Student Sliced Inverse Regression

Florence Forbes

Team Mistis, INRIA Rhône-Alpes, France
http://mistis.inrialpes.fr/~forbes

December 2016

Joint work with Alessandro Chiancone and Stéphane Girard
1 Sliced Inverse Regression (SIR)
2 Gaussian inverse regression
3 Student inverse regression
4 Validation on simulations
5 Real data study
Outline

1. Sliced Inverse Regression (SIR)
2. Gaussian inverse regression
3. Student inverse regression
4. Validation on simulations
5. Real data study
High dimensional regression

- Given two r.v. $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$, estimate $G : \mathbb{R}^p \to \mathbb{R}$ such that

$$Y = G(X) + \xi$$

where $\xi$ is independent of $X$.

- When $p$ is large, curse of dimensionality.

Natural solution: reduce the dimension of $X$ with a PCA on $X$ but does not take $Y$ into account.
Sufficient dimension reduction

Given two r.v. $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$, $G : \mathbb{R}^p \to \mathbb{R}$ such that

$$Y = G(X) + \xi$$

where $\xi$ is independent of $X$.

Sufficient dimension reduction aims at replacing $X$ by its projection onto a subspace of smaller dimension without loss of information on the distribution of $Y$ given $X$.

The central subspace is the smallest subspace $S$ such that, conditionally on the projection of $X$ on $S$, $Y$ and $X$ are independent:

$$Y \perp X \mid \pi_S(X)$$
Assume $\dim(S) = 1$ for the sake of simplicity, *i.e.* $S = \text{span}(b)$, with $b \in \mathbb{R}^p$ $\implies$ **Single index model**:

$$Y = g(b^t X) + \xi$$

where $\xi$ is independent of $X$.

The estimation of a $p-$ variate function $G$ is replaced by the estimation of a univariate function $g$ and of an axis $b$.

**Goal of SIR** [Li, 1991] : to estimate a basis of the central subspace (*i.e.* $b$ in this case).
SIR : Basic principle

Idea :
- Find the direction $b$ such that $b^t X$ best explains $Y$.
- Conversely, when $Y$ is fixed, $b^t X$ should not vary.
- Find the direction $b$ minimizing the variations of $b^t X$ given $Y$.

In practice :
- The range of $Y$ is partitioned into $h$ slices $S_j$.
- Minimize the within slice variance of $b^t X$ under the normalization constraint $\text{var}(b^t X) = 1$.
- Equivalent to maximizing the between slice variance under the same constraint.

$\implies$ intuitively PCA on $E[X|Y = y]$ the inverse regression curve
SIR: Illustration
SIR : Estimation procedure

Given a sample \( \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \), the direction \( b \) is estimated by

\[
\hat{b} = \arg\max_{b} b^t \hat{\Gamma} b \quad \text{u.c.} \quad b^t \hat{\Sigma} b = 1. \quad (1)
\]

where \( \hat{\Sigma} \) is the estimated covariance matrix of \( X \) and \( \hat{\Gamma} \) is the between slice covariance matrix defined by

\[
\hat{\Gamma} = \sum_{j=1}^{h} \frac{n_j}{n} (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t, \quad \bar{X}_j = \frac{1}{n_j} \sum_{Y_i \in S_j} X_i,
\]

with \( n_j \) is proportion of observations in slice \( S_j \). The optimization problem (1) has an explicit solution : \( \hat{b} \) is the eigenvector of \( \hat{\Sigma}^{-1} \hat{\Gamma} \) associated to its largest eigenvalue.
**Experimental set-up:** \( \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) with \( n = 100 \)

\( X_i \sim \mathcal{N}_p(0, \Sigma) \) and \( Y_i = g(b^t X_i) + \xi \) where \( g \) is the link function \( g(t) = \sin(\pi t/2) \), \( b \) is the true direction, \( \xi \sim \mathcal{N}_1(0, 9 \times 10^{-4}) \)

- **Blue:** Projections \( b^t X_i \) on the true direction \( b \) versus \( Y_i \),
- **Red:** Projections \( \hat{b}^t X_i \) on the estimated direction \( \hat{b} \) versus \( Y_i \),
- **Green:** \( b^t X_i \) versus \( \hat{b}^t X_i \).

**Note:** Once \( b \) is estimated, use your favorite regression method to estimate \( g \)

\( \Rightarrow \) **SIR** is a "model free" method
Outline

1. Sliced Inverse Regression (SIR)
2. Gaussian inverse regression
3. Student inverse regression
4. Validation on simulations
5. Real data study
Single-index inverse regression model

Model introduced in [Cook, 2007].

\[ X = \mu + c(Y)Vb + \varepsilon, \]  

(2)

where

- \( \mu \) and \( b \) are non-random \( \mathbb{R}^p \)-vectors,
- \( \varepsilon \sim \mathcal{N}_p(0, V) \), independent of \( Y \),
- \( c: \mathbb{R} \to \mathbb{R} \) is a nonrandom coordinate function.

If \( c(.) \) is decomposed on \( h \) basis functions \( s_j(.) \),

\[ c(.) = \sum_{j=1}^{h} c_j s_j(.) = s^t(.) c, \]

where \( c = (c_1, \ldots, c_h)^t \) is unknown and \( s(.) = (s_1(.), \ldots, s_h(.))^t \), it follows

\[ X = \mu + s^t(Y)cVb + \varepsilon, \quad \varepsilon \sim \mathcal{N}_p(0, V), \]
Maximum Likelihood estimation of \( \{\mu, c, V, b\} \)

**Notation:**

- **\( W \):** the \( h \times h \) empirical covariance matrix of \( s(Y) \) defined by
  \[
  W = \frac{1}{n} \sum_{i=1}^{n} (s(Y_i) - \bar{s})(s(Y_i) - \bar{s})^t \quad \text{with} \quad \bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(Y_i).
  \]

- **\( M \):** the \( h \times p \) matrix defined by
  \[
  M = \frac{1}{n} \sum_{i=1}^{n} (s(Y_i) - \bar{s})(X_i - \bar{X})^t,
  \]

If \( W \) and \( \hat{\Sigma} \) are regular, then the ML estimators are:

- **Direction:** \( \hat{b} \) is the eigenvector associated to the largest eigenvalue \( \hat{\lambda} \) of \( \hat{\Sigma}^{-1} M^t W^{-1} M \),

- **Coordinate:** \( \hat{c} = W^{-1} M \hat{b} / \hat{b}^t \hat{V} \hat{b} \),

- **Location parameter:** \( \hat{\mu} = \bar{X} - \bar{s}^t \hat{c} \hat{V} \hat{b} \),

- **Covariance matrix:** \( \hat{V} = \hat{\Sigma} - \hat{\lambda} \hat{\Sigma} \hat{b} \hat{b}^t \hat{\Sigma} / \hat{b}^t \hat{\Sigma} \hat{b} \),
In the particular case of **piecewise constant basis functions**

\[ s_j(.) = \mathbb{I}\{. \in S_j\}, \ j = 1, \ldots, h, \]

standard calculations show that

\[ M^t W^{-1} M = \hat{\Gamma} \]

and thus the **ML estimator** \( \hat{b} \) of \( b \) is the eigenvector associated to the largest eigenvalue of \( \hat{\Sigma}^{-1} \hat{\Gamma} \).

\[ \Rightarrow \text{SIR method.} \]
1. Sliced Inverse Regression (SIR)

2. Gaussian inverse regression

3. Student inverse regression

4. Validation on simulations

5. Real data study
Student distributed error

Standard SIR is intrinsically Gaussian
\[ \rightarrow \text{sensitive to outliers due to light tails} \]

Increase robustness by considering an heavy tailed error term \( \varepsilon \):

**Generalized Student distribution**

\[ S_p(\varepsilon; \mu, V, \alpha) = \frac{\Gamma(\alpha+p/2)}{|\Sigma|^{1/2} \Gamma(\alpha)(2\pi)^{p/2}} \left[ 1 + \frac{\delta(\varepsilon, \mu, \Sigma)/(2)}{(2)} \right]^{-(\alpha+p/2)} \]

- heavy tailed
- **tractable** via a hierarchical representation (Gaussian scale mixture) and EM algorithm
Multi-index Student inverse regression model

\[ X = \mu + V B c(Y) + \varepsilon, \]  \hspace{1cm} (3)

- \( \mu \in \mathbb{R}^p \) and \( B \) a \( p \times d \) matrix with \( B B^T = I_d \),
- \( \varepsilon \sim \mathcal{S}_p(0, V, \alpha) \), independent of \( Y \),
- \( c : \mathbb{R} \rightarrow \mathbb{R}^d \) is a nonrandom coordinate function.

**Proposition**: \( B \) corresponds to the direction of the central subspace (up to a linear full rank transformation).

\[ c(.) = (c_1(.) \ldots c_d(.)), \text{ with } c_k(.) = \sum_{j=1}^{h} c_{jk} s_j(\cdot) = s^T(\cdot)c \]

\( \implies C \) is a \( h \times d \) matrix and (3) can be rewritten as

\[ X = \mu + V B C^T s(Y) + \varepsilon \text{ with } \varepsilon \sim \mathcal{S}_p(0, V, \alpha) \]

\( \theta = \{\mu, V, B, C, \alpha\} \) to be estimated
Maximum likelihood via EM algorithm

Given a sample \( \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \)
Use Gaussian scale mixture representation of the \( t \)-distribution, introducing additional latent variables \( U_1, \ldots, U_n \),

\[
(X_i | Y_i) \sim \mathcal{S}_p(\mu + VBC^T s_i, V, \alpha)
\]

where \( s_i = s(Y_i) \)

is equivalent to

\[
X_i | U_i = u_i, Y_i = y_i \sim \mathcal{N}_p(\mu + VBC^T s_i, V/u_i),
\]
\[
U_i | Y_i = y_i \sim \mathcal{G}(\alpha, 1).
\]
Alternate E and M steps.

E-step: \( \bar{u}_i(t) = E_{U_i}[U_i | X_i, Y_i; \theta^{(t-1)}] \) and \( \tilde{u}_i(t) = E_{U_i}[\log U_i | X_i, Y_i; \theta^{(t-1)}] \)

\( \tilde{u}_i(t) \) acts as a weight for \( X_i, Y_i \).

M-step: use "weighted versions" of matrices \( \hat{\Sigma}, W \), etc. If \( W \) and \( \hat{\Sigma} \) regular,

- Directions: \( \hat{B} \) is the eigenvectors associated to the largest eigenvalues of \( \hat{\Sigma}^{-1} M^T W^{-1} M \),
- Covariance matrix: \( \hat{V} = \hat{\Sigma} - (M^T W^{-1} M \hat{B})(\hat{B}^T M^T W^{-1} M \hat{B})^{-1}(M^T W^{-1} M \hat{B})^T \),
- Coordinates: \( \hat{C} = W^{-1} M \hat{B}(\hat{B}^T \hat{V} \hat{B})^{-1} \) and
- Location parameter: \( \hat{\mu} = \bar{X} - \hat{V} \hat{B} \hat{C}^T \bar{s} \).

When: \( s_j(.) = \mathbb{I} \{ . \in S_j \}, j = 1, \ldots, h, \implies \text{Student SIR algorithm} \)
EM algorithm: notation

- $W$ : the $h \times h$ weighted covariance matrix $W$ of $s(Y)$

\[
W = \frac{1}{n} \sum_{i=1}^{n} \bar{u}_i (s_i - \bar{s})(s_i - \bar{s})^T,
\]

- $M$ : the $h \times p$ weighted covariance matrix $M$ of $(s, X)$

\[
M = \frac{1}{n} \sum_{i=1}^{n} \bar{u}_i (s_i - \bar{s})(X_i - \bar{X})^T,
\]

- and $\Sigma$ the $p \times p$ weighted covariance matrix of $X$

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \bar{u}_i (X_i - \bar{X})(X_i - \bar{X})^T,
\]

with $\bar{X} = \frac{1}{\sum_{i=1}^{n} \bar{u}_i} \sum_{i=1}^{n} \bar{u}_i X_i$ and $\bar{s} = \frac{1}{\sum_{i=1}^{n} \bar{u}_i} \sum_{i=1}^{n} \bar{u}_i s_i$. 
Graphical considerations, e.g. [Liquet et al 2012] : not quantitative.
Cross validation : $d$ may vary depending on the regression approach selected.
Tests : most approaches.
Penalized likelihood criterion [Zhu et al. 2006] : the most natural in our setting.

Bayesian information criterion :

\[ BIC(d) = -2L(d) + \eta \log n , \]
where \( \eta = \frac{p(p+3)}{2} + 1 + \frac{d(2p-d-1+2h)}{2} \)

BIC provides correct selections but requires large enough sample sizes.
1. Sliced Inverse Regression (SIR)

2. Gaussian inverse regression

3. Student inverse regression

4. Validation on simulations

5. Real data study
Validation on simulations

**Proximity criterion** between the true directions $B$ and the estimated ones $\hat{B}$:

$$r(B, \hat{B}) = \frac{\text{trace}(BB^T \hat{B} \hat{B}^T)}{d}$$

evaluates the distance between the subspaces spanned by the columns of $B$ and $\hat{B}$

- $0 \leq r \leq 1$,
- a value close to 0 implies a low proximity. If $d = 1$, $r$ is the squared cosine between the two spanning vectors: $\hat{b}$ is nearly orthogonal to $b$,
- a value close to 1 implies a high proximity.

**Results**: Student SIR shows good performance, outperforming SIR when the distribution of $X$ is heavy-tailed and preserving good properties such as insensitivity to the number of slices.
Outline

1. Sliced Inverse Regression (SIR)
2. Gaussian inverse regression
3. Student inverse regression
4. Validation on simulations
5. Real data study
Galaxy data

Data:

- \( n = 362,887 \) different galaxies (all the original observations are considered)
- The response variable \( Y \) is the stellar formation rate.
- The predictor \( X \) is made of spectral characteristics of the galaxies and is of dimension \( p = 46 \).
- True central space unknown

Evaluation setting:

- 1000 random subsets of \( X \) of size \( n = 30,000 \)
- \( h = 100 \)
- Reference results computed on the whole data set with \( d = 3 \) (BIC): \( \hat{B}^{\text{SIR}}, \hat{B}^{\text{st-SIR}} \)
  with \( r(\hat{B}^{\text{SIR}}, \hat{B}^{\text{st-SIR}}) = 0.95 \) (almost same central space)
$r_{i}^{SIR}$ vs. $r_{i}^{st-SIR}$: almost all points are lying above the line $y = x$ indicating that Student SIR improves SIR results and significantly so for the subsets in the upper left corner.
Conclusion and future work

Non Gaussian SIR based on intrinsic inverse regression representation of SIR

- Maximum likelihood setting
- Alternative to robust estimators (Median, etc.)
- Higher computational cost than SIR due to EM iterations

Future work:

- Case $p > n$ still problematic due to inversion of large covariance matrices $\rightarrow$ regularization possible
- Selection of the central subspace dimension $d$ when $n$ is not large enough
- Extension to multivariate responses

Paper & Matlab code available at [https://hal.inria.fr/hal-01294982](https://hal.inria.fr/hal-01294982)
SIR and regularized SIR references


