Criteria for Variable Selection with Dependence

Aurélie Boisbunon, Stéphane Canu, Dominique Fourdrinier
Université de Rouen and INSA de Rouen

Abstract

- Derivation of a new criterion through loss estimation
- Valid under the spherical assumption allowing for dependence between observations
- Integration in a whole procedure from model exploration to model evaluation

Context

Linear regression model

$$Y = X\beta + \varepsilon \qquad \begin{cases} Y \in \mathbb{R}^n \\ X \in \mathbb{R}^n \times \mathbb{R}^p \text{ (fixed)} \\ \beta \in \mathbb{R}^p \\ \varepsilon \in \mathbb{R}^n \sim \mathcal{S}_n(0) \end{cases}$$

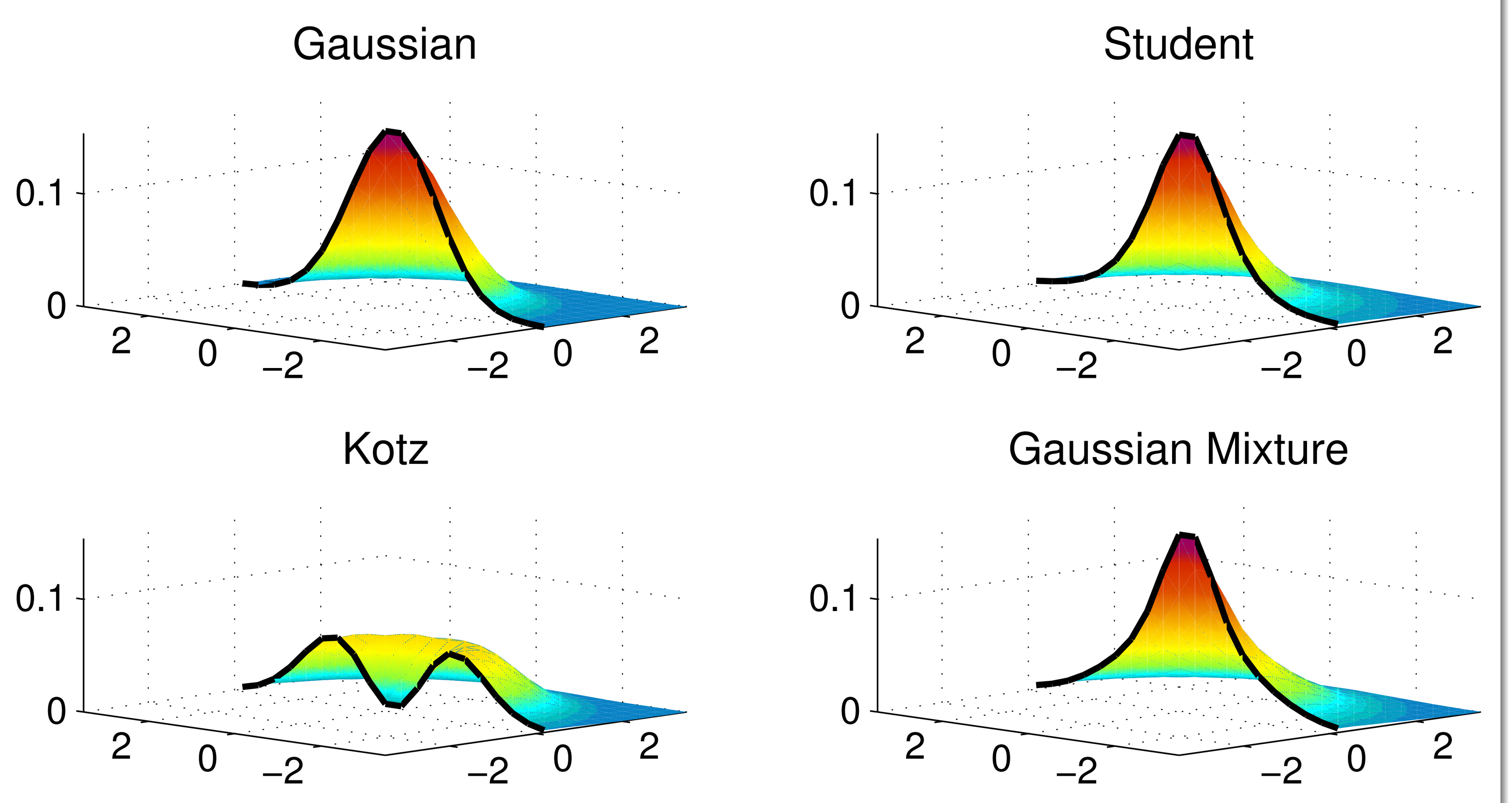
- Aim: sparse estimation of β

- Literature based exclusively on either
  - estimation of a sparse β
  - evaluation of several models
- Mostly rely on independence
  - generally not true in real examples

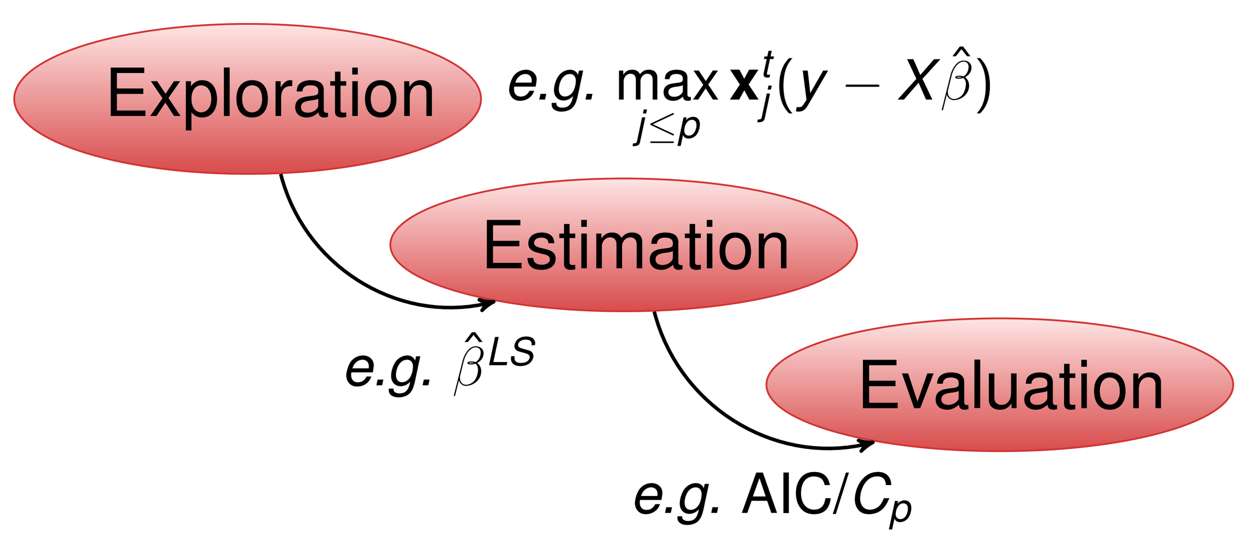
Framework

Spherically symmetric distributions $\mathcal{S}_n$

- No need to specify the form of the distribution
- Dependence between the components of Y
- Distributional robustness



Model Selection steps



Procedure

Firm Shrinkage / MC+

- Exploration: regularization path
- Estimation: nearly unbiased estimator

$$\hat\beta_j^{FS}(\lambda) = \begin{cases} 0 & |\hat\beta_j^{LS}| \leq \lambda \\ \alpha(\hat\beta_j^{LS} - \lambda \, sign(\hat\beta_j^{LS}))/(\alpha - 1) & \lambda < |\hat\beta_j^{LS}| \leq \alpha\lambda \\ \hat\beta_j^{LS} & |\hat\beta_j^{LS}| > \alpha\lambda \end{cases}$$

- $\lambda > 0 \ \rightarrow$ tunes sparsity, $\alpha > 1 \ \rightarrow$ tunes bias
- $\hat\beta^{LS} = (X^tX)^{-1}X^ty$ (least-squares estimator)

Evaluation: Loss estimation

- Loss function $L(\hat\beta, \beta) = \| \underbrace{X\hat\beta}_{\text{estimate}} - \underbrace{X\beta}_{\text{true}} \|^2$
- Estimation $\widehat{L}$ of L
  - Step 1: unbiased estimator $\widehat{L}_0 \setminus \forall\beta \ \mathbb{E}_Y[\widehat{L}_0] = \mathbb{E}_Y[L(\hat\beta, \beta)]$
  - Step 2: improvement $\widehat{L}_\rho \setminus \mathbb{E}_Y(\widehat{L}_\rho - L)^2 \leq \mathbb{E}_Y(\widehat{L}_0 - L)^2$

$$\widehat{L}_\rho = \|y - X\hat\beta^{FS}(\lambda)\|^2 + (2df - n)\frac{\|y - X\hat\beta^{LS}\|^2}{n-p} - \boldsymbol{\rho(y)}$$

$\hookrightarrow$ Correction function: $\rho(y) = C\gamma^{-1}(y)\|y - X\hat\beta^{LS}\|^4$

$\hookrightarrow$ $\gamma(y) = k \max_{i\leq p}\{(q_i^ty)^2 \setminus |q_i^ty| \leq \lambda\} + \sum_{j\leq p}(q_j^ty)^2 \mathbb{1}_{\{|q_j^ty|\leq\lambda\}}$

- Selection: $\hat\lambda = \arg\min_{\lambda\in\mathbb{R}_+} \widehat{L}_\rho(y, \lambda)$

Results

Example: $n = 40$ observations, $p = 5$ variables, $\beta = (2, 0, 0, 4, 0)^t$, $r = 5000$ replicates

$\varepsilon \sim \mathcal{N}_n(0, I_n)$

Subset	$\widehat{L}_\rho$	AIC	BIC	LOOCV	Real loss
{4}	26.12 (0.56)	20.18 (0.59)	40.05 (0.83)	14.42(16.18)	14.17 (0.43)
{1,4}	**44.41 (0.60)**	39.02 (0.74)	39.37 (0.49)	32.71(12.27)	54.29 (0.56)
{1,2,4}	1.33 (0.15)	7.57 (0.34)	3.66 (0.26)	5.68 (3.06)	7.46 (0.33)
{1,3,4}	1.30 (0.13)	7.83 (0.40)	3.73 (0.19)	6.93 (2.99)	7.63 (0.32)
{1,4,5}	2.13 (0.20)	7.73 (0.40)	3.73 (0.36)	6.49 (3.73)	7.87 (0.27)

Table: Percentage of selection with Firm Shrinkage

$\varepsilon \sim \mathcal{T}_n(\nu = 4)$

Subset	$\widehat{L}_\rho$	AIC	BIC	LOOCV	Real loss
∅	8.87 (0.43)	9.94 (0.65)	20.90 (0.74)	7.21 (3.12)	14.62 (0.45)
{4}	19.11 (0.29)	15.77 (0.37)	24.33 (0.45)	12.63 (8.99)	14.88 (0.50)
{1,4}	**38.01 (0.62)**	32.08 (0.74)	35.15 (0.82)	26.35(11.77)	46.08 (0.78)
{1,2,4}	0.00 (0.14)	6.08 (0.21)	2.74 (0.16)	5.82 (2.93)	4.65 (0.21)
{1,4,5}	1.63 (0.22)	6.21 (0.36)	2.83 (0.20)	6.58 (3.39)	4.50 (0.16)

Table: Percentage of selection with Firm Shrinkage

Conclusion

- STOP using AIC, BIC, and LOOCV
- USE $\widehat{L}_\rho$ instead
- Possible application to classification, clustering, etc.

References