sets. The boxplots of MARE associated with the three different methods are shown on Figure 6. It must be noted that this dataset is low dimensional: $p = 4$. However it is of interest that the number of e.d.r. spaces found is $\hat{D} = 2$. In Figure 8, the data is decomposed into the two components and the regression of the two link functions appears easier compared to the regression in Figure 7 where the cloud of points associated with SIR is thicker and not well shaped. Using different regression techniques (Gaussian kernel and polynomial regression), the results do not change significantly. On this dataset, Collaborative SIR performs better than SIR and cluster SIR. In addition, this result suggests that two subgroups are present in the data.

5.2 Galaxy dataset

The Galaxy dataset is composed by $n = 292,766$ different galaxies. Aberrant samples have been removed from the dataset after a careful observation of the histograms in each variable supervised by experts. The response variable Y is the specific stellar formation rate. The predictor X is of dimension $p = 46$ and is composed of spectral characteristics of the galaxies. For all the tests, the number of slices is set to $H = 1000$ and the number of samples in the first $H-1$ slices is the closest integer to n/H. We applied Collaborative SIR on the whole dataset to investigate the presence of subgroups and different directions. After different runs and numbers $c \in \{2,\dots,10\}$ of clusters, we observed two different subgroups ($\hat{D} = 2$) and hence directions $\hat{\beta}_1$, $\hat{\beta}_2$.

Best results are reported with $\hat{c} = 5$. The shapes of two nonlinear link functions appear on the projected data, see Figure 9. Clouds are thick but they show a very clear trend in the distributions. This dataset is a good example of how, in high dimension, two families can be found in a dataset using Collaborative SIR. The coefficients of the two estimated directions $\hat{\beta}_1$, $\hat{\beta}_2$ are displayed on Figure 10. It is interesting to observe how some variables are contributing in both linear combinations, but that there is also a reasonable difference in four variables (variables 2, 3, 6 and 23). Let us note

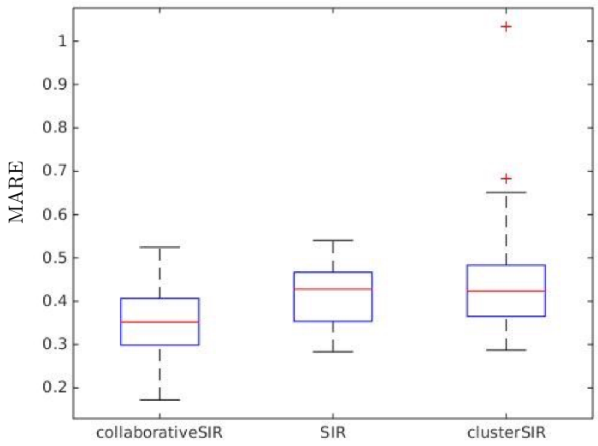

Figure 6: Boxplots of MARE for Collaborative SIR, SIR and Cluster SIR using 100 different initializations.

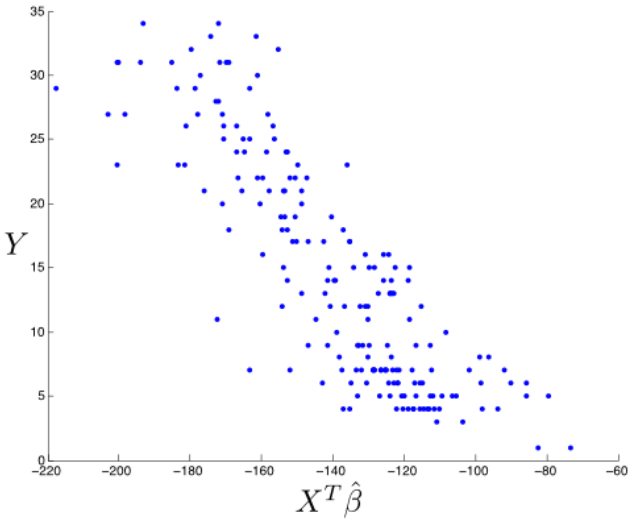

Figure 7: Graph of Y versus the projection along the direction $\hat{\beta}$ found by SIR.

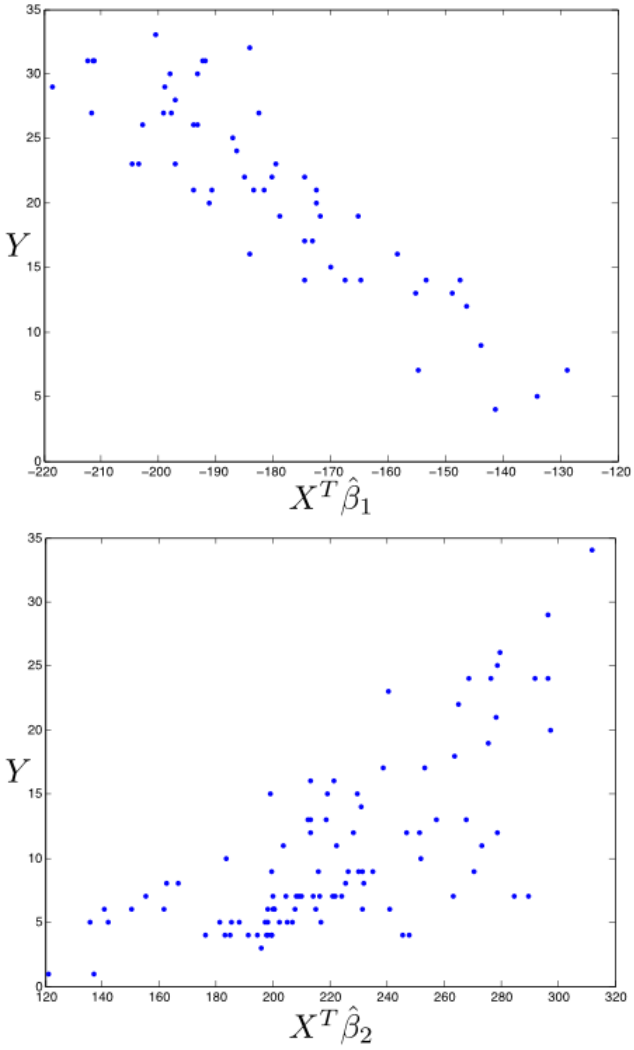

Figure 8: Top: Graph of Y versus the projection along the first direction $\hat{\beta}_1$ found by Collaborative SIR. Bottom: Graph of Y versus the projection along the second direction $\hat{\beta}_2$ found by Collaborative SIR. The estimated directions $\hat{\beta}_1, \hat{\beta}_2$ are nearly orthogonal: $\cos^2(\hat{\beta}_1, \hat{\beta}_2) = 0.01$.

that, variable 40, found to be relevant for both directions, is often used to estimate the specific stellar formation rate. Experts are working on a possible physical interpretation of the results. Even if the link functions look similar, we observe a significant difference in the coefficients of the two directions. This could lead to a better understanding and designing of further analysis of this kind of data.

5.3 Discussion on dimension k and the number of clusters c

In the whole paper, we presented results for unidimensional e.d.r. spaces ($k = 1$). It is worth noticing that the entire approach can be easily extended to a higher value of k, common to all the e.d.r. spaces. To this end, k should be estimated, which is a recurrent problem in SIR methodology [2, 19]. The graphical procedure introduced by [22] could be used for instance. It is then sufficient to consider a proximity measure between the linear subspaces, see [7] for a measure based on projectors. Our strategy can still be applied leading to a hierarchical merging tree. In case the dimension k varies along given mixture components, we suggest to set to zero the proximity between the corresponding e.d.r. spaces. Let us also note that SIR is a method to reduce dimensionality in order to "better" perform regression. When a regression is performed, the visualization of the results is crucial, that is one of the reasons for dimensionality reduction. If the dimension k is greater than two, visualization is not possible. This explains why SIR and its variants have mainly been applied with $k = 1$. Collaborative SIR is first dividing the predictors space into clusters, it seems natural to assume that dimension k locally would be smaller than globally. Considering $k = 1$ may not be a severe restriction if a visualization is needed. Finally, another drawback of increasing dimensionality is that the samples become more and more sparse and may not cover enough the surface we want to regress. Different regression techniques may then lead to dramatically different results.

We did not give an automatic way of selecting the number of clusters. In SIR literature [18], this selection is translated into an optimization problem. Nowadays, with the increasing capabilities of sensors, data are complex and complicated and it is hard to define a general criteria, ignoring previous knowledge, that could work for any kind of data. The number of clusters is deeply connected with how we want to group elements, the same data

can show two possible "correct" clusterings, depending on the task. Since SIR and collaborative SIR are fast and simple techniques, the user, using prior information, should orient the clustering and try different values for the parameters and empirically check which is the most suitable for the purpose. Developing flexible clustering capable of incorporating prior knowledge is one of our interests.

6 Conclusion and future work

Sliced Inverse Regression is an interesting and fast tool to explore data in regression, it is yet not so popular [9], but benefits from well established theory and simple implementation. If the link function turns out to be linear, SIR, not surprisingly, is outperformed by linear regression techniques. At the opposite, in case of evidence of non linearities, linear regression techniques force the model resulting in poor estimations. Collaborative SIR is meant to deal with the increasing complexity of the datasets that statisticians are asked to analyze. Often there is no reasonable criteria of gathering the samples, resulting in dataset that are, at least, a mixture of different phenomena and/or full of ambiguous samples. The hypothesis of having different families with different underlying models gives flexibility not affecting tractability. We encourage the community to improve our idea.

Acknowledgement

The authors thank Didier Fraix-Burnet for his contribution to the data. They are grateful to Vanessa Kuentz and Jérôme Saracco for providing their results on Horse-mussel dataset. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

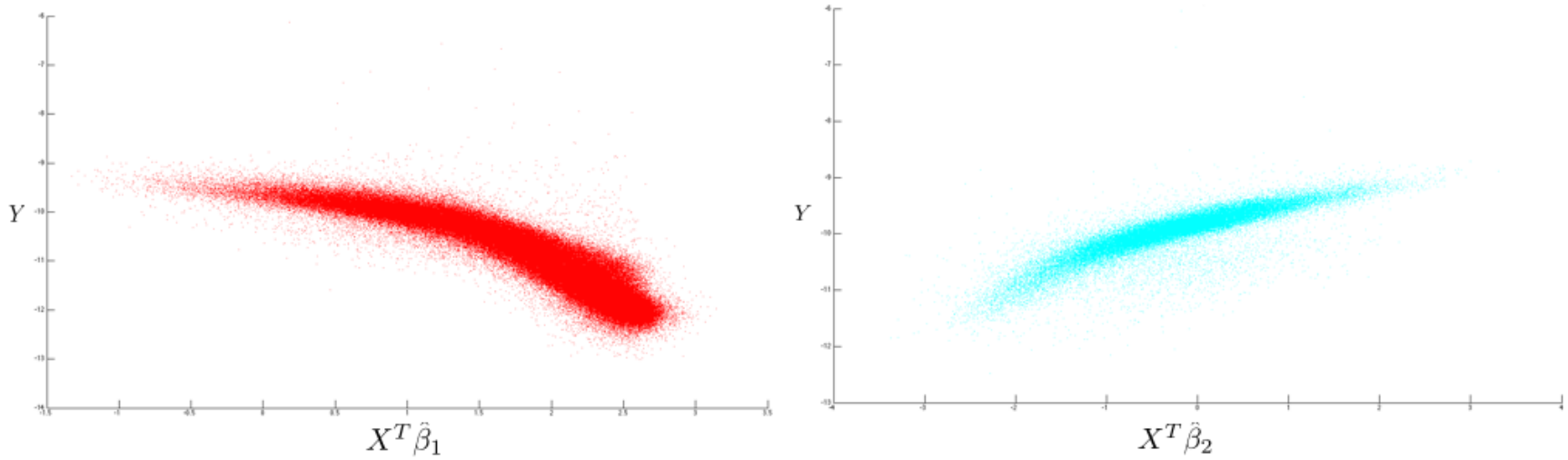

Figure 9: Left: Graph of Y versus the projection along the first e.d.r. direction $\hat{\beta}_1$ estimated by Collaborative SIR. Right: Graph of Y versus the projection along the second e.d.r. direction $\hat{\beta}_2$ estimated by Collaborative SIR. The nonlinear behavior of the two link functions appears clearly.

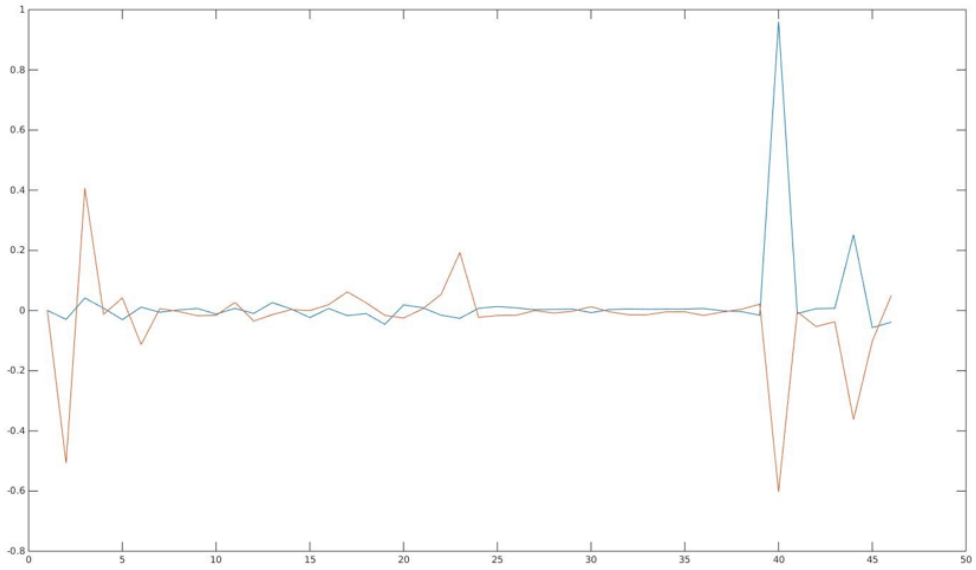

Figure 10: Comparison of e.d.r. directions $\hat{\beta}_1$ and $\hat{\beta}_2$. Many elements in the vectors are close to zero resulting in a variable selection. Differences in the two lines show how different variables contribute in regressing Y. The squared cosine between the two directions is 0.42.

References

[1] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38. ACM, 2015.

[2] M Pilar Barrios and Santiago Velilla. A bootstrap method for assessing the dimension of a general regression problem. *Statistics & probability letters*, 77(3):247–255, 2007.

[3] Laurent Bergé, Charles Bouveyron, and Stéphane Girard. HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, 46(6):1–29, 2012.

[4] Caroline Bernard-Michel, Sylvain Douté, Mathieu Fauvel, Laurent Gardes, and Stéphane Girard. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using regularized sliced inverse regression. *Journal of Geophysical Research: Planets (1991–2012)*, 114(E6), 2009.

[5] Caroline Bernard-Michel, Laurent Gardes, and Stéphane Girard. Gaussian regularized sliced inverse regression. *Statistics and Computing*, 19(1):85–98, 2009.

[6] Charles Bouveyron, Stéphane Girard, and Cordelia Schmid. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007.

[7] Marie Chavent, Stéphane Girard, Vanessa Kuentz-Simonet, Benoit Liquet, Thi Mong Ngoc Nguyen, and Jérôme Saracco. A sliced inverse regression approach for data stream. *Computational Statistics*, 29(5):1129–1152, 2014.

[8] Marie Chavent, Vanessa Kuentz, Benoît Liquet, and Jérôme Saracco. A sliced inverse regression approach for a stratified population. *Communications in Statistics-Theory and Methods*, 40(21):3857–3878, 2011.

[9] Chun-Houh Chen and Ker-Chau Li. Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8(2):289–316, 1998.

[10] Francesca Chiaromonte and Jessica Martinelli. Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, 176(1):123–144, 2002.

[11] R Dennis Cook and Sanford Weisberg. *Applied regression including computing and graphics*, volume 488. John Wiley & Sons, 2009.

[12] Raphaël Coudret, Stéphane Girard, and Jérôme Saracco. A new sliced inverse regression method for multivariate response. *Computational Statistics & Data Analysis*, 77:285–299, 2014.

[13] Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984.

[14] Naihua Duan and Ker-Chau Li. Slicing regression: a link-free regression method. *The Annals of Statistics*, 19(2):505–530, 1991.

[15] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013.

[16] Peter Hall and Ker-Chau Li. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21(2):867–889, 1993.

[17] Tailen Hsing and Raymond J Carroll. An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20(2):1040–1061, 1992.

[18] Vanessa Kuentz and Jérôme Saracco. Cluster-based sliced inverse regression. *Journal of the Korean Statistical Society*, 39(2):251–267, 2010.

[19] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[20] Lexin Li, R Dennis Cook, and Christopher J Nachtsheim. Cluster-based estimation for sufficient dimension reduction. *Computational Statistics & Data Analysis*, 47(1):175–193, 2004.

[21] Lexin Li and Hongzhe Li. Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, 20(18):3406–3412, 2004.

[22] Benoît Liquet and Jérôme Saracco. A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Computational Statistics*, 27(1):103–125, 2012.

[23] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[24] Jérôme Saracco. An asymptotic theory for sliced inverse regression. *Communications in Statistics-Theory and Methods*, 26(9):2141–2171, 1997.

[25] Luca Scrucca. Regularized sliced inverse regression with applications in classification. In *Data Analysis, Classification and the Forward Search*, pages 59–66. Springer, 2006.

[26] Luca Scrucca. Class prediction and gene selection for dna microarrays using regularized sliced inverse regression. *Computational Statistics & Data Analysis*, 52(1):438–451, 2007.

[27] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J Candes. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

[28] Laurens JP Van der Maaten, Eric O Postma, and H Jaap van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71, 2009.

[29] Wenxuan Zhong, Peng Zeng, Ping Ma, Jun S Liu, and Yu Zhu. Rsir: regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21(22):4169–4175, 2005.

[30] Li-Ping Zhu. Extending the scope of inverse regression methods in sufficient dimension reduction. *Communications in Statistics-Theory and Methods*, 40(1):84–95, 2010.