Sélection de variables en régression SIR (sliced inverse regression) par seuillage doux/dur de la matrice d'intérêt

Hadrien Lorenzo[1,3] & Jérôme Saracco[1,2,3] & Clément Weinreich[1,2]

hadrien.lorenzo@u-bordeaux.fr

[1] *ASTRAL Team, Inria, Talence*
[3] *OptimAI team, IMB, CNRS UMR 5251*

Tuesday June 16th 2022

SIR, a semi-parametric model

Theoretical context : The semi-parametric single index model from Duan and Li 1991 as

$$y = f(\beta' x) + \epsilon \quad (1)$$

where:

- y is a univariate response variable,
- $x \in \mathbb{R}^p$, covariates, such as $\mathbb{E}(x) = \mu$ and $\mathbb{V}(x) = \Sigma$,
- ϵ is independent of x,
- f the link function and $\beta \in \mathbb{R}^p$ the euclidean parameter are unknown.

f being unknown, β is not fully identifiable.
However, it is possible to estimate the space generated by β, called EDR (Effective Dimension Reduction) space.
Note : The model (1) can be generalized to a non-additive and heteroscedastic noise.

Estimation of the EDR space and f

The estimation of the SIR model involves 2 steps:

Estimation of the EDR space

$$\Gamma = \mathbb{V}\left[\mathbb{E}\{x|T(y)\}\right] = \sum_{h=1}^{H} p_h(m_h - \mu)(m_h - \mu)'$$

- T a slicing function which cuts the Y support into H slices $\{s_1, \ldots, s_H\}$
- $p_h = P(Y \in s_h)$ and $m_h = \mathbb{E}[X \mid Y \in s_h]$,
- The principal eigenvector of $\Sigma^{-1}\Gamma$, denoted $b \in \mathbb{R}^p$, is an EDR direction.

$\Longrightarrow$ The principal eigenvector $\hat{b}_{SIR}$ of $\hat{\Sigma}^{-1}\hat{\Gamma}$ is an estimated EDR direction. This estimation, suffers from the curse of dimensionality.

Estimation of f

Use of a non-parametric kernel estimator on $(y, \hat{b}'_{SIR}x)$.

Soft thresholding

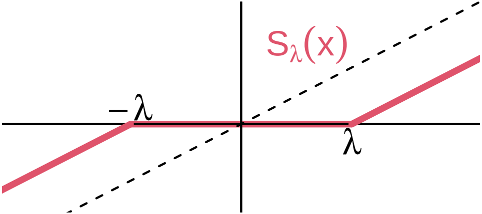


Figure: Soft thresholding

$$S_\lambda(x) = sign(x) \times \begin{cases} |x| - \lambda & \text{if } |x| - \lambda > 0, \\ 0 & \text{else.} \end{cases} \quad (2)$$

$\Longrightarrow$ Soft thresholding: continuity, but bias for high values.

Hard thresholding

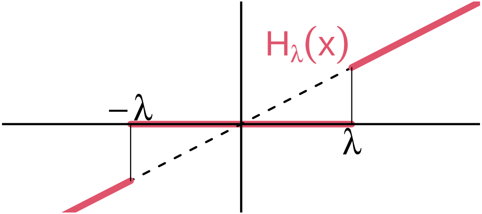


Figure: Hard thresholding

$$H_\lambda(x) = \begin{cases} x & \text{if } |x| - \lambda > 0, \\ 0 & \text{else.} \end{cases} \tag{3}$$

$\Longrightarrow$ Hard thresh.: no bias for high values, but discontinuity.

ST-SIR and HT-SIR estimators

- $\hat{b}_{ST-SIR}(\lambda)$: principal eigenvector of $S_\lambda(\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n)$.
- $\hat{b}_{HT-SIR}(\lambda)$: principal eigenvector of $H_\lambda(\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n)$.

The choice of the thresholding hyper-parameter λ must provide a balance between

- correct variable selection,
- low distortion of the estimated direction $\hat{b}_{SIR}$ too much.

$\hookrightarrow \hat{\lambda}_{\text{opt}} \implies$ selection of $\hat{p}^\star$ selected variables.

Before variable selection...

- $\hat{b}_{SIR}$: SIR estimator based on the p variables.
- $\hat{b}_{HT-SIR} := \hat{b}_{HT-SIR}(\hat{\lambda}_{opt-HT})$.
- $\hat{b}_{ST-SIR} := \hat{b}_{ST-SIR}(\hat{\lambda}_{opt-ST})$.

... after variable selection

1. Consider the $\hat{p}^{\star}$ selected variables (based on $\hat{\lambda}_{opt-ST}$)).
2. $\hat{b}^{\star}_{SIR}$: estimated EDR direction using the“reduced” SIR model based on the selected $\hat{p}^{\star}$ variables.

Example: the simulated regression model

$$y = (x'\beta)^3 + \epsilon,$$

- $\beta = (1, \dots, 1, 0, \dots, 0)' \in \mathbb{R}^p$, here $p = 20$ and $p^\star = 10$
- $x \sim \mathcal{N}(0, \mathbb{I}_p)$
- $\epsilon \sim \mathcal{N}(0, 10)$ and $\epsilon \perp\!\!\!\perp x$.

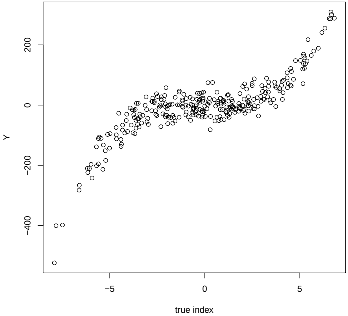
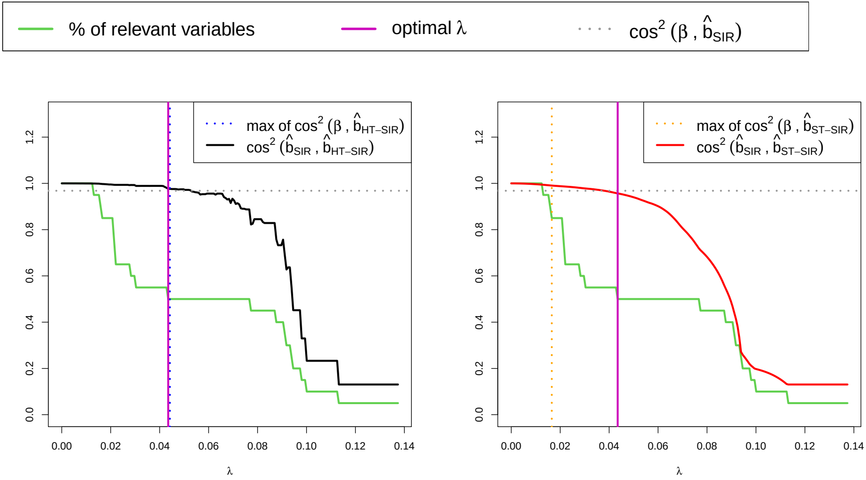


Figure: Sample size n=300, Noise to signal ratio = 0.1

Simple case: comparison between HT-SIR and ST-SIR



(a) HT-SIR (b) ST-SIR

Overall results for that case

HT-SIR and **ST-SIR**, similar results in **selection**:

- $\hat{p}^* = 10$ variables selected over the $p = 20$ variables.
- List of the $\hat{p}^* = 10$ selected variables :
  $X1, X2, X3, X4, X5, X6, X7, X8, X9, X10$

Very good **estimation** of the EDR direction:

- $cos^2(\beta, \hat{b}_{HT-SIR}) = 0.98$
- $cos^2(\beta^\star, \hat{b^\star}_{SIR}) = 0.99$

Simulation plan

Same regression model: $y = (x'\beta)^3 + \epsilon$

- $\beta = (1, \dots, 1, 0, \dots, 0)' \in \mathbb{R}^p$,
- $x \sim athcalN(0, \mathbb{I}_p)$
- $\epsilon \sim \mathcal{N}(0, 10)$ and $\epsilon \perp\!\!\!\perp x$.

Simulations with various values of (n, p, p^*):

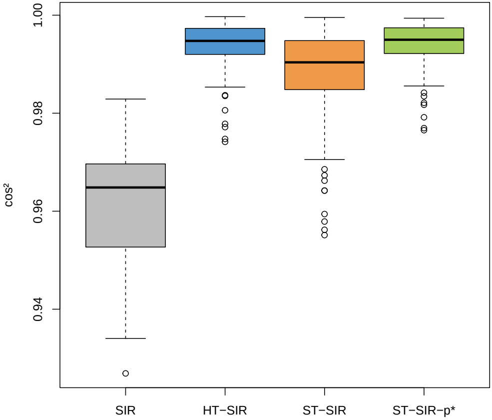
- $n \in \{200, 300, 500\}$
- p and $p^\star$ so that $\dfrac{p^\star}{p} = \dfrac{1}{5}$

  $\hookrightarrow \quad (p, p^\star) \in \{(25, 5), (50, 10), (100, 20)\}$
- Noise to Signal ratio: $\mathbb{V}(\epsilon)/\mathbb{V}(y) \in \{0.1, 0.01\}$

$N = 100$ replications considered for each case.

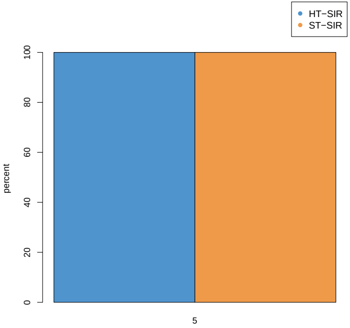
Simulations with $n = 500$, $p = 25$, $p^\star = 5$, NTS ratio$= 0.1$

Comparison of cos^2

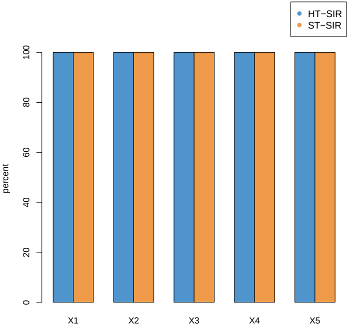


Simulations with $n = 500$, $p = 25$, $p^\star = 5$, NTSratio$= 0.1$

Selection performances



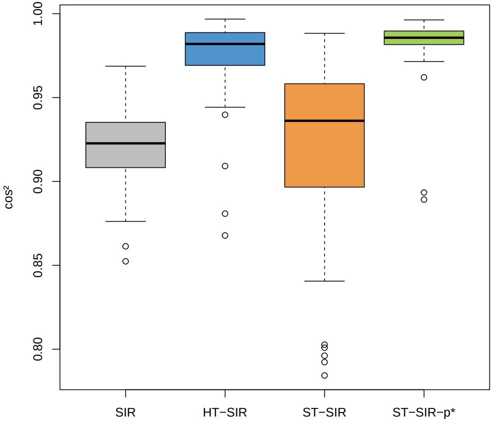
(a) Size of the reduced model



(b) Variables selected in the reduced model

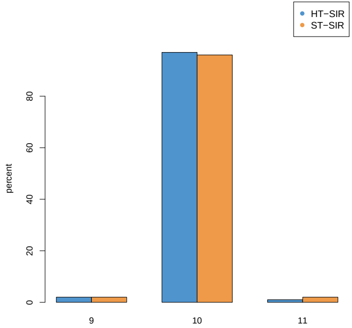
Increase p from 25 to 50 and $p^\star$ from 5 to 10

Comparison of cos^2

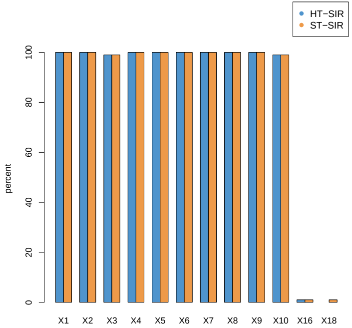


Increase p from 25 to 50 and $p^\star$ from 5 to 10

Selection performances



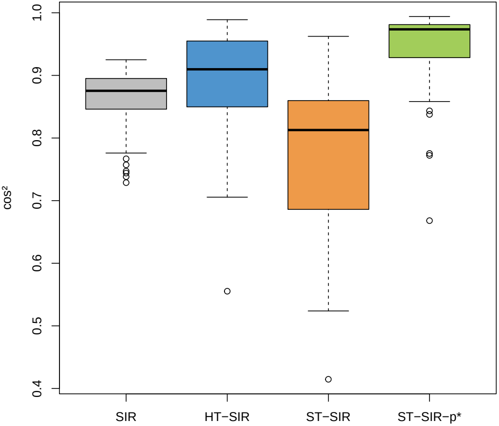
(a) Size of the reduced model



(b) Variables selected in the reduced model

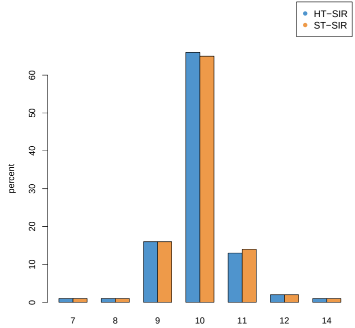
Decrease n from 500 to 300

Comparison of cos^2

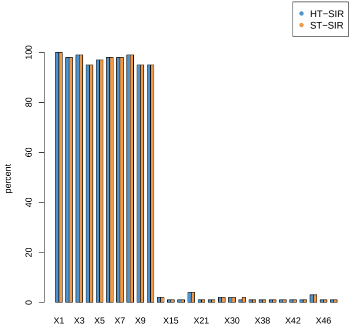


Decrease n from 500 to 300

Selection performances



(a) Size of the reduced model



(b) Variables selected in the reduced model

Concluding remarks

- No significant difference in variable selection between HT-SIR and ST-SIR.
- Efficient for $p < n$ for the two approaches.
- Bootstrap could stabilize the results and make them more robust (under investigation).
- Other thresholding methods (such as SCAD) could also offer interesting results (under investigation)
- An R package is under development!

References I

Duan, N. and K.-C. Li (1991). “Slicing regression: a link-free regression method”. In: *The Annals of Statistics* 19, pp. 505–530.

Appendix

Choose the optimal lambda - Exemple 1

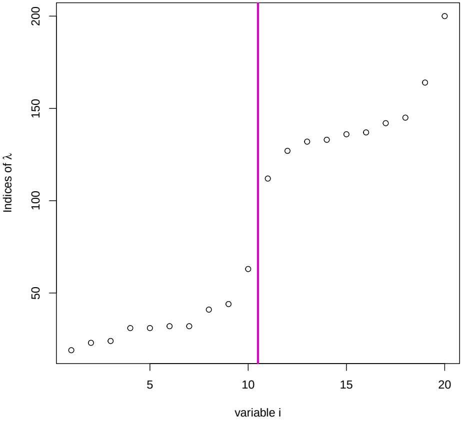


Figure: Index of the lambda from which the variable i is useless