Computational Statistics
Model selection in sparsity

georges.tod@outlook.com

March 2020

1 Introduction

Let us define a linear model,

$$Y = X\beta + \epsilon$$

where the observed variable $Y \in \mathbb{R}^n$, the explanatory variable $X \in \mathbb{R}^{n\times m}$, the coefficients vector $\beta \in \mathbb{R}^m$ and $\epsilon \sim N_n(0, \sigma^2 I_n)$ represents some noise.
In high dimensions, i.e. when $m > n$, the solution from classical least-squares fails because $X'X$ cannot be inverted anymore. Penalized versions of least-squares such as ridge regression or least absolute shrinkage and selection operator (LASSO) can be then applied. In this work, it is assumed that β is sparse therefore motivating the application of LASSO, for which, β is estimated by,

$$\hat{\beta} = \text{argmin}_\beta\Big(\frac{1}{2}||Y - X\beta||_2^2 + \lambda||\beta||_1\Big) \tag{1}$$

Ideally, we would like $\hat{\beta} = \beta$, such that the prediction error is just the noise but this cannot be guaranteed. However, a number of parameters that minimizes an estimation of the prediction error of the model can be selected[1] as a trade-off between model *closeness* and *complexity*. Our numerical experiments illustrate in particular the impact of knowing (or not) a priori the standard deviation of the noise term.
In the following sections, the use of two solvers implemented in ThreshLab is illustrated, namely iterative soft threshloding (ITST) and least angle regression (LAR). Synthetic data is used and has been generated using ThreshLab's function 'setupsimulationYisXbetaplussigmaZ' with a degree of sparsity of 0.05 and a signal to noise ratio of 10. For the sake of notation simplicty, when not specified, the used norm is L_2.

2 Iterative Soft Thresholding (ITST)

An approximate solution to equation 1 can be found in this case by solving the problem in two steps. First, the residual error is minimized with a fixed

[1] In theory, the true number of parameters can be found when $n \to +\infty$.

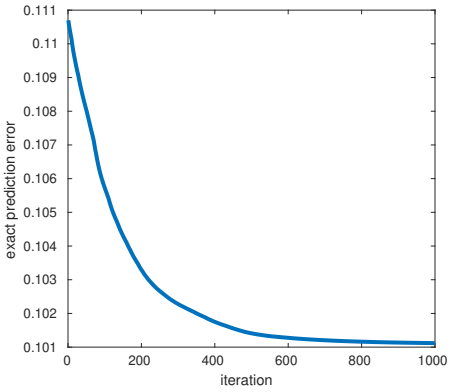

Figure 1: Exact model prediction error (PE) at each iteration for a fixed value of the threshold $\lambda = 0.12$. Here at $m = 1000, n = 200$, some improvement has been achieved after 10^3 iterations.

threshold λ. Secondly, the value of the threshold that minimizes an estimation of the prediction error, namely Mallow's C_p, is found.

2.1 $\hat{\beta}$ at fixed λ, minimizing the residual error

Iteratively, the residual error $||Y - X\hat{\beta}||^2$ is computed until it cannot be reduced anymore for a fixed value of λ. By introducing the iteration number $r \geq 1$ and an initial guess, $\hat{\beta}_0 = (0, \ldots, 0)' \in \mathbb{R}^m$, each iteration updates along the highest descent of the residual error,

$$\begin{aligned} \hat{\beta}_{r-1/2} &= \hat{\beta}_{r-1} - \frac{\partial(||Y - X\hat{\beta}_{r-1}||^2)}{\partial\hat{\beta}_{r-1}} \\ &= \hat{\beta}_{r-1} + 2X'(Y - X\hat{\beta}_{r-1}) \end{aligned} \tag{2}$$

Finally, the norm L_1 of $\hat{\beta}$ is restricted by thresholding or shrinking it component wise,

$$\hat{\beta}_{r,i} = \begin{cases} 0 & \text{if } -\lambda \leq \hat{\beta}_{r-1/2,i} \leq \lambda \\ \hat{\beta}_{r-1/2,i} + \lambda & \text{if } \hat{\beta}_{r-1/2,i} < -\lambda \\ \hat{\beta}_{r-1/2,i} - \lambda & \text{if } \hat{\beta}_{r-1/2,i} > \lambda \end{cases}$$

This procedure is called 'soft' thresholding as no discontinuity is introduced when $|\hat{\beta}_{r-1/2,i}| = \lambda$. To assess the performance of the model using the estimated parameters, the prediction error is introduced, see Figure 1 for illustration,

$$PE(\hat{\beta}_r) = \frac{1}{n}||X\beta - X\hat{\beta}_r||^2 \tag{3}$$

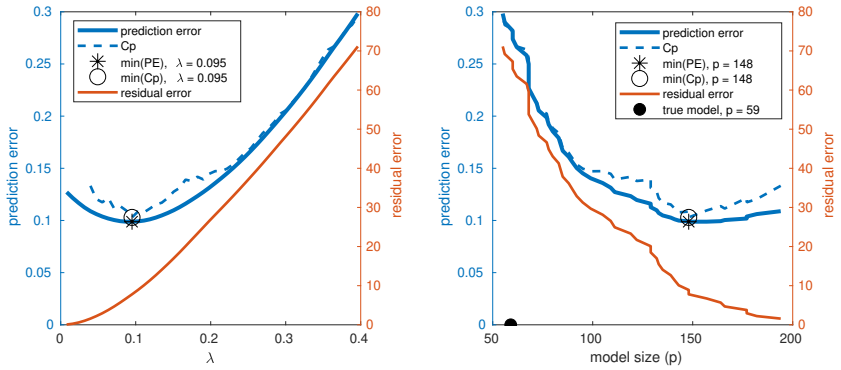

Figure 2: Choosing the threshold λ that minimizes the prediction error or its estimation (Mallow's C_p) allows to select the model size (i.e. non zeros components). Here at $m = 1000, n = 200$.

2.2 $\hat{\beta}$ at varying λ, minimizing the prediction error

In the previous section, it was assumed the threshold λ as known and fixed. However, by chosing a better threshold, a better estimation of $\hat{\beta}$ can be found. The threshold is chosen such that the prediction error is minimized, leading to a balance between model *closeness* and *complexity*. When working with real data, i.e. when the data generating process (DGP) is unknown, it will not be possible to compute the exact prediction error, however Mallow's C_p can provide us an estimation of it. If we know the variance of the noise term, then the non-studentized version of Mallow's C_p can be used; in most practical applications this is not the case and the studentized version of it will need to be used. This section shows some of the implications of using either by numerical experiments.

Below we introduce the *non studentized* Mallow's C_p,

$$\Delta(\hat{\beta}_p) = \underbrace{\frac{1}{n}||Y - X\hat{\beta}_p||^2}_{\text{closeness}} + \sigma^2 \underbrace{\left(\frac{2p}{n} - 1\right)}_{\text{complexity}} \tag{4}$$

where p stands for the model size. On Figure 2, it can be seen that Mallow's C_p minima captures quite well the minima of the prediction error, even for $n = 200$. Therefore, for that given threshold, both PE and C_p give the same model size ($p = 148$). However, it does not match the data generating process model for which $p = 59$. As the sample size increases, the variance of C_p and the PE decrease. Even though, the chosen model using C_p will still not be the one of the DGP, the prediction error does seem to tend to 0. Simulations for $n > m$ are not carried out to stay in the high dimensional case.

Now when the noise term is unknown, it needs to be estimated. We can then

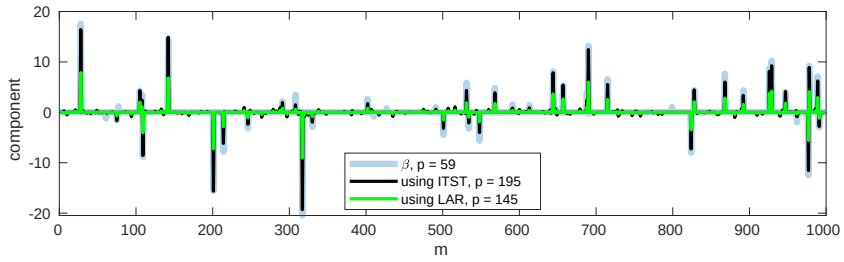

Figure 3: A couple of solutions for $n = 800$ at $\min \Delta(\hat{\beta}_p)$ using both solvers. Both estimations capture main β components with some shrinkage. LAR captures a higher number of zeros in this case.

apply the *studentized* Mallow's C_p,

$$\Delta_s(\hat{\beta}_p) = \frac{||Y - X\hat{\beta}_p||^2}{\hat{\sigma}^2} + 2p - n \tag{5}$$

Where $\hat{\sigma}^2$ is model indepedent. In order to estimate it, the noise needs to be isolated. A rough approximation can be obtained by using the model that gives the smallest residual error (see Figure 2) which is the 'fullest' model at hand,

$$\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}_{\nu_{full}}||^2}{n - \nu_{full}} \tag{6}$$

The point used for the estimation at each sample size can be visualized on the right hand side of each subfigure of Figure 5. Once, the studentized Mallow's C_p is computed it can be compared to a scaled version of the prediction error (see Figure 5),

$$PE_{scaled}(\hat{\beta}_p) = \frac{n}{\sigma^2} PE(\hat{\beta}_p)$$

For small sample size values, for example at $n = 200$ $(m/n = 5)$ the estimation of the prediction can be very bad. For $n = 400$ $(m/n = 2.5)$ a single simulation out of 20 gives a terrible result. At higher sample sizes, the C_p starts behaving in a more repeatable way. From $n = 800$ $(m/n = 1.25)$ and onwards, Figure 5 gives the intuitive feeling the bias w.r.t. exact prediction error will not converge to 0. However, the variance of $\min \Delta(\hat{\beta}_p)$ does seem to converge to 0. In conclusion, a good estimation of the noise or a large sample size is important to be able to use the non studentized version of Mallow's C_p.

3 Least Angle Regression (LAR)

An intuitive way of describing the algorithm is descibed by Hastie, Tibshirani and Friedman[2]. As for the ITST, the initial solution is a vector full of zeros

[2]Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics

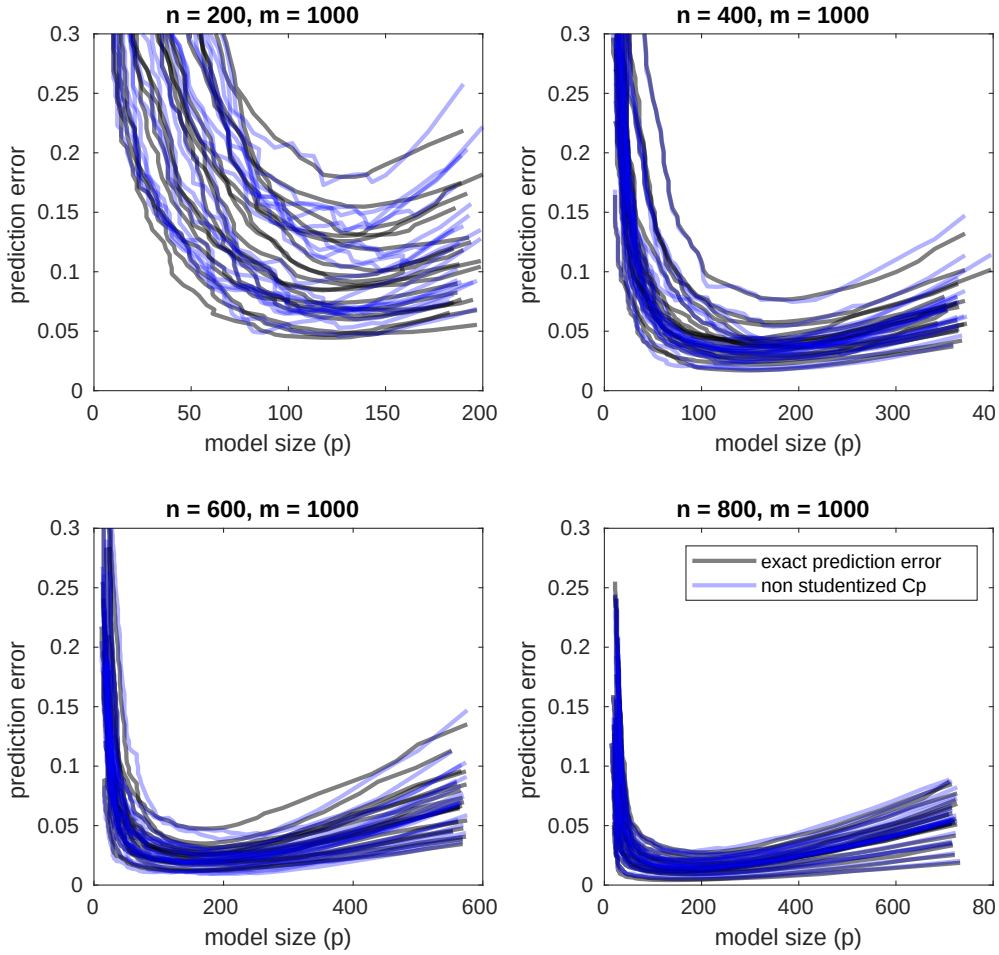

Figure 4: Impact of varying sample size on non studentized Mallow's C_p. Each simulation is ran 20 times. As the sample size increases both C_p and the prediction error tend to 0.

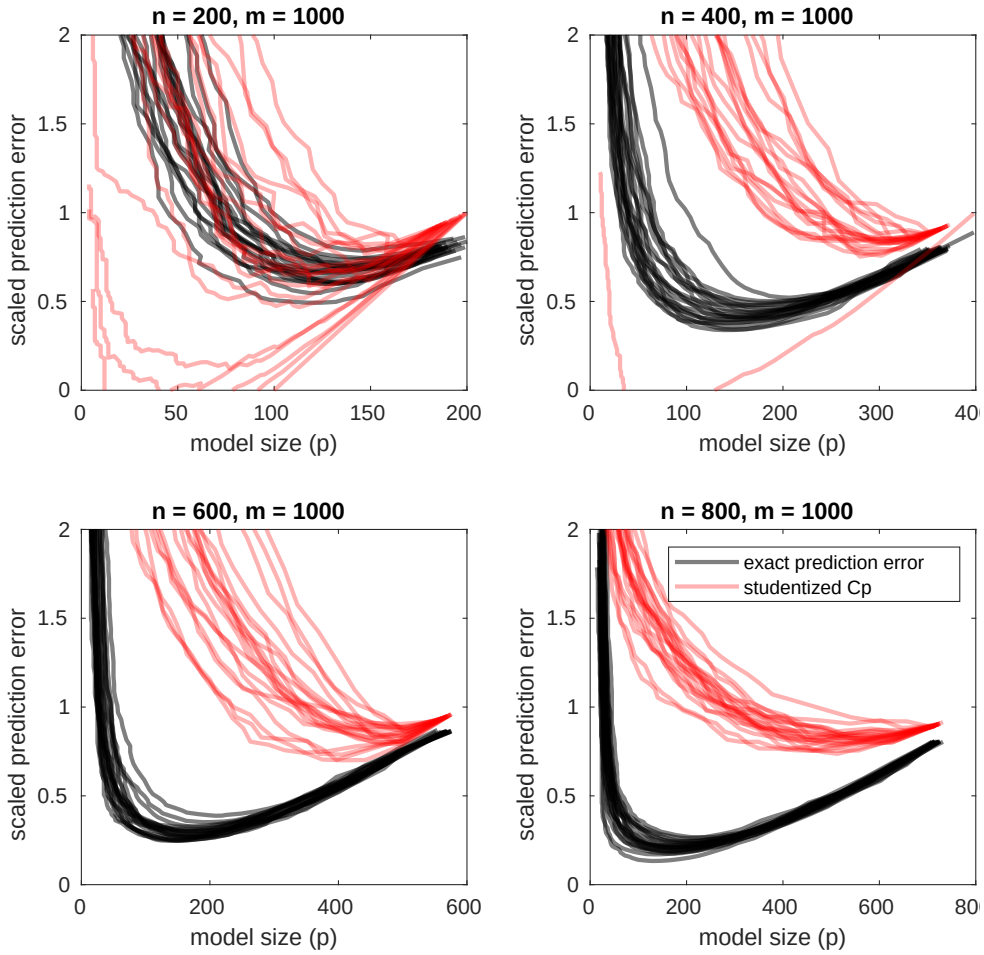

Figure 5: Impact of varying sample size on studentized Mallow's C_p. Each simulation is ran 20 times. As the sample size increases, the variance of the estimation using C_p decreases, however it does not converge to $n/\sigma^2 PE$.

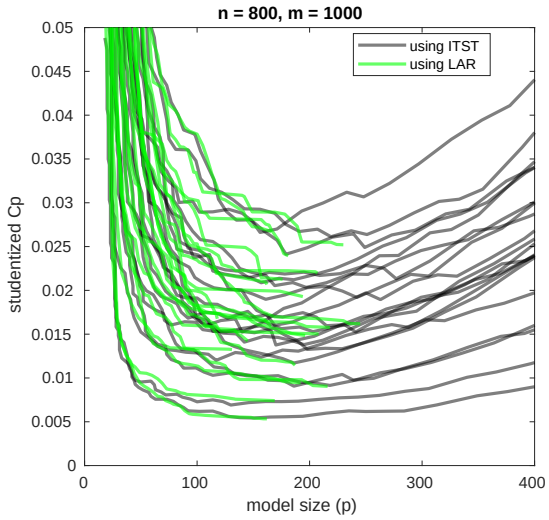

Figure 6: Comparing LAR and ITST solvers at $n = 800$. Solutions are quite similar. Per implemententation, LAR stops as soon as it has reached a minima.

that is updated component wise in a least-squares sense (equation 2). The main difference is that this is done among predictors in a competitive fashion where the updates are carried according to their correlation with the residuals. As a stopping criterion, the minimum of Mallow's C_p is used.
On Figure 6 some simulations results are reported to compare LAR and ITST solvers. Qualitatively, the solutions are quite similar. On Figure 3, two solutions are compared for $n = 800$ at $\min \Delta(\hat{\beta}_p)$. Both estimations capture main β components with some shrinkage but in this case LAR captured a higher number of zeros. Interestingly both solvers are uncapable of capturing some of the smaller components of the real parameter vector and unnecessary components appear as non-zero as well. Again, it is expected that this becomes less and less the case at fixed m when $n \to +\infty$.