# Collaborative Sliced Inverse Regression

Alessandro Chiancone, Stéphane Girard, Jocelyn Chanussot

# Collaborative Sliced Inverse Regression

Alessandro Chiancone[a,b,c], Stephane Girard[a], Jocelyn Chanussot[b]

[a] *Laboratoire Jean Kuntzmann & INRIA Rhone-Alpes, team Mistis, Inovallee, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France*
[b] *GIPSA-Lab, Grenoble INP, Saint Martin d'Heres, France*
[c] *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 because 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. SIR, not surprisingly, is not capable of dealing with a mixture of Gaussians with different underlying models whereas our approach is able to correctly investigate the mixture. A study on simulated data as well as two real applications is provided.

*Keywords:* Mixture models, inverse regression, sufficient dimension reduction

## 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. Van der

Maaten et al. [21] provide an interesting review concluding that even if the variety of non-linear methods is huge, Principal component Analysis (PCA) [16], 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, to catch the intrinsic dimensionality of the data, which is the minimum numbers of parameters needed to describe it [9]. 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). Encoded in $Y$ there is additional information 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 $Y$.
Regression framework is characterized by the assumption of a link function between $X$ and $Y$ i.e. $Y = f(X, \epsilon)$, where $\epsilon$ 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 $Y = f(X^T\beta_1, ..., X^T\beta_k, \epsilon)$ the effective dimension reduction (e.d.r) space is what we are looking for [13]. 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 $\beta_1, ..., \beta_k$. Sliced Inverse Regression (SIR) [13] 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 [12] or [20] for a detailed discussion and references). Our hypothesis is that the e.d.r. space is not unique all over the data and 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? [4].
In section 2 we rapidly 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. Motivation and main problem are described. Asymptotic results are established under mild conditions. The simulation study, section 4, is where the performances of

Collaborative SIR are shown and analyzed under 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 conclusion are finally drawn encouraging the community to improve our idea.

## 2. Sliced Inverse Regression (SIR)

Back in 1991, Li [13] called SIR 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 are nowadays of extreme importance because human beings are still, unfortunately, limited by a perception which only allows us to display 3 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 Li proposed is:

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

where $Y \in \mathbb{R}$ is the response variable, $X$ is a random variable, $X \in \mathbb{R}^p$ ($\Sigma = \mathrm{Cov}(X)$, $\mu = \mathbb{E}(X)$). $\epsilon$ is a random error independent of $X$. If $k \ll p$ the functions 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:

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

Duan and Li [8] showed that the centered inverse regression curve is contained in the $k$-dimensional linear subspace of $\mathbb{R}^p$ spanned by $\Sigma\beta_1, ..., \Sigma\beta_k$. If we consider a monotone transformation $T(\cdot)$ of $Y$, the matrix $\Sigma^{-1}\Gamma$ is degenerated in any direction orthogonal to $\beta_1, ..., \beta_k$, 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 $\Gamma$, Li [13] used a slicing procedure as candidate for $T(\cdot)$. Dividing the range of $Y$ in non-overlapping slices, $s^1, ..., s^H (H > 1)$. $\Gamma$ can then be written as:

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

where $p^h = P(Y \in s^h)$ and $m^h = \mathbb{E}(X|Y \in s^h)$. The estimator $\hat{\Gamma}$ can then be defined substituting $p^h, m^h$ with the corresponding sample versions. The $k$ eigenvectors corresponding to the largest eigenvalues of $\hat{\Sigma}^{-1}\hat{\Gamma}$ are the estimation of a basis of the e.d.r. space.

*2.1. Limitations*

SIR's theory is well established and comes fully equipped by asymptotic results [11, 17]. Two main limitations affect the building:

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

When the number of samples is $n \leq p$ the sample covariance matrix is singular, and when the variables are highly correlated (e.g. in hyperspectral images) the covariance matrix is ill conditioned. To compute the e.d.r directions the inversion of $\hat{\Sigma}$ must be achieved, recently many papers faced this problem and provided solutions ([5, 15, 18, 19, 22]). An homogeneous framework to perform regularized SIR has been proposed in [1] where, depending on the choice of the prior covariance matrix, the above mentioned techniques can be obtained and extended.

The $(LDC)$, less studied in literature, is the central assumption of the theory and it depends on the unobserved e.d.r. directions, therefore it cannot be directly checked [23]. It can be proved that if $X$ is elliptical distributed the condition holds. This condition is much stronger than $(LDC)$ but easier to verify in practice since it does not depend on the $\beta_1$,...,$\beta_k$. Good hope comes from a result of Hall and Li [10] 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 [7] where the authors show that high dimensional dataset are nearly normal in most of the low dimensional projections. If $X$ follows an elliptical distribution the $(LDC)$ condition holds, 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. Kuentz and Saracco [12] using an idea from [14] proposed to clusterize the space to look locally for ellipticity rather than globally. Chavent et al. [3] introduced categorical predictors to

distinguish different populations. This is our very start, assuming $X$ from a mixture model we focus on decomposing the mixture and we extend the basic model to improve SIR's capability to explore complex datasets.

## 3. Collaborative SIR

First, we give a motivation and introduce in subsection 3.1 the population version of Collaborative SIR. Second, a sample version in different steps is detailed and an algorithm is outlined (subsections 3.2-3.5). For sake of simplicity we will focus on the case when $k = 1$ i.e. the effective dimension reduction space is of dimension one.

### *3.1. Population version*

In SIR the underlying model 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 hypothesis considering the data a realization from a weighted sum of distributions with different parameters. As mentioned before, in such case there is no straightforward way to check if the $(LDC)$ holds. Let $X$ be a random vector, $X \in \mathbb{R}^p$, from a mixture model and be $Z$ an unobserved latent random variable $Z \in \{1, ..., c\}$, 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, $Y \in \mathbb{R}$, $F$ is an unknown deterministic function $F : \{1, ..., c\} \to \{1, ..., D\}$, $D \in \mathbb{N}$. The functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, ..., D$ are unknown link functions between $X$ and $Y$. Finally $\epsilon_i$ are random errors $\forall i \ \epsilon_i \in \mathbb{R}$, i.e. each component is allowed to have a different related error.

Under the 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 $\beta_1, ..., \beta_D$. 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) $\forall i = 1, ..., c\ \mathbb{E}(X^T b | X^T \beta_{F(i)}, Z = i)$ 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 matrix $\Sigma_i = Cov(X|Z = i)$ 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 [13].
If $F : \{1, ..., c\} \to \{1, ..., D\}$ is known, the inverse image of the elements $j \in \{1, ..., D\}$ can be defined:

$$F^{-1}(j) = \{i \in \{1, ..., c\} \text{ s.t. } F(i) = j\},$$

since $F$ is not required to be injective, an e.d.r direction $\beta_i$ may be associated with several components. Suppose that $\{b_i, i \in F^{-1}(j)\}$ are observed, given the proximity criteria

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

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

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

where $B_j = [b_{i, i \in F^{-1}(j)}]$. Using Lagrange multipliers is easy to show that vector $v$ must be an eigenvector of the matrix $(B_j^T B_j)$ and, since we want to maximize, it will be the one associated with the largest eigenvalue. The following lemma motivates this argument.

**Lemma 1.** *Assuming the (LCD) and model* (2) *the eigenvector* $\tilde{\beta}_j$ *associated to the only non-zero eigenvalue of the matrix* $[B_j B_j^T]$ *is collinear with* $\beta_j$.

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

$[B_j B_j^T] = \sum_{i \in F^{-1}(j)} \alpha_i^2 \beta_j \beta_j^T = \|\alpha\|^2 \beta_j \beta_j^T$. This concludes the proof. □

This lemma shows that $\tilde{\beta}_j$ is an e.d.r. direction for each $j$ and the precedent argument gives a strategy to estimate the directions $\beta_j$ based on the

proximity criteria (3).

**Remark.** If $D = 1$ then $F^{-1}(1) = \{1, ..., c\}$, the e.d.r. direction and the link functions do not vary through all the mixture. This specific case is addressed in [12].

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

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

$$\bar{X}_i = \frac{1}{n_i} \sum_{t \in \mathcal{C}_i} X_t, \; \hat{\Sigma}_i = \frac{1}{n_i} \sum_{t \in \mathcal{C}_i} (X_t - \bar{X}_i)(X_t - \bar{X}_i)^T, \text{ for each } i = 1, ..., c.$$

To obtain an estimator for $\Gamma_i$, we introduce as in classical SIR a slicing. For each $\mathcal{C}_i$ we can define the slicing $T_i$ of $Y_i$ into $H_i \in \mathbb{N}$ slices ($H_i > 1 \; \forall i = 1, ..., c$). Let $s_i^1, ..., s_i^{H_i}$ 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 $p_i^h = P(Y \in s_i^h | Z = i)$, $m_i^h = \mathbb{E}(X | Z = i, Y \in s_i^h)$. Let us recall that $\mu_i = \mathbb{E}(X|Z = i)$ and $\Sigma_i = Cov(X|Z = i)$, as defined in section 3.1. Let $n_{h,i} = \sum_{t \in \mathcal{C}_i} \mathbb{I}[Y_t \in s_t^h]$, where $\mathbb{I}$ is the indicator function. Replacing $p_i^h, m_i^h$ with the corresponding sample versions, it is possible to estimate $\Gamma_i$:

$$\hat{\Gamma}_i = \sum_{h=1}^{H_j} \hat{p}_i^h (\hat{m}_i^h - \bar{X}_i)(\hat{m}_i^h - \bar{X}_i)^T,$$

where $\hat{p}_i^h = \dfrac{n_{h,i}}{n_i}$ and $\hat{m}_i^h = \dfrac{1}{n_{h,i}} \sum_{t \in \mathcal{C}_i} X_t \mathbb{I}[Y_t \in s_t^h]$. The estimated e.d.r. directions are then $\hat{b}_1, ..., \hat{b}_c$ 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 $\beta_j$:

- (i) $\hat{B}_j = [\hat{b}_{i,i\in F^{-1}(j)}]$, $i \in \{1, ..., c\}$, $\hat{B}_j$ is a $p \times |F^{-1}(j)|$ matrix;
- (ii) $\hat{\beta}_j$ $\forall j = 1, ..., D$ is the major eigenvalue of $\hat{B}_j^T \hat{B}_j$.

Asymptotic results can be establish similarly to Chavent et al. [2]. We fix $j \in \{1, ..., D\}$ and consider $\{X_t,\ t \in \bigcup_{i\in F^{-1}(j)} \mathcal{C}_i\}$ and a sample size $n^j = \sum_{i\in F^{-1}(j)} n_i$ which tends to $\infty$. The following three assumption are considered:

- (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 $i$, the support of $\{Y_t, t \in \mathcal{C}_i\}$ is partitioned into a fixed number $H_t$ of slices such that $p_i^h > 0, h = 1, \ldots, H_t$.
- (A3) For each $i$ and $h = 1, \ldots, H_t$, $n_{h,i} \to \infty$ (and therefore $n_i \to \infty$) as $n \to \infty$.

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

*(i)* $\widehat{\beta}_j = \beta_j + O_p(\underline{n}^{j^{-1/2}})$, *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 [13] each estimated EDR direction $\widehat{b}_i$ converges to $\beta_j$ at root $\underline{n}^j$ rate: that is, for $i \in F^{-1}(j)$, $\widehat{b}_i = \beta_j + O_p(\underline{n}^{j^{-1/2}})$. We then have $\widehat{B}_j^T \widehat{B}_j = B_j^T B_j + O_p(n^{j^{-1/2}})$. Therefore the principal eigenvector of $\widehat{B}_j^T \widehat{B}_j$ converges to that corresponding to $B_j^T B_j$ at the same rate: $\hat{\beta}_j = \beta_j + O_p(\underline{n}^{j^{-1/2}})$. The estimated e.d.r. direction $\hat{\beta}_j$ converges to an e.d.r. direction at root $\underline{n^j}$ rate. □

(*ii*) The proof is similar to the one of Chavent et al. [2], Theorem 2.

In the following sections 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 $\hat{b}_1, ..., \hat{b}_c$.

**Definition**. Let $V = \{v_1, v_2, ..., v_{|V|}\}$ be a set of vectors in dimension $p$ with associated weights $w_i$. We define the quantity $\lambda(V)$:

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

$$= \text{largest eigenvalue of } \frac{1}{w_V} \sum_{i=1}^{|V|} w_i v_i v_i^T$$

where $w_V = \sum_{i=1}^{|A|} w_i$ is the normalization. Vector $v$ maximizing $\lambda(V)$ is the most collinear vector to our set of vectors given the proximity criteria (3) and the weights $w_i$. To build the hierarchy we consider the following iterative algorithm initialized with the set $A = \{\{\hat{b}_1\}, ..., \{\hat{b}_c\}\}$:

**while** $card(A) \neq 1$

**Let** $a, b \in A$ **such that** $\lambda(a \cup b) > \lambda(c \cup d) \forall c, d \in A$
$A = (A \setminus \{a, b\}) \bigcup a \cup b$
**end**

the weights are set equal to the number of samples in each components, i.e. $w_i = n_i$, $i = 1, ..., c$. At each step the cardinality of the set $A$ decreases merging the most collinear sets of directions (Fig. 1). 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 ($\hat{b}_1, \hat{b}_2$ are selected to be merged in Fig. 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 $A = \{\{\hat{b}_1, \hat{b}_2\}, \{\hat{b}_3\}, ..., \{b_{12}\}\}$ in Fig. 1)

Therefore it is possible to infer the number $D$ of underlying e.d.r. spaces analyzing the values of $\lambda$ in the hierarchy (Fig. 2) looking for a discontinuity that will occur when two sets with different underlying $\beta_j$ (i.e. non collinear) are merged. We automatically estimate $D$ with the following procedure:

($i$) Draw a line from the first value of the graph $(1, \lambda_1)$ to the last $(c, \lambda_c)$.

($ii$) Compute the distance between points in the graph and the line.

($iii$) Select the merging point maximizing that distance. $\hat{D} = c-$number of merge selected.

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

*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 $\hat{\beta}_1, ..., \hat{\beta}_{\hat{D}}$. In the following paragraph a procedure for the estimation $\hat{F}$ of the function $F$ is detailed.

Once the candidates $\hat{\beta}_1, ..., \hat{\beta}_{\hat{D}}$ are estimated, the whole data ($X$,$Y$) is considered to estimate $F$. Starting from $i \in \{1, ..., \hat{c}\}$ the goal is to find $j \in \{1, ..., \hat{D}\}$ such that $F(i) = j$, under certain conditions. The $\hat{D}$ covariance matrices of the distributions $(X_t^T \hat{\beta}_j, Y_t)$, $t \in \mathcal{C}_i$, $j \in \{1, ..., \hat{D}\}$ are considered. The idea is to select the direction that best explains $Y_t$, $t \in \mathcal{C}_i$ among the estimated directions $\hat{\beta}_1, ..., \hat{\beta}_{\hat{D}}$.
Let us assume $f_j$ functions "locally" linear ($A4$): $f_j$ can be approximated with piecewise linear functions so that $Y_t = f_j(X_t^T \beta_j) = k_i X_t^T \beta_j$, $\forall t \in \mathcal{C}_i$, $i \in F^{-1}(j)$.

**Lemma 2.** *Let $j \in \{1, ..., D\}$. Under assumption ($A4$) the e.d.r. direction $\beta_j$ is the vector minimizing the second eigenvalue of the covariance matrix of the pairs $(X^T \beta_s, Y)_{s=1,...,D}$.*

*Proof.* We have that:

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

where the scalar product and the norm are induced by $\Sigma$. The characteristic polynomial is $p(\lambda) = \lambda^2 - \lambda(\|\beta_s\|^2 + k_j^2\|\beta_j\|^2) + k_j^2(\|\beta_s\|^2\|\beta_j\|^2 - \langle\beta_s,\beta_j\rangle^2)$. We have $\Delta = (\|\beta_s\|^2 - k_j^2\|\beta_j\|^2) + 4k_j^2\langle\beta_s,\beta_j\rangle^2 > 0$. From Cauchy-Schwarz inequality $\lambda_1, \lambda_2 \geq 0$ and $\lambda_2 = 0$ if and only if the equality holds. Since $\beta_s$, $s = 1, ..., D$ are linearly independent it follows that $\lambda_2 = 0$ if and only if $\beta_s = \beta_j \Leftrightarrow s = j$. □

In practice, fixed $i = \{1, ..., \hat{c}\}$, vectors $\hat{\beta}_j$, $j = 1, ..., \hat{D}$ are the candidates for $(X_t, Y_t)$, $t \in \mathcal{C}_i$. Lemma 2 is stating that under the assumption $(A4)$ the vector $\hat{\beta}_j$ minimizing the second eigenvalue of $(X_t^T\hat{\beta}_s, Y_t)$, s=1,...,D, $t \in \mathcal{C}_i$ is such that $j = F(i)$. We require the functions to be locally linear, if the functions are approximately linear the estimation will work. 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 $D$ is overestimated $\hat{D} > D$ (e.g. due to instabilities in the estimation of the direction in some components) in the simulation we observed that the estimation of $F$ mitigates this error often avoiding to select the aberrant directions $\beta_j$, $j > D$.

### *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 [12]. Twenty initial random centroids are chosen as initialization of k-means, the one minimizing the sum of squares is retained.

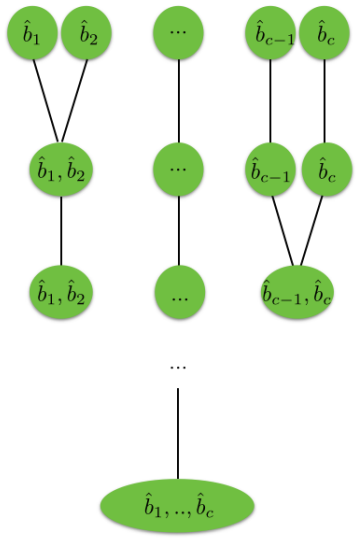


Figure 1: Hierarchy built following the proximity criteria (3).

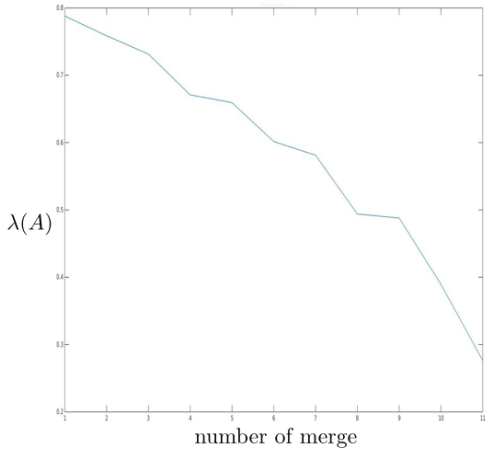


Figure 2: Cost function $\lambda(A)$, the number $D$ of unknown e.d.r directions decreases at each step by one. $\hat{D} = c-$number of merge selected. In the example above $c = 12$. The algorithm selects merge step 9 which corresponds to the correct estimation of the parameter: $\hat{D} = 3$.

## 4. Simulation study

We performed a study on simulated data, this was the opportunity 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 compared to SIR performed independently in each cluster.

The first experiment is performed on the same dataset to study 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 different conditions.

### *4.1. Test case A*

To study the sensitivity to clustering $n = 2500$ samples from Gaussian mixture model are drawn with uniform mixing proportions and $c = 10$ components. Each component follows a Gaussian distribution $\mathcal{N}(\mu_i, \Sigma_i)$, $\Sigma_i = Q_i \Delta Q_i^t$ where $Q_i$ is a matrix drawn from the uniform distribution on the set of orthogonal matrices and $\Delta_{ii} = (\frac{(p+1-i)}{p})^{\theta_i}$. The parameter $\theta_i$ is randomly drawn from the standard uniform distribution. To prevent too close centroids, each entry of the $\mu_i$ is the result of adding two samples from the standard uniform distribution. In figure 3 the projection on the two first principal components of the considered mixture is reported, different colors represent different components. Data in figure 3 appear mixed and 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. $Y$ is simulated as follows:

- For each $i \in \{1, ..., c\}$, one of the two possible directions $\beta_j \in \{\beta_1, \beta_2\}$ is randomly selected with probability $1/2$.
- $Y_t = sinh(X_t^T \beta_j) + \epsilon$, $\forall t \in \mathcal{C}_i$, $i \in F^{-1}(j)$ 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 when 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 estimated directions for Collaborative SIR $\{\hat{\beta}_{\hat{F}(1)}, ..., \hat{\beta}_{\hat{F}(c)}\}$ are considered. The average of the squared cosines (3) between the estimated and real direction $\{\beta_{F(1)}, ..., \beta_{F(c)}\}$ is computed (see column 2 Table 1). The 100 results are then averaged. In the cases where clustering has zero error the average of the quality measure is 0.8958. Averaging only on the runs of k-means with more than 10 percent of error the quality measure decreases to 0.8273. This shows that even if, not surprisingly, 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's theory.

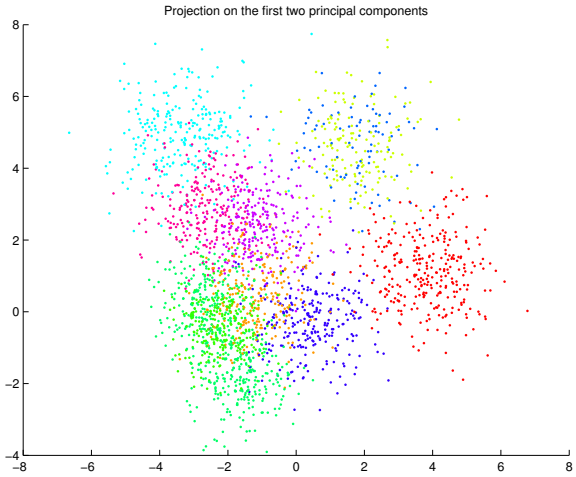


Figure 3: Projection on the two first principal components of the considered mixture, different colors represent different components.

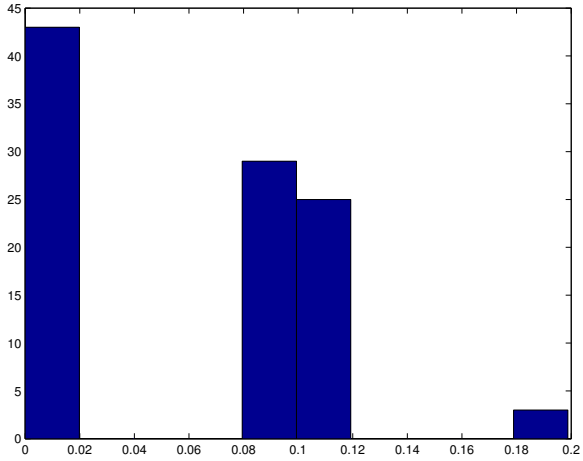


Figure 4: Histograms of the percentage of badly clustered samples over 100 runs of k-means.

*4.2. Test case B*

To investigate the strengths and limitations of the method 100 different mixture of Gaussian models for different numbers of total samples (10000, 5000, 2500) are generated. Only the case where $n = 2500$, dimension $p = 200$, $D = 2$, $c = 10$ 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. 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, e.g. at dimension $p = 50$ the performance are still stable and accurate. Analyzing the histograms of the differences of the average of the squared cosines (Table 1) between Collaborative SIR and SIR (figure 5) it is evident that Collaborative SIR is always improving the quality of the estimation leading to a significant difference. Averaging the 100 quality measures results in: $0.50 \pm 0.05$ for SIR and $0.80 \pm 0.07$ for Collaborative SIR. Since the quality measure is bounded to 1, a relevant improvement is found using Collaborative SIR. In figure 6 we show the estimation $\hat{D}$ of the number of e.d.r. spaces. The estimation is concentrated around the true value, $D = 2$.

Table 1: Quality measure

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

*4.3. Simulation results*

In the simulations the sensitivity to clustering and the effective gain in using Collaborative SIR is analyzed. Several tests changing the dimension $p$, and the collinearity of the $\beta_j$ were carried out. As soon as the directions get collinear our model is no more identifiable, despite that, the results are not affected. When the vectors are, in the limit, collinear the e.d.r spaces simply reduce to one. Non orthogonal e.d.r. directions and multiple e.d.r. spaces ($D = 3$) have been analyzed reporting good results in case of orthogonality and non orthogonality of the $\beta_j$'s. Simulations are interesting but cannot cover the complexity of the real application. In the following, two real dataset where Collaborative SIR shows its capabilities are discussed and analyzed.

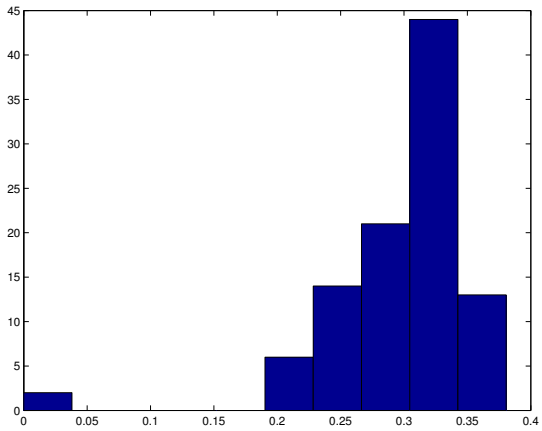


Figure 5: Histograms of the difference between the quality measure (table 1) of Collaborative SIR and SIR obtained over 100 different dataset.

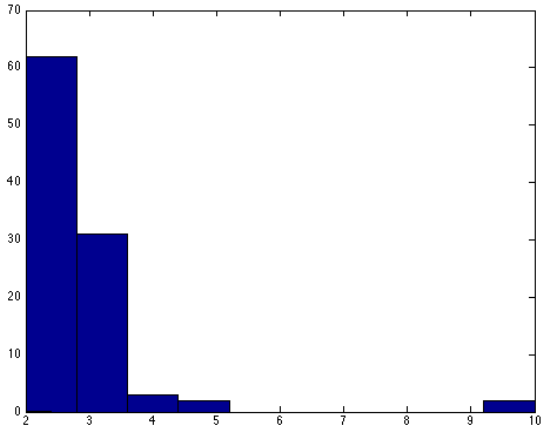


Figure 6: Histograms of the number $\hat{D}$ of estimated e.d.r spaces, $D = 2$.

## 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 Kuentz and Saracco [12], is considered. Second, a dataset composed of different parameters 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 different numerical measures of the shell: length, width, height and weight ($p = 4$, a detailed description can be found in Cook and Weisberg [6]). The response variable $Y$ to predict is the weight of the edible portion of the mussel. To compare to [12] the discrete response variable was transformed into a continuous variable $Y = Y + \epsilon$, $\epsilon \sim N(0, 0.01^2)$. The clustering obtained by [12] was adopted and the number of slices set to four: $H_i = 4$ for all $i \in \{1, ..., 5\}$. 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 (we fitted a polynomial of degree 2)

(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 our estimation.

We computed 100 different training and validation set. In figure 7 the box plots of the three different methods are shown. 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 and the regression of the two link functions appears easier compared to the regression in figure 9 where the cloud of point 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 performed better than SIR and cluster SIR. In addition, this result suggests that two subgroups are present in the data.

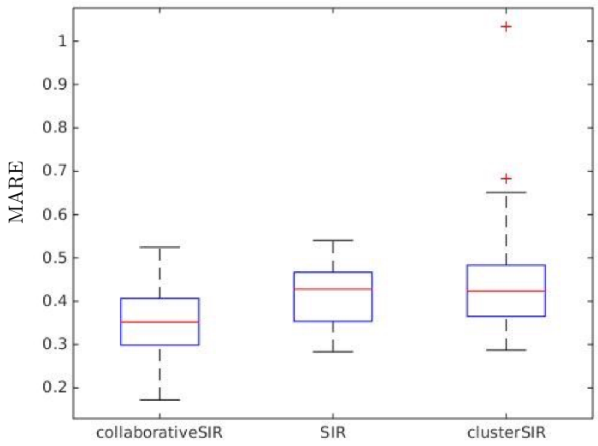


Figure 7: Box plots of MARE for Collaborative SIR, SIR and Cluster SIR using 100 different initializations.

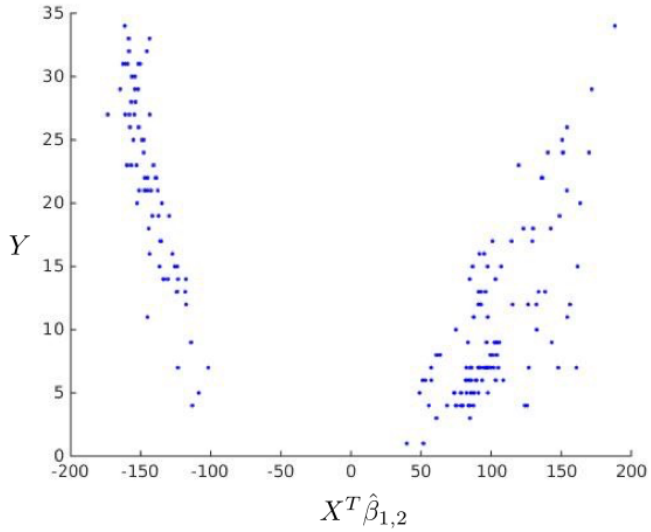


Figure 8: Graph of $Y$ and the projection along the two directions $\hat{\beta}_1, \beta_2$ found by Collaborative SIR.

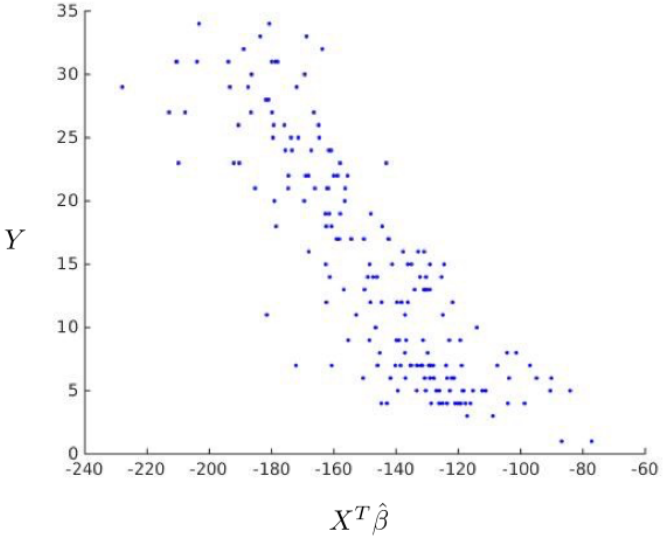


Figure 9: Graph of $Y$ projection along the direction $\hat{\beta}$ found by SIR.

*5.2. Galaxy dataset*

The Galaxy dataset is composed by $n = 292766$ 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. We applied Collaborative SIR on the whole dataset to investigate the presence of subgroups and different directions.
After different runs and number $\hat{c} = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ of clusters we observed two different subgroups and hence directions $\hat{\beta}_1, \hat{\beta}_2$.

Best results are reported with $\hat{c} = 5$, in figure 10 the two non linear link functions are shown. 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.
In figure 11 the distribution of the coefficient of the two directions is presented. It is interesting to observe how some variables are contributing in both linear combinations but that there is a reasonable difference in four variables (variables 2, 3, 6 and 23). The $d4000_n$ (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 coefficient of the two directions. This could lead to a better understanding and designing of further analysis of this kind of data.

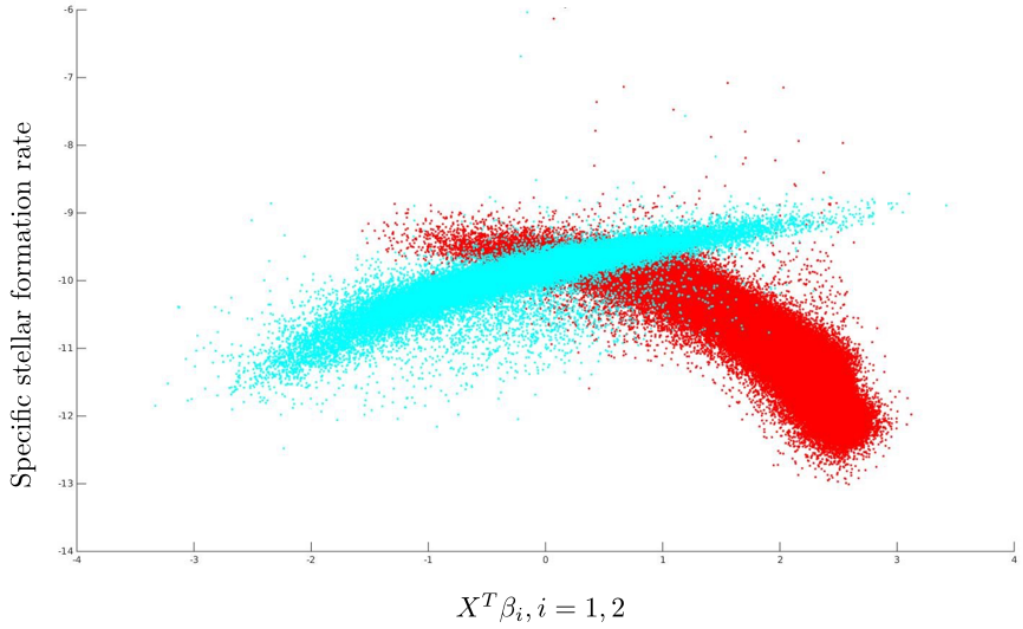


Figure 10: Graph of $Y$ and the projection along the two e.d.r. directions $\hat{\beta}_1, \beta_2$ found by Collaborative SIR. It is evident the nonlinear behavior of the two link functions.

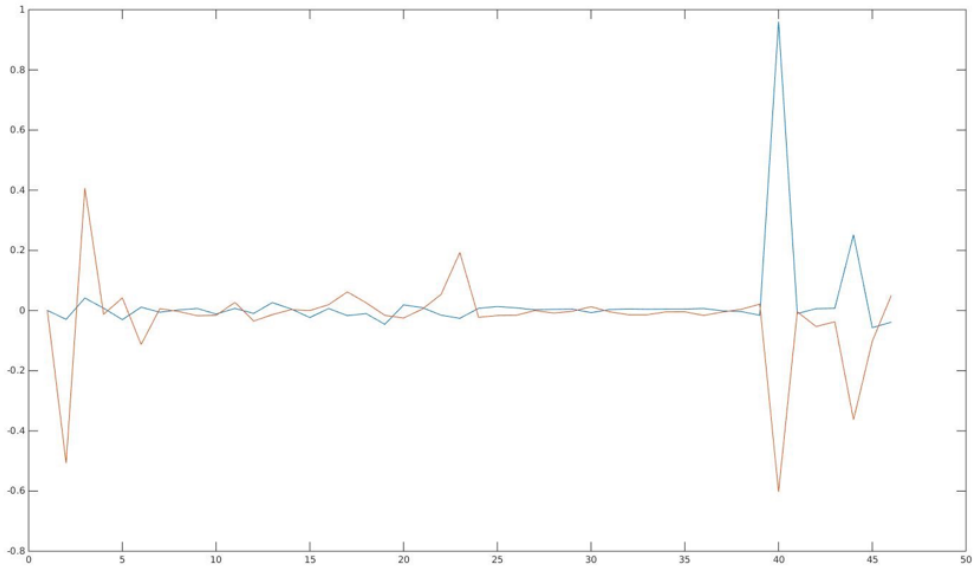


Figure 11: Differences in e.d.r directions $\hat{\beta}_1$ and $\hat{\beta}_2$. Many elements in the vectors are close to zero resulting in a variable selection. Differences in the two lines show how different variables contribute in regressing $Y$.

*5.3. Discussion on dimension k and the number of clusters c*

In the whole paper we presented results for dimension $k = 1$ ($Y = f(\beta_1^T X, ..., \beta_k^T X)$), the assumption is that e.d.r. spaces are one-dimensional. It is worth noticing that the entire approach can be easily extended to a higher $k$, it is sufficient to give a proximity measure between the linear subspaces (e.g. $Trace$ in [2]). But that might raise a question: it is worth it? SIR is a method to reduce dimensionality 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 2 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 i.e. that considering $k = 1$ is not a severe restriction if a visualization is needed. Finally another drawback of increasing dimensionality is that the samples become be more and more sparse and not cover enough the surface we want to regress, different regression techniques may lead to dramatically different results. The problem of dimension $k$ could be the reason why SIR is not yet widely used.
We did not give an automatic way of selecting the number of clusters. In SIR literature Kuentz and Saracco [12] translate the selection in an optimization problem. Nowadays, with the increasing capabilities of sensors, data are complex and complicated and 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" clustering, 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 [4] but has well established theory and simple implementation. If the link function turns out to be linear SIR, not surprisingly, is outperformed by linear regression techniques, but in case of

evidence of non linearity, linear regression techniques force the model resulting in poor estimations. Collaborative SIR is meant to deal with the increasing complexity of the dataset 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. A robustified version of SIR will be our main field of research for the next period.

**Acknowledgement**

The authors thank Didier Fraix-Burnet for his contribution to the data. They are grateful to Vanessa Kuentz and Jerome 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).

**References**

[1] Bernard-Michel, C., Gardes, L., Girard, S., 2009. Gaussian regularized sliced inverse regression. Statistics and Computing 19 (1), 85–98.

[2] Chavent, M., Girard, S., Kuentz-Simonet, V., Liquet, B., Nguyen, T. M. N., Saracco, J., 2014. A sliced inverse regression approach for data stream. Computational Statistics 29 (5), 1129–1152.

[3] Chavent, M., Kuentz, V., Liquet, B., Saracco, J., 2011. A sliced inverse regression approach for a stratified population. Communications in Statistics-Theory and Methods 40 (21), 3857–3878.

[4] Chen, C.-H., Li, K.-C., 1998. Can sir be as popular as multiple linear regression? Statistica Sinica 8 (2), 289–316.

[5] Chiaromonte, F., Martinelli, J., 2002. Dimension reduction strategies for analyzing global gene expression data with a response. Mathematical Biosciences 176 (1), 123–144.

[6] Cook, R. D., Weisberg, S., 2009. Applied regression including computing and graphics. Vol. 488. John Wiley & Sons.

[7] Diaconis, P., Freedman, D., 1984. Asymptotics of graphical projection pursuit. The Annals of Statistics 12 (3), 793–815.

[8] Duan, N., Li, K.-C., 1991. Slicing regression: a link-free regression method. The Annals of Statistics 19 (2), 505–530.

[9] Fukunaga, K., 2013. Introduction to statistical pattern recognition. Academic press.

[10] Hall, P., Li, K.-C., 1993. On almost linearity of low dimensional projections from high dimensional data. The Annals of Statistics 21 (2), 867–889.

[11] Hsing, T., Carroll, R. J., 1992. An asymptotic theory for sliced inverse regression. The Annals of Statistics 20 (2), 1040–1061.

[12] Kuentz, V., Saracco, J., 2010. Cluster-based sliced inverse regression. Journal of the Korean Statistical Society 39 (2), 251–267.

[13] Li, K.-C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86 (414), 316–327.

[14] Li, L., Cook, R. D., Nachtsheim, C. J., 2004. Cluster-based estimation for sufficient dimension reduction. Computational Statistics & Data Analysis 47 (1), 175–193.

[15] Li, L., Li, H., 2004. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics 20 (18), 3406–3412.

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

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

[18] Scrucca, L., 2006. Regularized sliced inverse regression with applications in classification. In: Data Analysis, Classification and the Forward Search. Springer, pp. 59–66.

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

[20] Soltanolkotabi, M., Elhamifar, E., Candes, E. J., 2014. Robust subspace clustering. The Annals of Statistics 42 (2), 669–699.

[21] Van der Maaten, L. J., Postma, E. O., van den Herik, H. J., 2009. Dimensionality reduction: A comparative review. Journal of Machine Learning Research 10 (1-41), 66–71.

[22] Zhong, W., Zeng, P., Ma, P., Liu, J. S., Zhu, Y., 2005. Rsir: regularized sliced inverse regression for motif discovery. Bioinformatics 21 (22), 4169–4175.

[23] Zhu, L.-P., 2010. Extending the scope of inverse regression methods in sufficient dimension reduction. Communications in Statistics-Theory and Methods 40 (1), 84–95.