Empirical properties of the Lasso by simulations
Part 2

Franck Picard[⋆]

[⋆] *Laboratoire de Biométrie et Biologie Évolutive*, Univ. Lyon 1

ECAS, High Dimensional Statistics Course, October 2017

Practice, Part 2

1. Compete the Lasso with other selection methods
2. Modify the Lasso: adaptive lasso and group lasso
3. Compare performance in low/high dimension

First competitors ‣ Competitors

- **OLS** (negative control): no selection. Use slight ridge regularization for high dimensional cases

  $\rightarrow$ *How does the Lasso compete with the "worst" method ?*

- **Oracle** (positive control): knows the true null and non-null positions. Perform OLS on J_0.

  $\rightarrow$ *How does the Lasso compete with the "best" method ?*

- **Stepwise**: variable selection based on an iterative algorithm (ℓ_0)

  $\rightarrow$ What is the gain of using the Lasso (ℓ_1) compared with ℓ_0 selection ?

Modifications of the Lasso: the adaptive Lasso

Competitors

- Two-step procedure to account for bias and to perform a component-wise selection.
- First estimate $\widehat{\beta}_{\text{init}}$, using the Lasso or Ridge regression
- Construct weights such that $w_j = \widehat{\beta}_{j,\text{init}}$, and solve

$$\widehat{\beta}_\lambda = \text{Argmin}_\beta \left(\frac{1}{n} \|Y - X\beta\|_2^2 + \sum_{j=1}^{p} \frac{\lambda}{|w_j|} |\beta_j| \right).$$

- If $\widehat{\beta}_{j,\text{init}}$ is big, the penalty $\lambda/|w_j|$ will be small, hence a reduced shrinkage for β_j

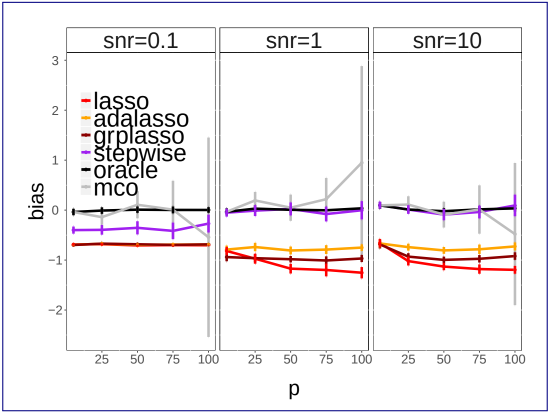
The Group Lasso ‣ Competitors

- When covariates can be grouped to be shrunk/selected
- Prior information that can be put in a structured penalty

$$\widehat{\beta}_\lambda = \mathrm{Argmin}_\beta \left(\frac{1}{n} \| Y - X\beta \|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{n_k \sum_{j \in G_k} \beta_j^2} \right).$$

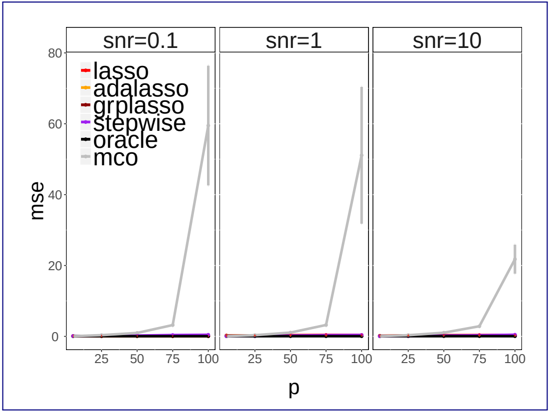
- One could consider a group-wise calibration of λ
- Requires some prior knowledge

Estimation quality: Bias



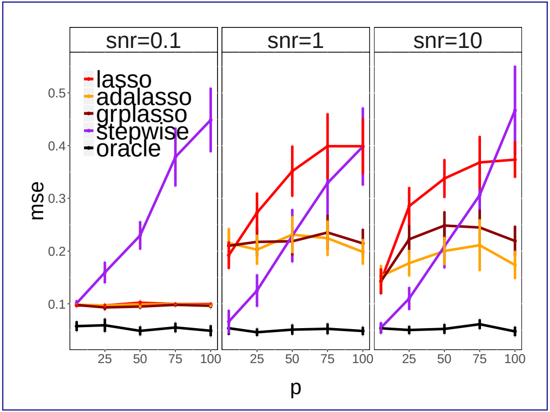
The Lasso is more biased than other methods but the the adaptive version corrects the bias.

Estimation quality: MSE



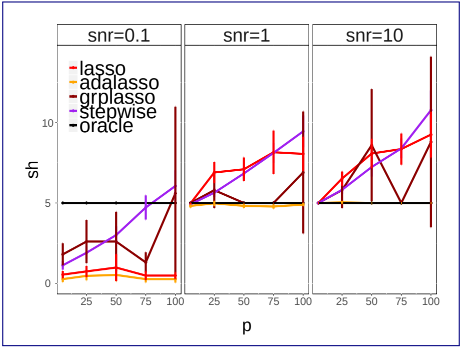
Performing no selection explodes the MSE !

Estimation quality: MSE-bis



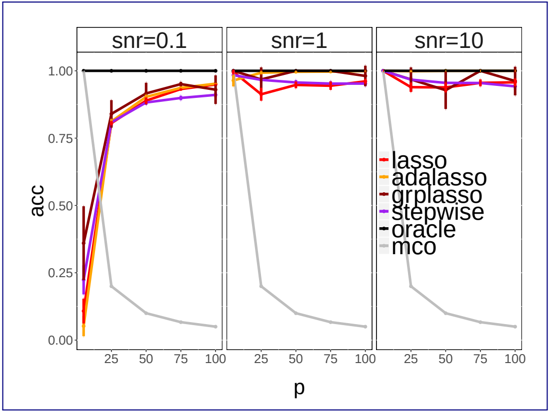
Stepwise precision not robust to increase in p. Adaptivity increases precision.

Estimation quality: Model Selection



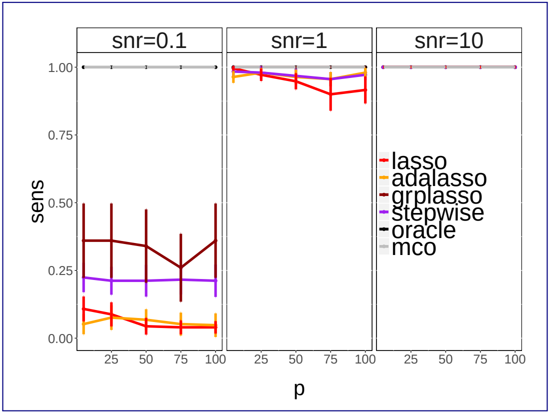
The Lasso is more conservative when signal is low but overestimates the dimension when there is signal. The adaptive lasso is accurate for model selection

Estimation quality: Accuracy



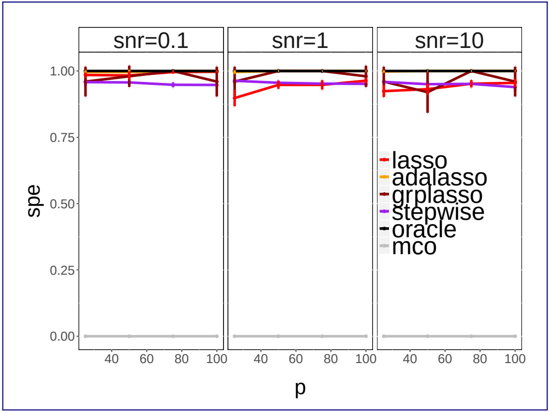
Accuracies are comparable between variable selection methods

Estimation quality: Sensitivity



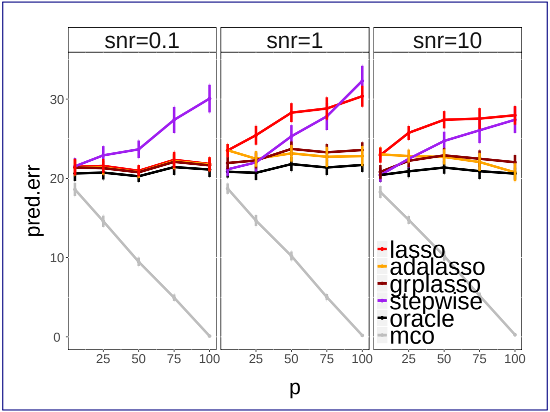
The Lasso is more conservative (less sensitive)...

Estimation quality: Specificity



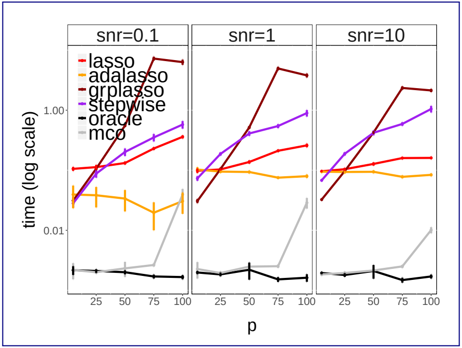
...with comparable specificities

Estimation quality: Prediction Error



Prediction errors are comparable (lasso/stepwise) when there is some signal. The adaptive lasso is more accurate in prediction

Estimation quality: Time of execution



The complexity of the stepwise method is prohibited for large datasets !
Be cautious when comparing execution time (depends on implementation)

What about high dimensional models ?

- We explored only situations when $n \leq p$, what about $n > p$?
- The situation becomes complex because the information is no longer contained in the SNR, but also in a mix between n, p and p_0
- In the context of linear regression, M. Wainwright introduced the notion of rescaled sample size $\frac{n}{p_0 \log(p-p_0)}$
- Question : what would it take to recover the support of β in terms of rescaled sample size ?
- $\mathbb{S}_{\pm}(\beta)$ is the vector of signs of β such that:

$$\mathbb{S}_{\pm}(\beta_i) = \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{if } \beta_i = 0 \end{cases}$$

Results on signed support

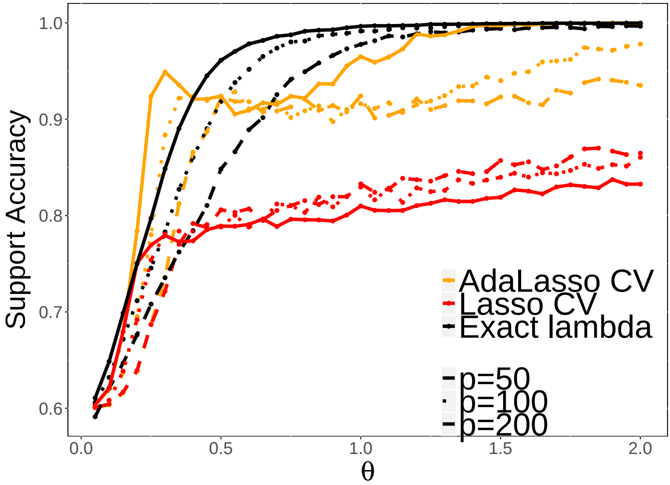
- M. Wainwright shows the existence of two constants that depend on $\Sigma = \mathbb{V}(X)$, $0 < \theta_\ell(\Sigma) \leq \theta_u(\Sigma) < \infty$ such that for a given value of the lasso regularization hyperparameter

$$\lambda_n = \sqrt{\frac{2\sigma^2 \log(p_0) \log(p - p_0)}{n}}$$

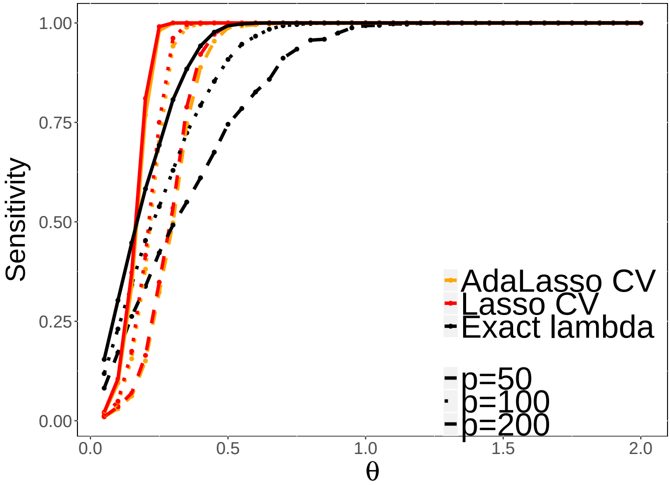
- if $n/(2p_0(\log(p - p_0))) > \theta_u(\Sigma)$ then it is always possible to find a value of λ such that the lasso has a unique solution $\widehat{\beta}$ with $\mathbb{P}\{\mathbb{S}_\pm(\beta^*) = \mathbb{S}_\pm(\widehat{\beta})\}$ tending to 1.
- if $n/(2p_0(\log(p - p_0))) < \theta_\ell(\Sigma)$, then whatever the value of $\lambda > 0$, no solution of the lasso will recover the signed support of β^*, $\mathbb{P}\{\mathbb{S}_\pm(\beta^*) = \mathbb{S}_\pm(\widehat{\beta})\}$ tends to 0.
- if $\Sigma = I$, then $\theta_\ell(I) = \theta_u(I) = 1$.

Rescaled Sample size and Accuracy

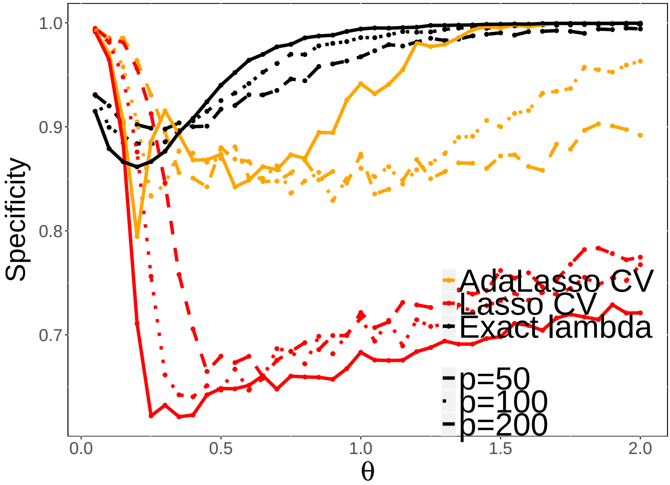
Results on Support Accuracy



Rescaled Sample size and Sensitivity



Rescaled Sample size and Specificity



Conclusions

- Quite complex to compare methods based on multiple criteria : model selection, selection accuracy, prediction, time of execution
- Overall, the adaptive Lasso seems to perform well on all criteria. Simple to implement
- All results highlight the importance of **calibration** in practice
- Performance in high dimension depend on a mix between p, p_0, n and SNR