Collaborative Sliced Inverse Regression

Rencontres d'Astrostatistique

13 Novembre 2014, Grenoble

Alessandro Chiancone

Stephane Girard

Jocelyn Chanussot

Sliced Inverse Regression

SIR limitations

Cluster SIR

Collaborative SIR

Motivation

Given a response continuous variable $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ explanatory variables we look for a functional relationship between the two:

$$Y = f(X, \epsilon), \ \epsilon \text{ independent of } X$$

Motivation

Given a response continuous variable $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ explanatory variables we look for a functional relationship between the two:

$$Y = f(X, \epsilon), \ \epsilon \text{ independent of } X$$

It is reasonable to think that X contains more information that what we need for our task of predicting Y, in addition when the dimension p is large regression is a very hard task when we do not have information about f.

Motivation

Given a response continuous variable $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ explanatory variables we look for a functional relationship between the two:

$$Y = f(X, \epsilon), \ \epsilon \text{ independent of } X$$

It is reasonable to think that X contains more information that what we need for our task of predicting Y, in addition when the dimension p is large regression is a very hard task when we do not have information about f.

Given that a possible solution is to make a strong assumption:

$$\boxed{Y = f(\beta_1^T X, \beta_2^T X, ..., \beta_k^T X, \epsilon)}$$

The space $S = Span\{\beta_1, \beta_2, ..., \beta_k\}$ is called effective dimension reduction (e.d.r.) space. Only few linear combinations of the predictors, $k \ll p$, are sufficient to regress Y.

Idea of Sliced Inverse Regression

Suppose k=1 for simplicity: $Y = f(\beta^T X, \epsilon)$

We want to find the direction β that best explains Y.

In other words if Y is fixed then $\beta^T X$ should not vary. Consequently our goal is to find a direction β which minimizes the variations of $\beta^T X$ given Y.

Idea of Sliced Inverse Regression

Suppose k=1 for simplicity: $Y = f(\beta^T X, \epsilon)$

We want to find the direction β that best explains Y.

In other words if Y is fixed then $\beta^T X$ should not vary. Consequently our goal is to find a direction β which minimizes the variations of $\beta^T X$ given Y.

To estimate $\beta^T X | Y$ we arrange Y in h slices each with the same number of samples, our goal is to find a direction β which minimizes the within-slice variance of $\beta^T X$ under the constraint $var(\beta^T X) = 1$.

Idea of Sliced Inverse Regression

Suppose k=1 for simplicity: $Y = f(\beta^T X, \epsilon)$

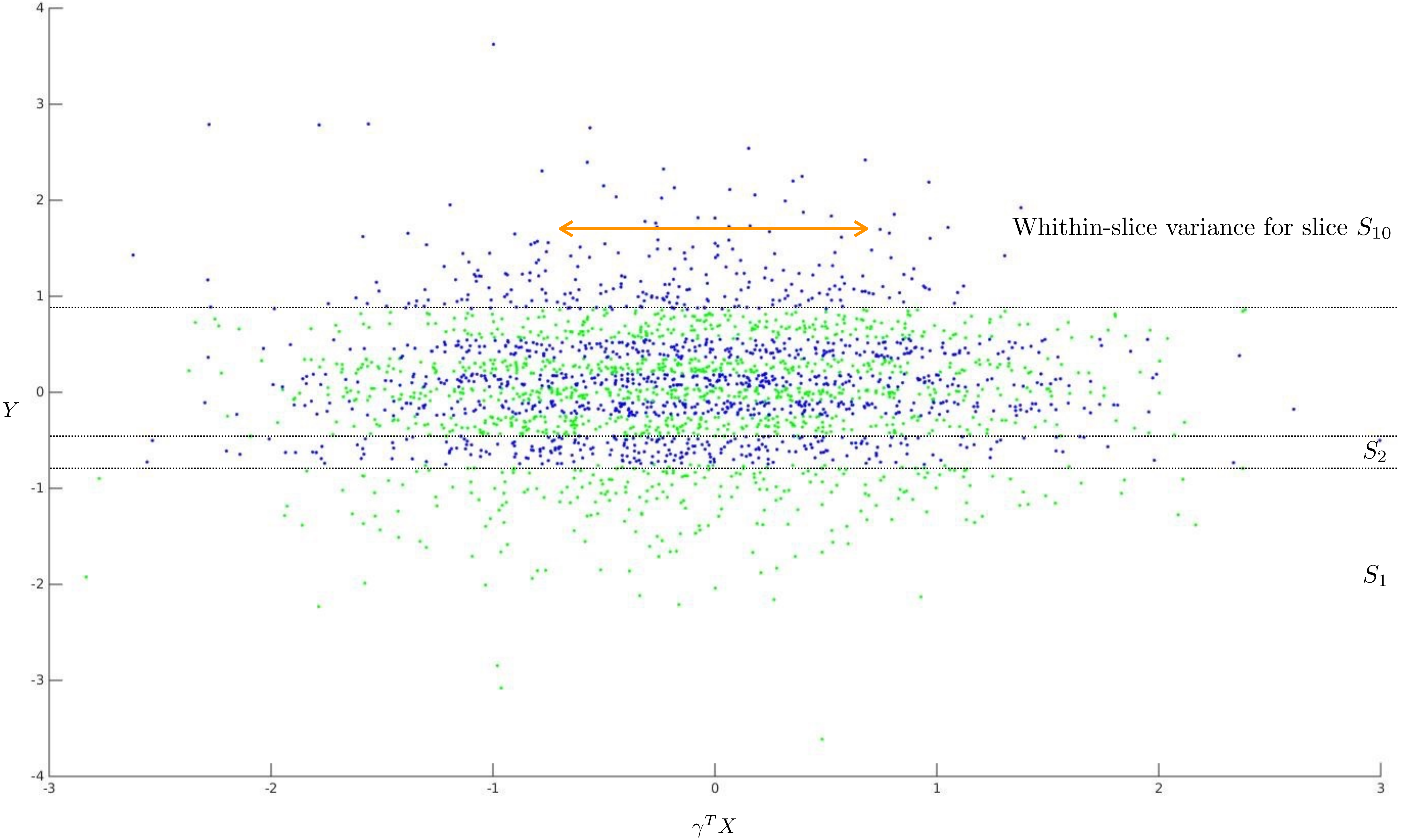
We want to find the direction β that best explains Y.

In other words if Y is fixed then $\beta^T X$ should not vary. Consequently our goal is to find a direction β which minimizes the variations of $\beta^T X$ given Y.

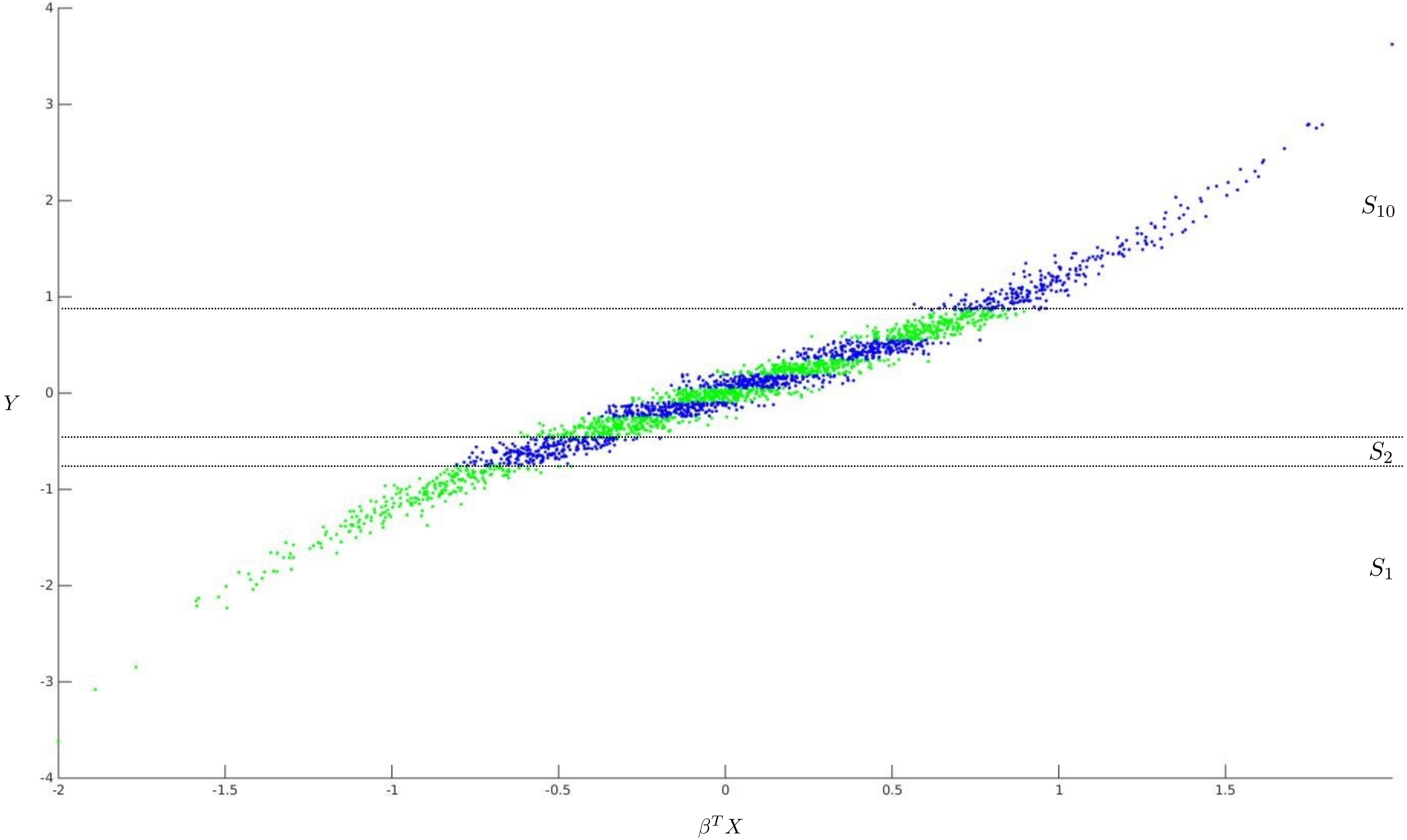
To estimate $\beta^T X|Y$ we arrange Y in h slices each with the same number of samples, our goal is to find a direction β which minimizes the within-slice variance of $\beta^T X$ under the constraint $var(\beta^T X) = 1$.

Since $\hat{\Sigma} = \hat{B} + \hat{W}$, where $\hat{\Sigma}, \hat{B}, \hat{W}$ are respectively the sample covariance matrix, the between-slice covariance matrix and the within-slice covariance matrix an equivalent approach is to maximize the between-slice variance under the same constraint.

Idea of Sliced Inverse Regression



Idea of Sliced Inverse Regression



Sliced Inverse Regression

Let us go back to the general case $Y = f(\beta_1^T X, \beta_2^T X, ..., \beta_k^T X, \epsilon)$. Since f is unknown it is not possible to directly retrieve $\beta_1, \beta_2, ..., \beta_k$ but a basis of the e.d.r. space S. Sliced Inverse Regression [1] solves this problem under the following so called Linearity Design Condition:

(LDC) $\forall b \in \mathbb{R}^p$ $\mathbb{E}(b^T X | \beta_1^T X, \beta_2^T X, ..., \beta_k^T X) = c_0 + c_1 \beta_1^T X + ... + c_k \beta_k^T X$ for some constants $c_0, ..., c_k$.

[1] LI, Ker-Chau. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 1991, 86.414: 316-327.

Sliced Inverse Regression

Let us go back to the general case $Y = f(\beta_1^T X, \beta_2^T X, ..., \beta_k^T X, \epsilon)$.
Since f is unknown it is not possible to directly retrieve $\beta_1, \beta_2, ..., \beta_k$ but a basis of the e.d.r. space S. Sliced Inverse Regression [1] solves this problem under the following so called Linearity Design Condition:

(LDC) $\forall b \in \mathbb{R}^p$ $\mathbb{E}(b^T X | \beta_1^T X, \beta_2^T X, ..., \beta_k^T X) = c_0 + c_1 \beta_1^T X + ... + c_k \beta_k^T X$ for some constants $c_0, ..., c_k$.

The centered inverse regression curve $\mathbb{E}(X|Y) - \mathbb{E}(X)$ is then contained in the linear subspace spanned by $\Sigma\beta_{i,i=1,...,k}$ where $\Sigma = Cov(X)$.

[1] LI, Ker-Chau. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 1991, 86.414: 316-327.

Sliced Inverse Regression

Let us go back to the general case $Y = f(\beta_1^T X, \beta_2^T X, ..., \beta_k^T X, \epsilon)$. Since f is unknown it is not possible to directly retrieve $\beta_1, \beta_2, ..., \beta_k$ but a basis of the e.d.r. space S. Sliced Inverse Regression [1] solves this problem under the following so called Linearity Design Condition:

(LDC) $\forall b \in \mathbb{R}^p$ $\mathbb{E}(b^T X | \beta_1^T X, \beta_2^T X, ..., \beta_k^T X) = c_0 + c_1 \beta_1^T X + ... + c_k \beta_k^T X$ for some constants $c_0, ..., c_k$.

The centered inverse regression curve $\mathbb{E}(X|Y) - \mathbb{E}(X)$ is then contained in the linear subspace spanned by $\Sigma\beta_{i,i=1,...,k}$ where $\Sigma = Cov(X)$.

This implies that the $\Gamma = Cov(\mathbb{E}(X|Y) - \mathbb{E}(X))$ is degenerated in any direction orthogonal to the directions $\Sigma\beta_{i,i=1,...,k}$ and furthermore that the k eigenvectors associated with the k largest eigenvalues are the $\Sigma\beta_{i,i=1,...,k}$.

[1] LI, Ker-Chau. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 1991, 86.414: 316-327.

SIR IN PRACTICE

SIR is quite simple to implement:

- Split Y in h slices
- Estimate Γ using the slices, $\hat{\Gamma}$ is the between-slice covariance matrix
- Compute the eigendecomposition of the matrix $\hat{\Sigma}^{-1}\hat{\Gamma}$, where $\hat{\Sigma}$ is the empirical covariance matrix of X
- Select the eigenvectors corresponding to the highest eigenvalues.

SIR LIMITATIONS

Recently many papers pointed out limitations of SIR since the eigendecomposition can be challenging when the covariance matrix Σ is ill conditioned. Many solutions have been proposed starting from preprocessing the data using PCA [1] to a more comprehensive approach to regularize SIR [2].

[1] LI, Lexin; LI, Hongzhe. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics, 2004, 20.18: 3406-3412.

[2] BERNARD-MICHEL, Caroline; GARDES, Laurent; GIRARD, Stéphane. Gaussian regularized sliced inverse regression. Statistics and Computing, 2009, 19.1: 85-98.

SIR LIMITATIONS

Recently many papers pointed out limitations of SIR since the eigendecomposition can be challenging when the covariance matrix Σ is ill conditioned. Many solutions have been proposed starting from preprocessing the data using PCA [1] to a more comprehensive approach to regularize SIR [2].

The weakest point of SIR is the (LDC) which cannot be verified in practice because it depends on the true directions $\beta_1, ..., \beta_k$. The condition holds in case of elliptic symmetry and more generally it has been shown [3] that if the dimension p tends to infinity the condition is always approximately verified.

[1] LI, Lexin; LI, Hongzhe. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics, 2004, 20.18: 3406-3412.

[2] BERNARD-MICHEL, Caroline; GARDES, Laurent; GIRARD, Stéphane. Gaussian regularized sliced inverse regression. Statistics and Computing, 2009, 19.1: 85-98.

[3] HALL, Peter; LI, Ker-Chau. On almost linearity of low dimensional projections from high dimensional data. The annals of Statistics, 1993, 867-889.

Cluster SIR

The LDC is verified when X follows an elliptically symmetric distribution (e.g. multi normality of X).

When X follows a Gaussian mixture model the condition does not globally hold but it is verified locally (i.e. in each mixture).

Kuentz & Saracco [1] clusterized X to force the condition to hold locally in each cluster and then combine the result in each cluster to obtain the final e.d.r. directions

[1] KUENTZ, Vanessa; SARACCO, Jérôme. Cluster-based sliced inverse regression. Journal of the Korean Statistical Society, 2010, 39.2: 251-267.

Cluster SIR

The LDC is verified when X follows an elliptically symmetric distribution (e.g. multi normality of X).

When X follows a Gaussian mixture model the condition does not globally hold but it is verified locally (i.e. in each mixture).

Kuentz & Saracco [1] clusterized X to force the condition to hold locally in each cluster and then combine the result in each cluster to obtain the final e.d.r. directions

---

Our work is based on this intuition. We first clusterize the predictor space X then a greedy merging algorithm is proposed to assign each cluster to its e.d.r space taking into account the size of the cluster on which SIR is performed.

[1] KUENTZ, Vanessa; SARACCO, Jérôme. Cluster-based sliced inverse regression. Journal of the Korean Statistical Society, 2010, 39.2: 251-267.

SIR - Cluster SIR

Collaborative SIR

$k = 1$

SIR - $Y = f(\beta^T X)$

Cluster SIR- $X = X_1 \cup X_2 \cup .... \cup X_c$, where c is the number of clusters

$Y_i = f(\beta^T X_i)$, $i = 1, ..., c$

The e.d.r. direction and the linking function do not change depending on the cluster.

SIR - Cluster SIR

Collaborative SIR

$k = 1$

SIR - $Y = f(\beta^T X)$

Cluster SIR- $X = X_1 \cup X_2 \cup .... \cup X_c$, where c is the number of clusters

$Y_i = f(\beta^T X_i)$, $i = 1, ..., c$

The e.d.r. direction and the linking function do not change depending on the cluster.

$X = X_1 \cup X_2 \cup ... \cup X_c$

$Y_i = f_i(\gamma_i^T X_i)$, $\gamma_i \in \{\beta_1, ..., \beta_D\}$

The number D $(D \leq c)$ of e.d.r. directions is unknown.

The e.d.r. directions and the link function may change depending on the cluster.
A merging algorithm is introduced to infer the number D based on the collinearity of the vectors β_i.

Collaborative SIR : Merging

The directions $\gamma_1, ..., \gamma_c$ are obtained applying SIR independently in each cluster.

A hierarchical structure is built to infer the number D of e.d.r. directions using a proximity criteria.

The most collinear vector to a set of vectors $A = \{\gamma_1, ..., \gamma_c\}$ given the proximity criterion $m(a,b) = cos^2(a,b) = (a^T b)^2$ is the solution of the following problem:

$$\lambda(A) = \max_{a \in \mathbb{R}^p} \sum_{\gamma_t \in A} w_t m(\gamma_t, a) \text{ s.t. } \|a\| = 1$$

$$= \text{largest eigenvalue of } \sum_{\gamma_t \in A} w_t \gamma_t \gamma_t^T$$

where w_t are weights and sum to one.

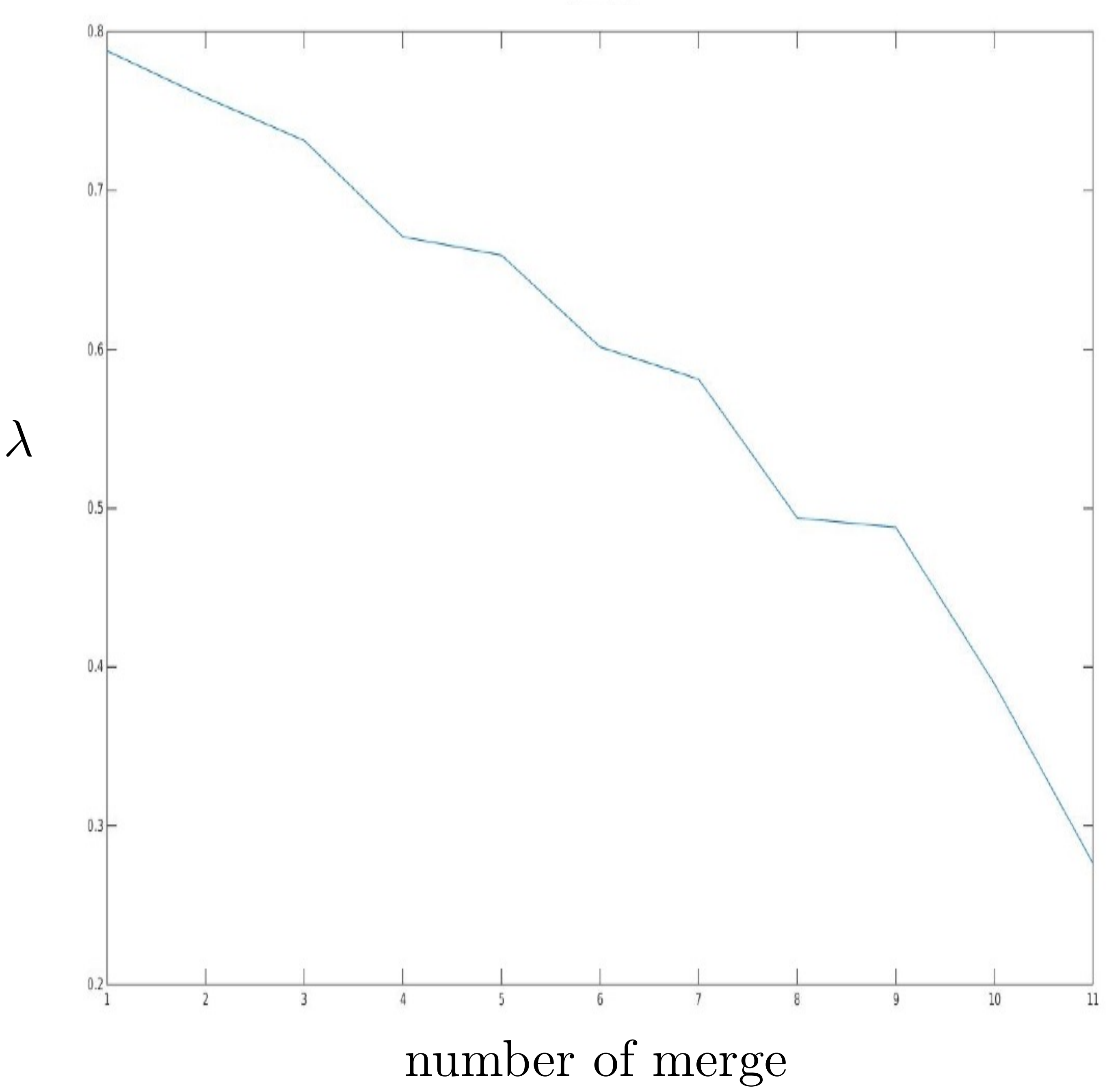
Merging

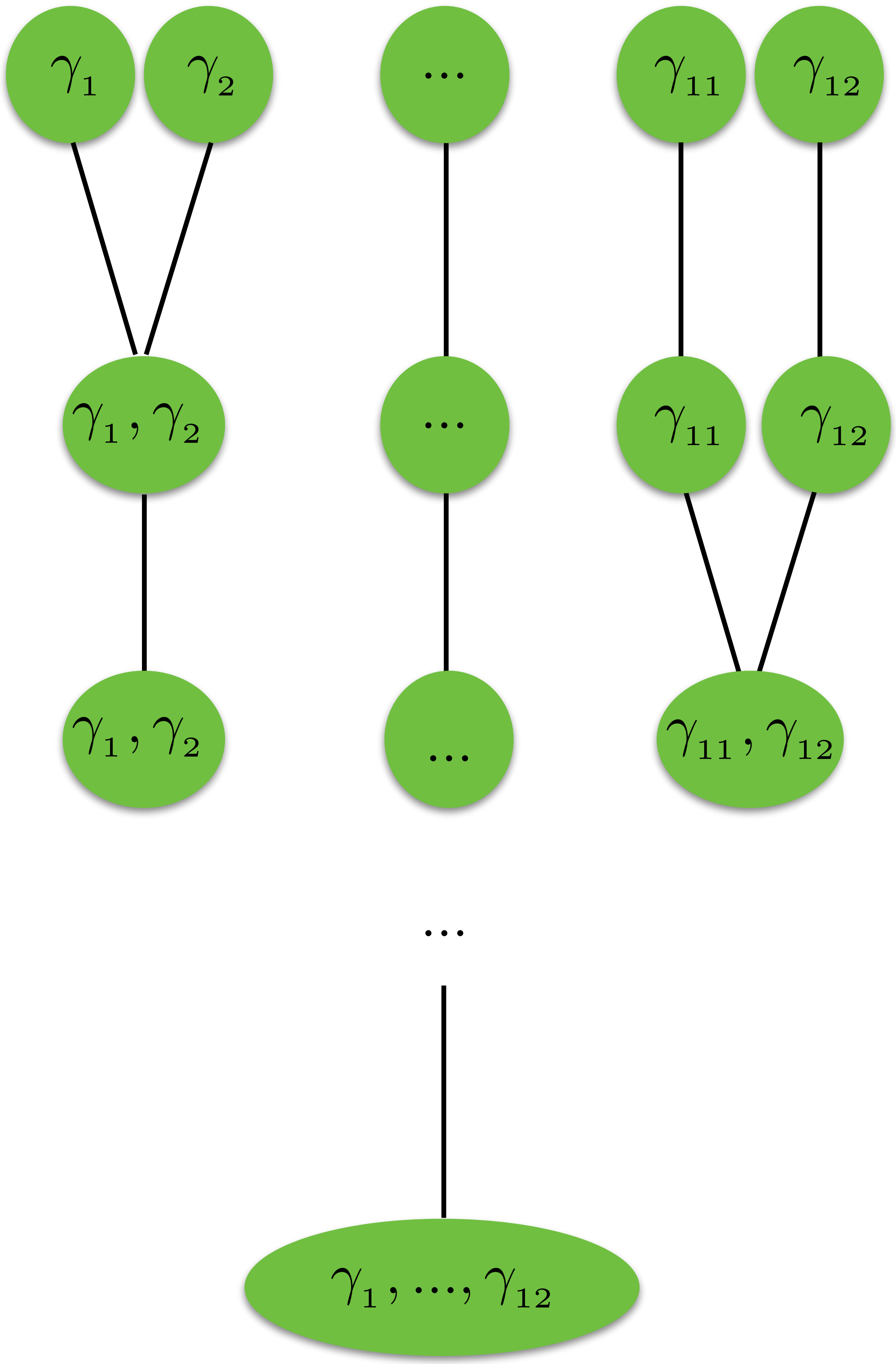
To build the hierarchy we consider the following iterative algorithm initialized with the set $A = \{\{\gamma_1\}, ..., \{\gamma_c\}\}$:

```
while $card(A) \neq 1$
Let $a, b \in A$ such that $\lambda(a \cup b) > \lambda(c \cup d) \forall c, d \in A$
$A = (A \setminus \{a, b\}) \cup a \cup b$
end
```

at each step the cardinality of the set A decreases merging the most collinear sets of directions. Therefore it is possible to infer the number D of underlying e.d.r. spaces analyzing the values of λ in the hierarchy.

Merging


0.8
0.7
0.6
0.5
0.4
0.3
0.2
λ
1
2
3
4
5
6
7
8
9
10
11
number of merge


γ_1
γ_2
...
γ_{11}
γ_{12}
γ_1, γ_2
...
γ_{11}
γ_{12}
γ_1, γ_2
...
γ_{11}, γ_{12}
...
$\gamma_1, ..., \gamma_{12}$

Regularization

After merging, each cluster is assigned one of the D e.d.r. directions $\hat{\beta}_1, ..., \hat{\beta}_D$

For each X_i, $i = 1, ..., c$ we consider the D e.d.r directions and we analyze the two-dimensional distributions:

$(Y_i, \hat{\beta}_j^T X_i)$ $\forall j = 1, ..., D$

Then we select the direction $\hat{\beta}_{j^*}$, $j^* = \min_{j=1,\ldots,D} \lambda_{2,j}$, where $\lambda_{2,j}$ is the second eigenvalue of the covariance matrix $Cov(Y_i, \hat{\beta}_j^T X_i)$.

This step reconsiders the data to select the best direction from the pool of the D estimated directions

Experimental Results

We simulated 100 different datasets following a gaussian mixture model:

- 10 mixture components
- uniform mixing proportions
- 2500 samples
- dimension $p = 240$

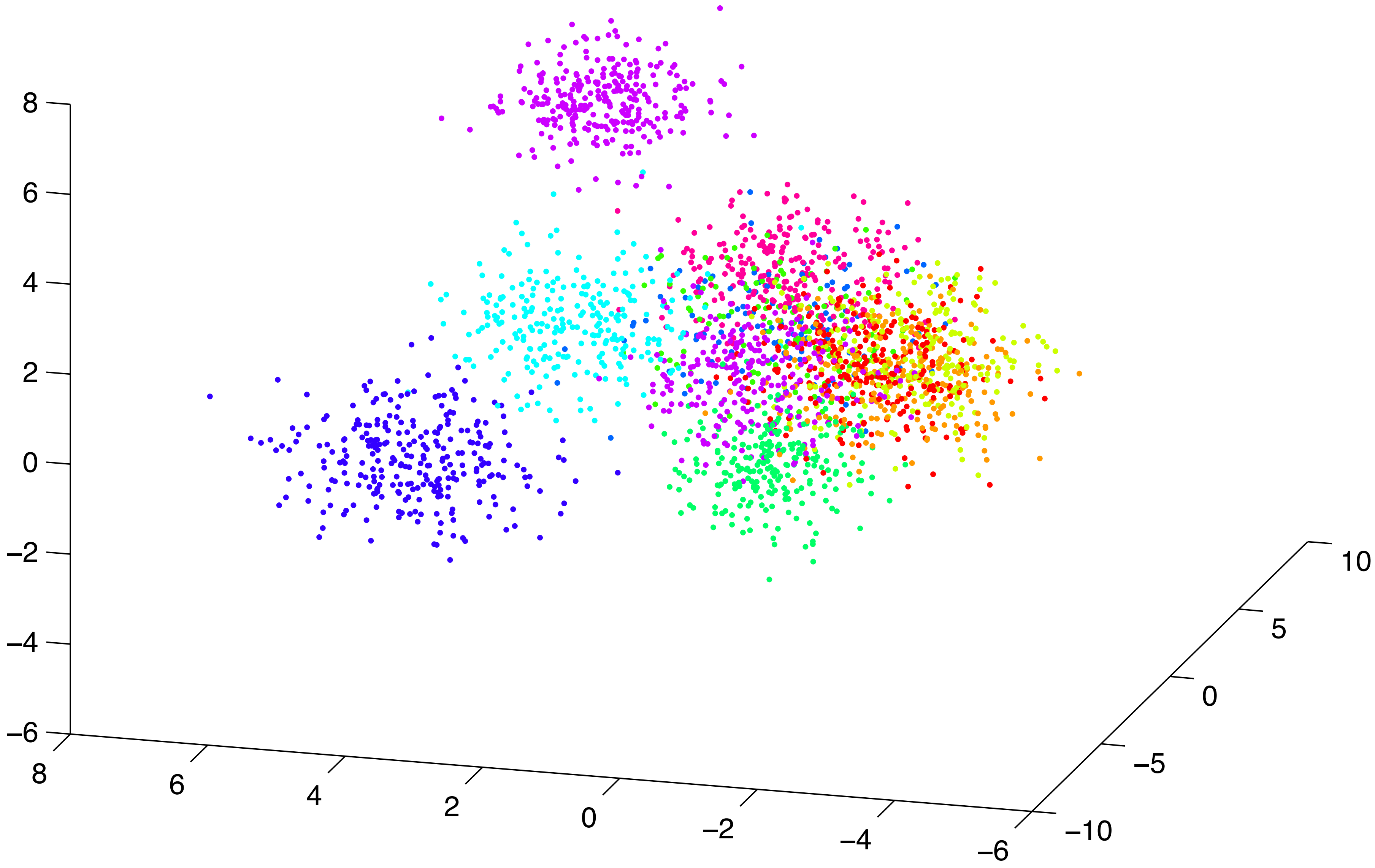
The response variable Y is generated using the hyperbolic sin link function: $Y_i = \sinh(\gamma_i^T X_i) + \epsilon$, ϵ independent of X.

We will show the case where $\gamma_i \in \{\beta_1, \beta_2\}$ i.e. the number of e.d.r. directions $D = 2$

The proximity criteria $m(\hat{\beta}, \beta) = \cos^2(\hat{\beta}, \beta) = (\hat{\beta}^T \beta)^2$ is evaluated to assess the quality of the estimation [1]

[1] CHAVENT, Marie, et al. A sliced inverse regression approach for data stream. Computational Statistics, 2013, 1-24.

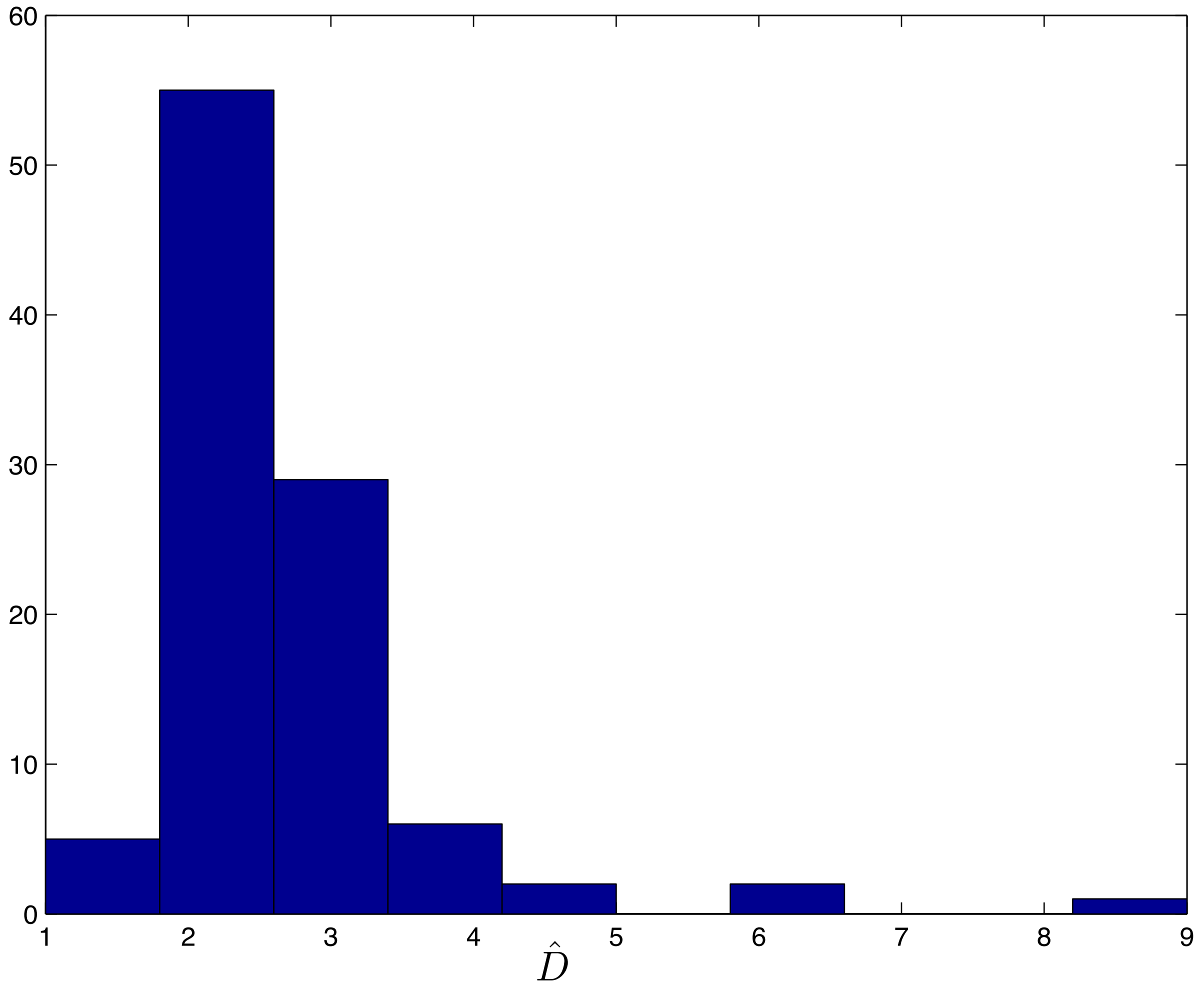
Experimental Results

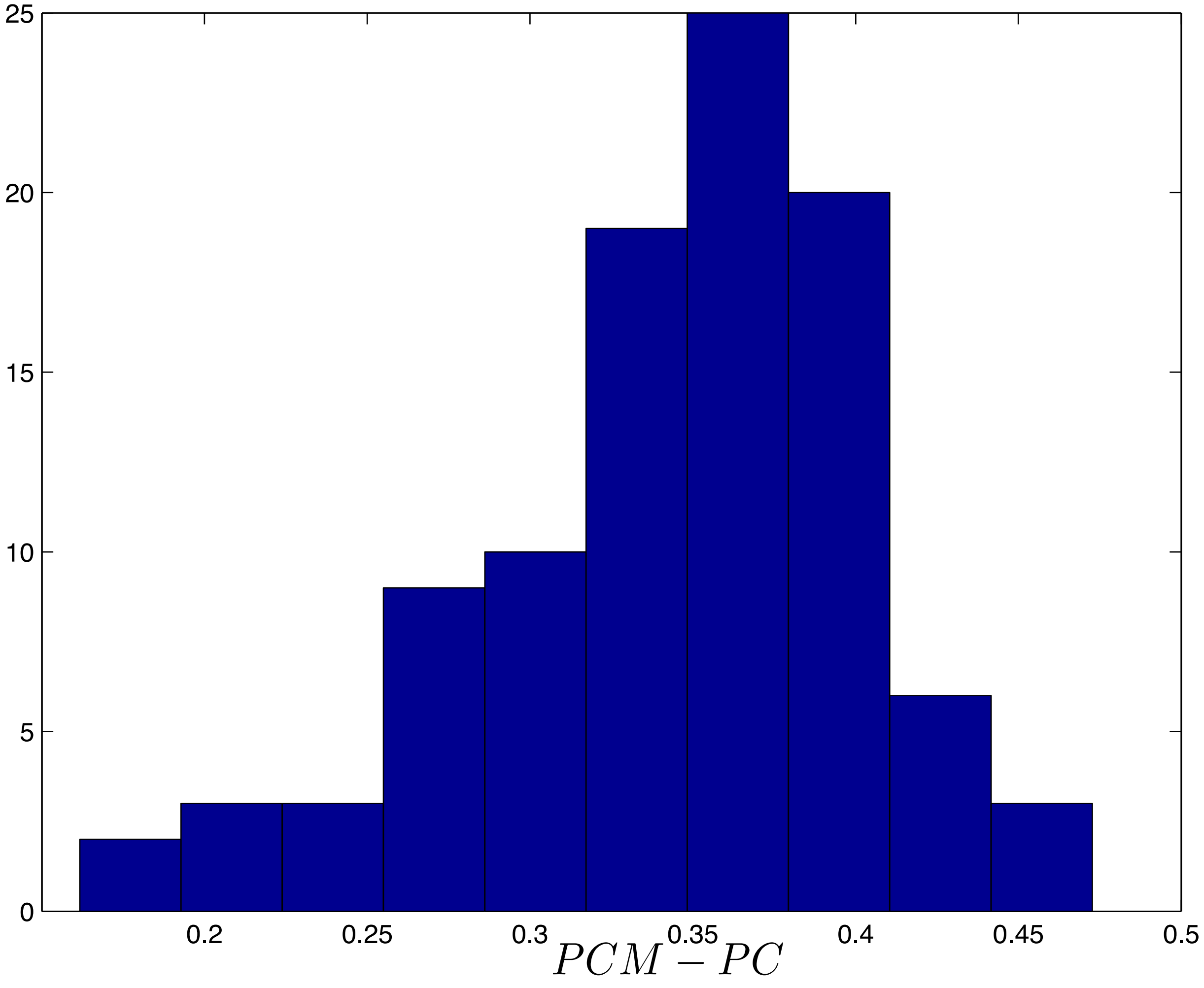


Experimental Results

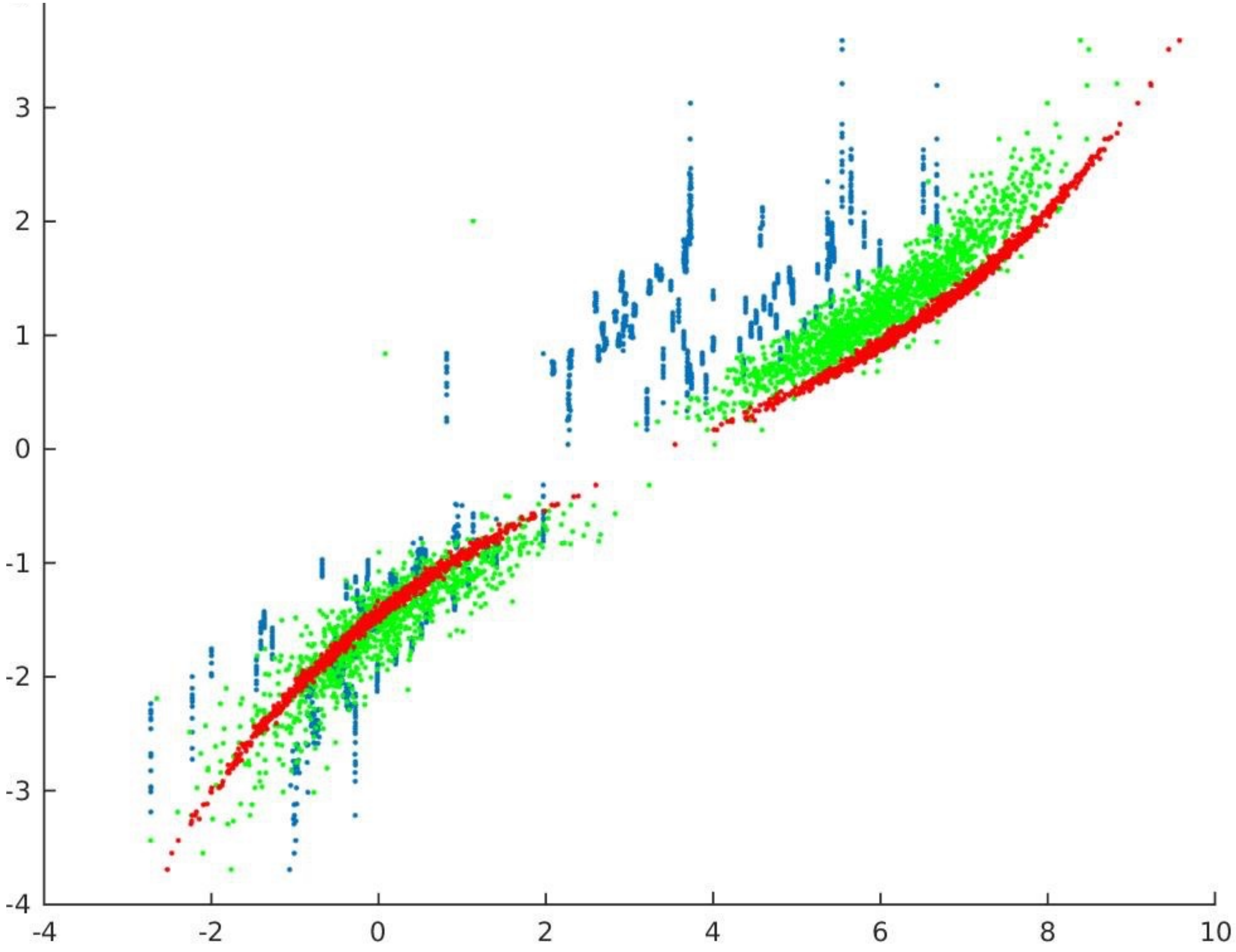
The proximity criteria is computed using the estimation obtained in each cluster independently, PC, and after collaborative SIR, PCM, for each run of K-means:

The average PC is 0.4499 with a standard deviation of 0.0690
The average PCM is 0.7939 with a standard deviation of 0.0960





Experimental Results



Experimental Results

We consider the data example of horse mussels present in cluster SIR. The observations correspond to $n = 192$ horse mussels captured in the Malborough Sounds at the Northeast of New Zealand's South Island. The response variable Y is the muscle mass, the edible portion of the mussel, in grams. The predictor X is of dimension $p = 4$ and measures numerical characteristics of the shell: length, width, height, each in mm, and mass in grams.

We repeated the following algorithm 100 times:

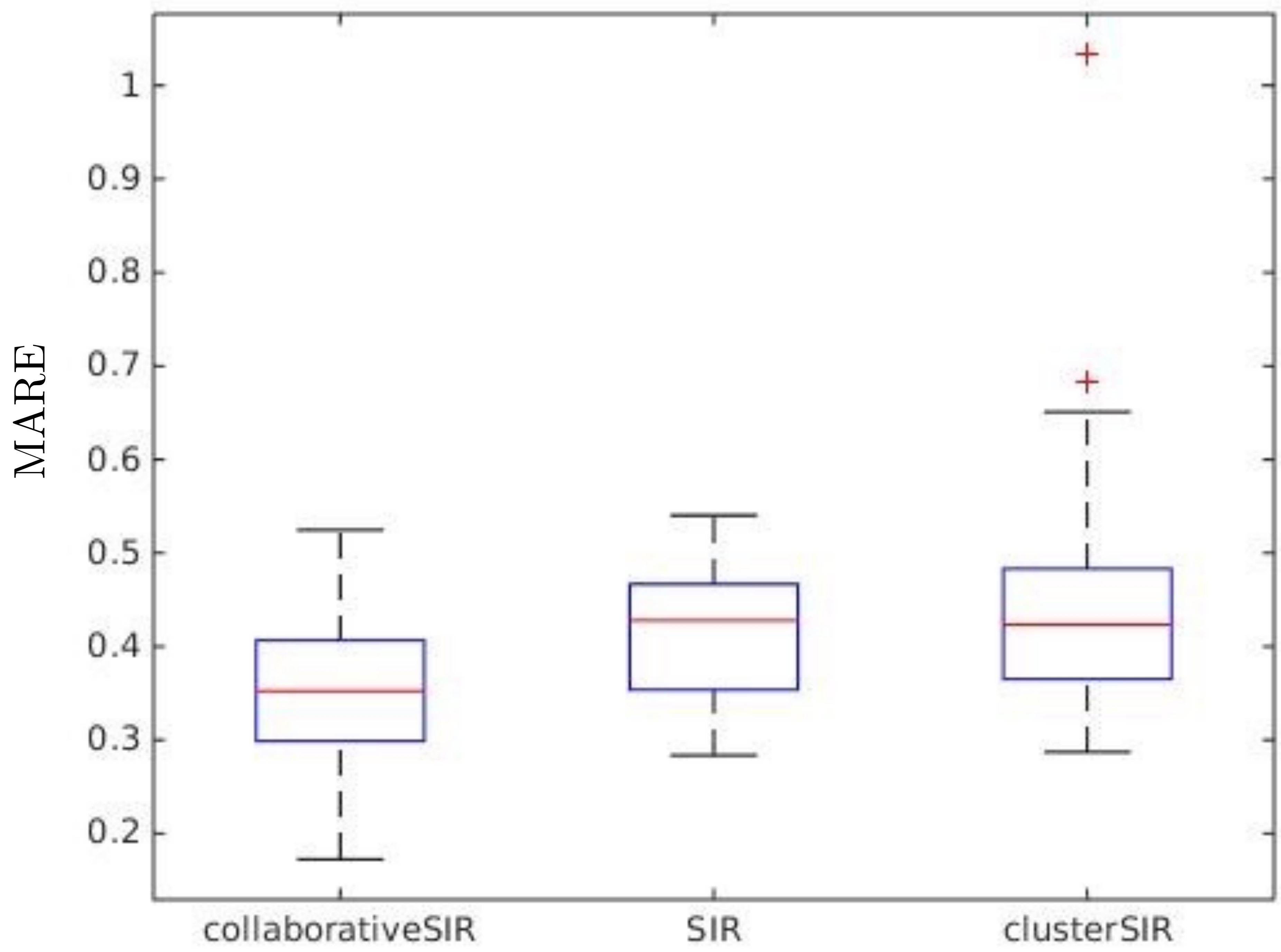
(1) Randomly select 80% of training and 20% of test.

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

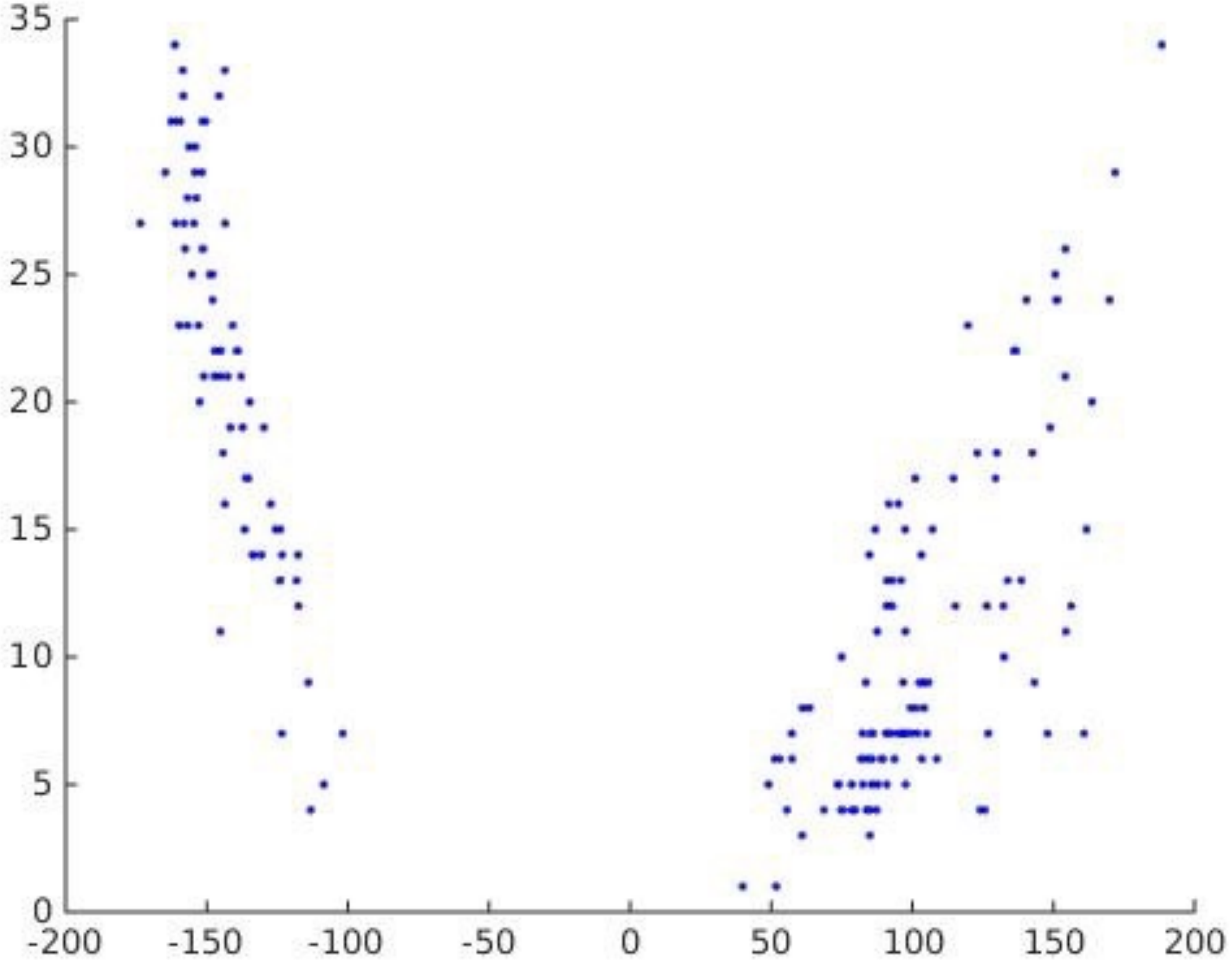
(3) Regress the functions using the training samples

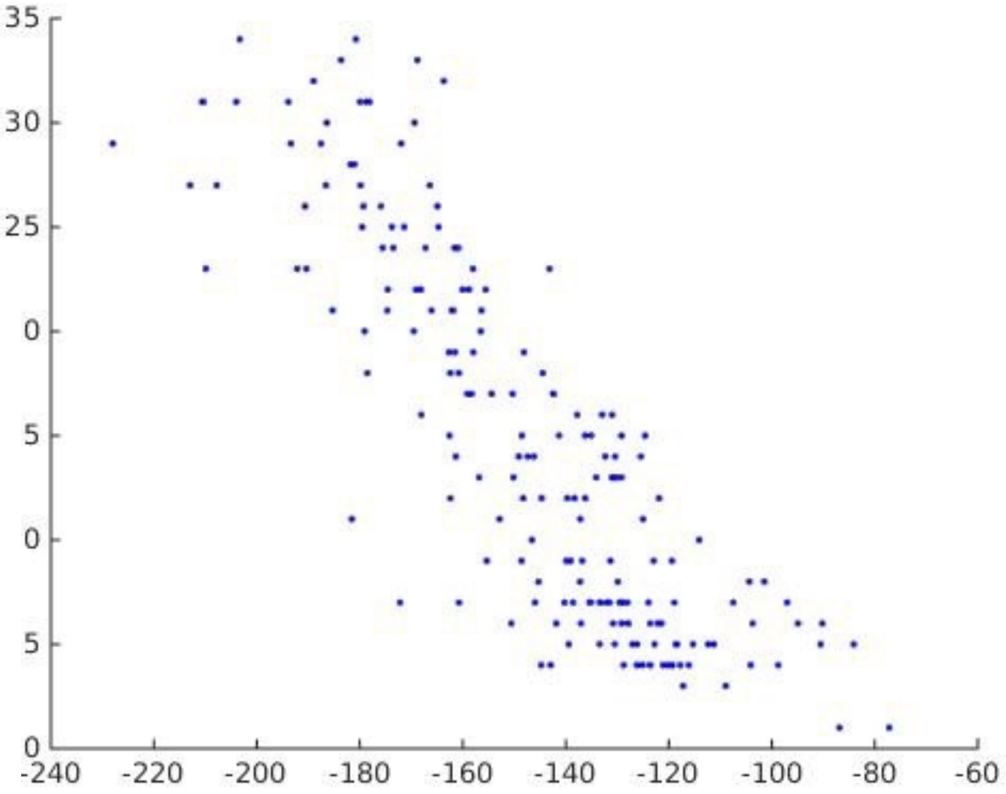
(4) Compute the Mean Absolute Relative Error (MARE) on the test

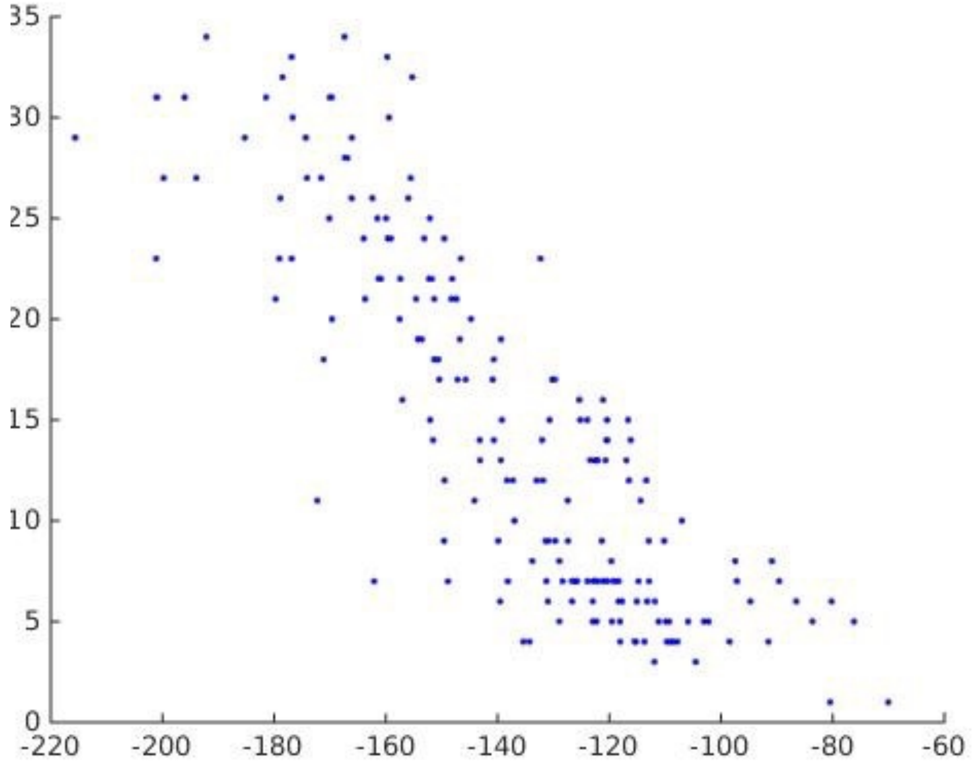
Experimental Results

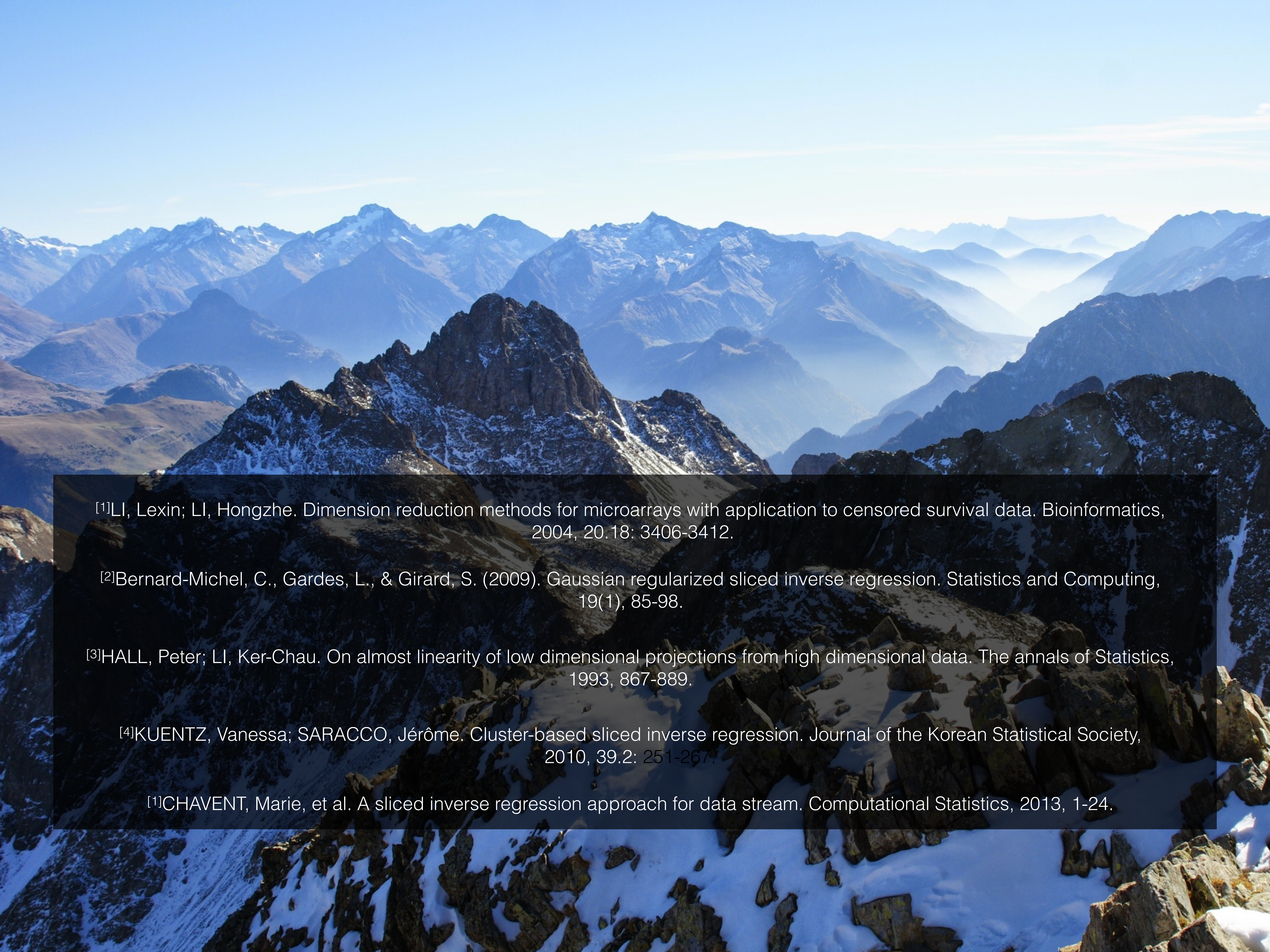


Experimental Results









[1]LI, Lexin; LI, Hongzhe. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics, 2004, 20.18: 3406-3412.

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

[3]HALL, Peter; LI, Ker-Chau. On almost linearity of low dimensional projections from high dimensional data. The annals of Statistics, 1993, 867-889.

[4]KUENTZ, Vanessa; SARACCO, Jérôme. Cluster-based sliced inverse regression. Journal of the Korean Statistical Society, 2010, 39.2: 251-267.

[1]CHAVENT, Marie, et al. A sliced inverse regression approach for data stream. Computational Statistics, 2013, 1-24.