Adaptive ridge regression for variable selection

Jean-Michel Becu[1], Christophe Ambroise[2], Cyril Dalmasso[2], Yves Grandvalet[1]

1 - Heudiasyc (UTC), 2 - LaMME (UEVE)

jean-michel.becu@hds.utc.fr

A two step approach

Goal

Retrieve variables that explain output by **significance testing** with **false discovery rate** control.

Proposition

Transform an high dimensional problem to an easier problem by a **two step** approach, on two **independant** subsamples :

- **Screening :** Select variables with a sparse approach (Lasso)
- **Cleaning :** Significance test for selected variables during the screening step

Schematic protocol

$$\overbrace{\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \xrightarrow[\text{fit model}]{\text{step I } (\mathcal{D}_1)} \{S_\lambda\}_{\lambda\in\Lambda} \xrightarrow[\text{select model}]{\text{step II } (\mathcal{D}_1)} S_{\hat\lambda}}^{\text{screening stage}} \overbrace{\xrightarrow[\text{test support}]{\text{step III } (\mathcal{D}_2)} \hat S}^{\text{cleaning stage}}$$

Notations

$\mathcal{D} = \{X, \mathbf{y}\}$

$S_\lambda = \{j \in \{1, ..., p\} | \hat\beta_j(\lambda) \neq 0\}$

$\{S_\lambda\}_{\lambda\in\Lambda}$: All subsets estimated by Lasso, for a Λ grid on $\mathcal{D}_1$

$S_{\hat\lambda}$: Best subset chosen by k-folds cross-validation on $\mathcal{D}_1$

$\hat S$: Final subset of selected variables: $S_{\hat\lambda}$ is cleaned on $\mathcal{D}_2$ to control FDR.

The adaptive ridge solution

Equivalence with Lasso

Lasso : $\hat{\boldsymbol\beta}(\lambda) = \underset{\boldsymbol\beta\in\mathbb{R}^p}{\arg\min} \|X\boldsymbol\beta - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol\beta\|_1$

Specific Penalty : $\omega_j = \dfrac{\lambda}{|\hat\beta_j(\lambda)|}$

Ridge : $\hat{\boldsymbol\beta}(\omega) = \underset{\boldsymbol\beta\in\mathbb{R}^p}{\arg\min} \|X\boldsymbol\beta - \mathbf{y}\|_2^2 + \sum_{j=1}^{p}\omega_j\beta_j^2 = \hat{\boldsymbol\beta}(\lambda)$

ℓ_1-variant regression	Adaptive ridge penalty
Lasso	$\frac{\lambda}{\lvert\beta_j\rvert}$
Elastic-net	$\lambda\left(\frac{\alpha}{\lvert\beta_j\rvert} + 1 - \alpha\right)$
Group Lasso	$\frac{\lambda}{\sqrt{\lvert G(j)\rvert \sum_{m\in G(j)} \beta_m^2}}$
Sparse Group Lasso	$\lambda\left(\frac{\alpha}{\lvert\beta_j\rvert} + \frac{1-\alpha}{\sqrt{\lvert G(j)\rvert \sum_{m\in G(j)} \beta_m^2}}\right)$

$G(j)$ correspond to the group of the j^{th} variable.

Fisher test for variables in $S_{\hat\lambda}$

Estimated statistic

$$F_j = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}_1\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}_1\|^2}$$

Simulated statistic under H0

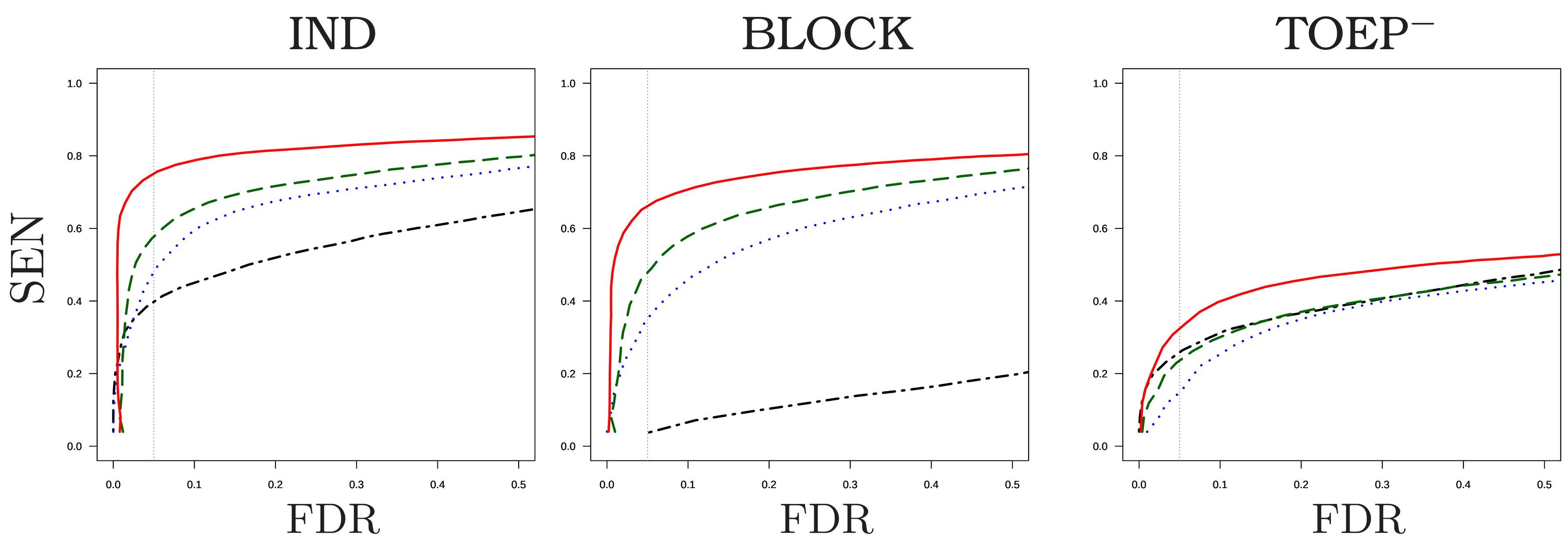
$$F_j^* = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}_1^*\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}_1^*\|^2}$$

Let $\tilde X = \{\mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_j^*, \mathbf{x}_{j+1}, ..., \mathbf{x}_p\}$, $\mathbf{x}_j^*$ is the permuted vector $\mathbf{x}_j$ and $\hat{\mathbf{y}}_1^* = \tilde X\left(\tilde X^\top \tilde X + \Omega\right)^{-1}\tilde X^\top \mathbf{y}$

Simulation design	IND		BLOCK		TOEP⁻	
	FPR	SEN	FPR	SEN	FPR	SEN
permutation F-test	5.1	92.4	3.9	86.7	4.7	81.9
standard F-test	9.9	93.1	11.8	89.6	15.4	87.1
standard t-test	8.0	94.0	12.4	93.1	7.9	85.1

IND Independant variables, **BLOCK** Block correlation structure, **TOEP-** Toeplitz correlation structure with 50% of negative correlation. For all $max(\rho_{j\neq k}) = 0.5$.

Results in a high-dimensional setting



$$\text{SEN} = \mathbb{E}\left[\frac{TP}{TP+FN}\mathbb{I}_{\{(TP+FN)>0\}}\right], \text{FDR} = \mathbb{E}\left[\frac{FP}{TP+FP}\mathbb{I}_{\{(TP+FP)>0\}}\right]$$

Simulation design	IND		BLOCK		TOEP⁻	
	FDR	SEN	FDR	SEN	FDR	SEN
Screening (Lasso)	76.7	87.5	76.0	83.9	79.9	56.5
AR* cleaning (—)	4.2	**76.1**	2.8	64.8	4.3	**39.6**
Ridge cleaning (- -)	4.6	57.9	3.6	49.8	4.7	27.2
OLS cleaning (···)	3.9	48.3	3.1	37.1	3.7	25.3
Univar (-·-)	4.4	40.4	86.4	**71.0**	4.2	28.4

AR* : Adaptive Ridge

Conclusion

- An efficient way to select variables in high dimensional setting
- An adaptive approach currently applied to Lasso regression, but it's applicable to several variants of the lasso.
- Could be used for graphical inference or more simply to prediction

References

[BA] Becu, Grandvalet, Ambroise and Dalmasso , 2015. Beyond Support in Two-Stage Variable Selection. *Statistics and Computing.*

[WR] Wasserman and Roeder, 2009. High-dimensional variable selection. *Annals of Statistics.*