# The Illusion of Independence: High Dimensional Data, Shrinkage Methods and Model Selection

Daniel Coutinho Pedro Souza (Orientador)
Marcelo Medeiros (Co-orientador)

November 30, 2017


**Daniel Martins Coutinho**

# The Illusion of Independence: High Dimensional Data, Shrinkage Methods and Model Selection

Monografia de fim de curso

Orientadores: Pedro CL Souza e Marcelo Medeiros

Rio de Janeiro
Novembro de 2017

# Agradecimentos

Aos meus orientadores, Pedro CL Souza e Marcelo Medeiros
A equipe do Digital Lab Lojas Americanas, pelas muitas críticas
Aos meus pais, por todo o apoio
Aos muitos professores centrais na minha formação, em especial Rogério Werneck, Marcio Garcia e Juarez Figueiredo

# Contents

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..." (Dan Ariely)

# Chapter 1

# Introduction

> "The importance of big data doesn't revolve around how much data you have, but what you do with it." - *SAS, one of the biggest data analysis firm in the world*

It is a cliché to say that we live in the era of information and that never before we had access to this amount of data. Yet, using this data is still a challenge. There are a number of methods that were created to use a large number of variables to create models. This models might be used to forecast - forecasting inflation tomorrow, for example - or to infer causality, e.g. understanding the determinants of inflation. By a large number of variables, we mean that there are possibly more variables than observations, a situation in which traditional methods, like Ordinary Least Square, are useless. This kind of situation is called high dimensional data.

Among the many methods to work in a big data environment, the LASSO - which stands for Least Absolute Shrinkage and Select Operator - is one of the most popular methods. The original idea comes from Tibshirani (1996), a paper that was cited 21181 times, a proof of its popularity. The LASSO is a maximum likelihood estimator that penalizes the coefficients. In the case of a linear regression, the LASSO estimator for $\beta$ is:

$$\hat{\beta}_{LASSO}(\lambda) = \arg\min \sum_{i=1}^{n}(y_i - \beta' x_i)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{1.1}$$

Where we have n observations and p covariates and $|.|$ is the absolute value - so $\sum_{j=1}^{p}|\beta_j|$ is the $\ell_1$ norm. The $\ell_1$ norm does not only reduces the absolute value of the coefficients compared to the value of the estimation by OLS, but also performs the selection of variables, i.e. which are relevant to explain or predict values of $y$ and those that are not. We can rewrite 1.1 in another way that can be useful to understand what the LASSO does:

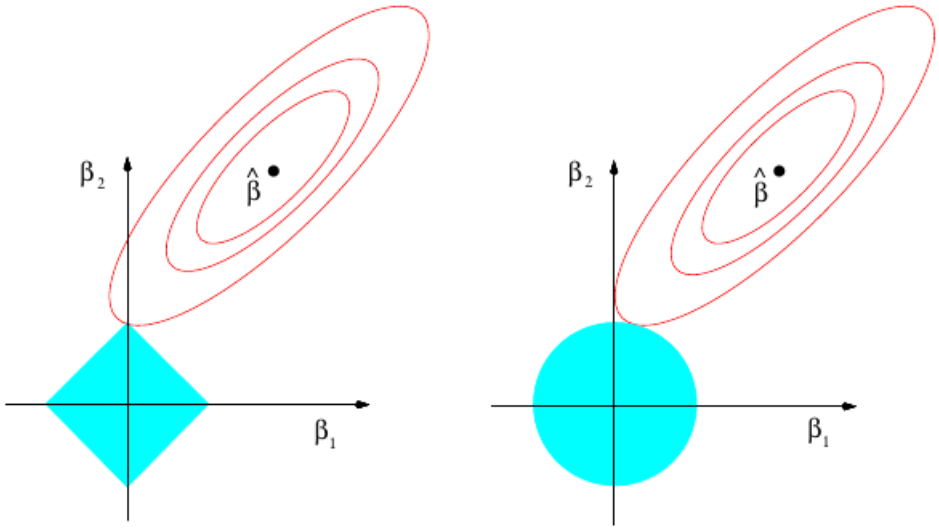
$$\text{Minimize} \sum_{i=1}^{n}(y_i - \beta' x_i)^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq t \tag{1.2}$$

Equation 1.2 makes explicit that the LASSO gives a budget, of value $t$, to the coefficients and that the value of $t$ has to be selected by the researcher. In equation

1.1, $\lambda$ is responsible to control the budget that the coefficients have, so we can drop the variable $t$. Since we have a budget, the coefficients will be shrunken, and some will be set to zero, leaving that variable out of the model.

The selection of variables depends strongly on the $\ell_1$ norm. For example, using the $\ell_2$ norm would lead to a shrinkage of the parameters, but none would be set to exactly zero. Using the $\ell_2$ leads to a estimator known as the ridge. Figure 1.1, from Hastie, Tibshirani and Wainwright (2015), illustrates what each norm does in the case of two variables. When using LASSO, the $\ell_1$ norm generates constraints with corners, which allows to set some coefficients to zero. Since the $\ell_2$ has no corners, no variable is set to exactly zero. On the other hand, one could think of a norm that is smaller than one, which is also possible. However, in this case, the optimization problem is not convex and can become hard to the computer to solve. The fact that the $\ell_1$ norm guarantees a convex problem that makes variable selection explains its popularity.

Figure 1.1: The LASSO constraint (on the right, in blue) vs the ridge constraint (on the left, in blue)




The main parameter to be selected by the researcher is $\lambda$, which is called the shrinkage (or penalty) parameter. If $\lambda \to \infty$, all $\beta$ are set to zero. If $\lambda = 0$, we are back on the OLS regression. Therefore, we need to select the shrinkage parameter in a way that we do not exclude every variable, since we will exclude variables that are actually relevant to explain or predict the variable $y$. On the other hand, we should not let every variable in the model, even when we can estimate it (in the case in which $n < p$), since we would suffer from overfit. In fact, we can be interested in selecting the right variables in the model and excluding every variable that is not relevant to explain $y$. We survey a number of ways to select the $\lambda$ in chapter 2.

The LASSO has several interesting properties concerning forecasting, screening - the inclusion of all relevant variables - and model selection - including all the relevant variables *and* excluding all of the irrelevant variables. While the conditions for good forecasting and screening are rather weak, the conditions for model selection are strong. We show that this conditions are hard to observe, even when we have

variables that are uncorrelated: a small sample, given the dimension of the data, will generate spurious correlation way too often. This is the theme of chapter 3.

In chapter 4, we will also discuss some alternatives estimators based on the LASSO, like the elastic net and the adaptive LASSO (adaLASSO), introduced by Zou (2006). Chapter 4 also discuss forecasting with LASSO and adaLASSO. The last chapter concludes.

# Chapter 2

# Selecting the shrinkage parameter

The main parameter to be selected when using LASSO is the shrinkage parameter, the $\lambda$ in equation 1.1. This controls the "budget" that the coefficients have: the bigger it is, the more the coefficients can increase. If $\lambda$ is too small, important variables to explain the dependent variable may be lost. On the other hand, if it is too big, many irrelevant variables will be included and the model will overfit.

One of the most popular ways to select the shrinkage parameter is using Cross Validation, which is discussed in Hastie, Tibshirani and Wainwright (2015). The idea is to split the data in $K$ groups: use one of them to estimate the model and calculate the sum of squared prediction errors for the other $K-1$ folds of data. Do it for every of the $K$ folds and select the model - which in the end is the value to the shrinkage parameter - that minimizes the squared of the prediction errors for the out of sample folds.

Another possibility is to use some information criteria, such as the Akaike Information Criteria (AIC), or Bayesian Information Criteria (BIC) to select the $\lambda$. For this to work, there must be a way to count how many degrees of freedom are lost. As Hastie, Tibshirani and Wainwright (2015) puts it "Somewhat miraculously, one can show that for the lasso, with a fixed penalty parameter $\lambda$, the number of nonzero coefficients $k_\lambda$ is an unbiased estimate of the degrees of freedom". Due to this result, one can easily apply information criterias to select the shrinkage parameter. We shall test AIC, BIC, the Hannan Quinn Criteria (HQC) and the criteria proposed by Fan and Tang (2013). In general, an information criteria is of the form:

$$\text{measure of model fitting} + a * \text{measure of model complexity} \tag{2.1}$$

For a linear model, a measure of model fit is the sum of squared errors; a measure of the model complexity is the number of non zero parameters. The value of $a$ is what changes between criterias: in AIC, $a = 2$; BIC sets $a = \log(n)$, and HQC $a = 2\log(\log(n))$. The Fan and Tang (2013) proposal is to set $a = \log(\log(n))\log(p)$, and we try a criteria based on this one, that sets $a = 2\log(\log(n))\log(p)$[1].

[1] We call this criteria Fan and Tang (2013)* or HQC*

We also test some other proposals, like the rigorous lasso proposed by Belloni, Chernozhukov and Hansen (2010) and Belloni et. al. (2012); the selection of the shrinkage parameter based on the theory, that indicates that $\lambda \asymp \sqrt{\frac{\log(p)}{n}}$; and the selection of the average $\lambda$, which we describe in more details now.

The average $\lambda$ method relies on the algorithm used on the `glmnet` package, by Jerome Friedman, Trevor Hastie, Noah Simon and Rob Tibshirani. When fitting a model, the algorithm fits the model for a number of lambdas, in such a way that the first model to be fitted has just one of the covariates. The algorithm fits models with different shrinkage parameters until it reaches a model in which all covariates are included. In this way, we have a path with decreasing values for the shrinkage parameter. The average $\lambda$ method takes the average of the path and estimate the model with this penalty parameter.

## 2.1 Simulations

Our simulations use the following DGP:

$$Y = \boldsymbol{X}\beta + \epsilon \tag{2.2}$$

Where $\boldsymbol{X}$ is an $nxp$ matrix, in which $n$ is the number of observations and $p$ is the number of variables. In this simulation, we set $n = 100$ and $p = 50$. We select ten variables to be the relevant variables and we set $\beta = 1$. The other 40 are irrelevant and have $\beta = 0$. Each variable comes from a normal distribution with mean 0 and variance 1. The variables are independent between each other. The error, $\epsilon$, also comes from a normal distribution with mean 0 and variance 1, and is independent from the other variables.

Table 2.1 shows the results of the simulations. Each method for selecting the shrinkage parameter in LASSO was replicated 30,000 times. The "Zeros Right" column shows how often each criteria excluded the irrelevant variables; the "Non Zeros Right" columns shows the proportion that the relevant variables were included and the column "Sparsity" shows the proportion that the criteria selected the right model, i.e. included all relevant variables and excluded the irrelevant variables. Every criteria includes all the relevant variables, so we can do screening with LASSO. However, the story is completely different when we are interested in model selection: the best criteria to select the right model is the HQC*, but it never gets more than 12% right. This is a worrying result: even in the best conditions - independent variables, homoscedasticity, gaussian distribution for everything - none of the methods applied to select $\lambda$ are capable of recovering the right model.

Table 2.3 shows the same simulations of Table 2.1, but we divide the results by the quartile of the maximum sample correlation between the relevant variables and the residual: 4Q is the highest quartile, i.e. the 25% biggest maximum correlation between a irrelevant variable and the residual and 1Q is the lowest quartile. Table 2.2 shows the maximum of the absolute correlation between residual and irrelevant variables for each quantile. Even in the lowest quantile, the correlation is quite high, at 0.19. Although the correlation is still high in the lowest quartile, the recovery of sparsity grows as the correlation falls: this is a strong evidence that our problem comes from the correlation between the residuals and the irrelevant variables.

Table 2.1: LASSO results with different criterias, with n = 100

| | Zeros Right | Non Zeros Right | Sparsity |
|---|---|---|---|
| AIC | 0.44 | 1.00 | 0.00 |
| BIC | 0.82 | 1.00 | 0.01 |
| HQC | 0.67 | 1.00 | 0.00 |
| CV | 0.84 | 1.00 | 0.01 |
| CV Last Block | 0.61 | 1.00 | 0.00 |
| Belloni | 0.92 | 1.00 | 0.04 |
| Avg $\lambda$ | 0.94 | 1.00 | 0.11 |
| Theory Based | 0.92 | 1.00 | 0.03 |
| HQC* | 0.94 | 1.00 | 0.12 |

The sparsity column shows clearly that the higher the sample correlation between the relevant variables and the residual, the worse the performance in recovering the true sparsity structure. This is true for every criteria, and shows a dramatical situation for the LASSO: even when we have variables that are theoretically independent, the LASSO has a terrible performance in recovering the true model. The best performance is from the HQC*, and even in the best case it just recovers the right model in 12% of the cases.

Table 2.2: Max of |Cor(irrelev,res)|, per quartile

| | 4Q | 3Q | 2Q | 1Q |
|---|---|---|---|---|
| LASSO | 0.2984 | 0.2608 | 0.2319 | 0.1928 |

What happens when we have more observations? Tables 2.1 and 2.1 repeats the simulations from table 2.1, but with sample sizes of 5000 observations and 10000 observations, respectively. In fact, the right sparsity is recovered more often for almost every criteria when the sample size grows. When the sample size grows, the average $\lambda$ and the Cross Validation are the best criterias available to recover the true sparsity.

However, the results are rather disappointing: five or ten thousand observations for 50 variables is not a high dimensional setting. As a matter of fact, 50 variables and 100 observations were not a high dimensional situation, and the result for model selection was terrible. Our objective, in the end, is to find methods that can make model selection in a high dimensional situation. So, why LASSO fails at the task of selecting the right variables in a low dimensional, extremely well behaved case? Table 2.3 gives a clue: the higher the correlation between the irrelevant variables and the residual, a relevant but hidden variable. The next chapter shows why this matters and, mostly important, why we observe a correlation between two variables that are, theoretically, independent.

Table 2.3: Maximum absolute correlation between irrelevant variables and residuals: LASSO

| | Sparsity | | | | Zeros Right | | | | Non Zeros Right | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4Q | 3Q | 2Q | 1Q | 4Q | 3Q | 2Q | 1Q | 4Q | 3Q | 2Q | 1Q |
| AIC | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.42 | 0.43 | 0.48 | 1.00 | 1.00 | 1.00 | 1.00 |
| BIC | 0.00 | 0.00 | 0.01 | 0.02 | 0.79 | 0.80 | 0.82 | 0.86 | 1.00 | 1.00 | 1.00 | 1.00 |
| HQC | 0.00 | 0.00 | 0.00 | 0.00 | 0.63 | 0.64 | 0.67 | 0.73 | 1.00 | 1.00 | 1.00 | 1.00 |
| CV | 0.00 | 0.00 | 0.01 | 0.03 | 0.81 | 0.82 | 0.84 | 0.88 | 1.00 | 1.00 | 1.00 | 1.00 |
| CV Last Block | 0.00 | 0.00 | 0.00 | 0.01 | 0.58 | 0.59 | 0.60 | 0.65 | 1.00 | 1.00 | 1.00 | 1.00 |
| Belloni | 0.01 | 0.02 | 0.03 | 0.07 | 0.91 | 0.91 | 0.92 | 0.94 | 1.00 | 1.00 | 1.00 | 1.00 |
| Avg $\lambda$ | 0.06 | 0.08 | 0.11 | 0.18 | 0.93 | 0.93 | 0.94 | 0.95 | 1.00 | 1.00 | 1.00 | 1.00 |
| Theory Based | 0.01 | 0.02 | 0.03 | 0.07 | 0.91 | 0.91 | 0.92 | 0.93 | 1.00 | 1.00 | 1.00 | 1.00 |
| HQC* | 0.08 | 0.10 | 0.13 | 0.16 | 0.93 | 0.94 | 0.94 | 0.95 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 2.4: LASSO, Sample Size = 5000

| | Zeros Right | Non Zeros Right | Sparsity |
|---|---|---|---|
| AIC | 0.63 | 1.00 | 0.00 |
| BIC | 0.95 | 1.00 | 0.19 |
| HQC | 0.87 | 1.00 | 0.03 |
| CV | 1.00 | 1.00 | 0.88 |
| CV Last Block | 0.63 | 1.00 | 0.01 |
| Belloni | 0.93 | 1.00 | 0.06 |
| Avg $\lambda$ | 1.00 | 1.00 | 1.00 |
| Theory Based | 0.95 | 1.00 | 0.13 |
| Fan and Tang (2013)* | 0.98 | 1.00 | 0.43 |

Table 2.5: LASSO, Sample Size = 10000

| | Zeros Right | Non Zeros Right | Sparsity |
|---|---|---|---|
| AIC | 0.64 | 1.00 | 0.00 |
| BIC | 0.96 | 1.00 | 0.22 |
| HQC | 0.87 | 1.00 | 0.03 |
| CV | 1.00 | 1.00 | 0.97 |
| CV Last Block | 0.64 | 1.00 | 0.01 |
| Belloni | 0.93 | 1.00 | 0.06 |
| Avg $\lambda$ | 1.00 | 1.00 | 1.00 |
| Theory Based | 0.95 | 1.00 | 0.13 |
| Fan and Tang (2013)* | 0.98 | 1.00 | 0.44 |

# Chapter 3

# The LASSO and model selection

We now turn to the task of understanding why the LASSO fails at model selection. For model selection, the model has to meet two conditions: the smaller coefficients cannot be too small. Since every relevant variable in our problem has the same $\beta$, this should not be the problem. The other condition is the irrepresentable condition, and it is particularly strong. It states that, for the set S of relevant variables:

$$\max_{j \in S^c} \quad \left| \left( \boldsymbol{X'}_S \boldsymbol{X}_S \right)^{-1} \boldsymbol{X'}_S \boldsymbol{X}_j \right| \leq 1 - \eta \tag{3.1}$$

For some $\eta > 0$. That puts a limitation on the sample covariance between the relevant variables and the irrelevant variables, and a rather strong one: $\left| (\boldsymbol{X'}_S \boldsymbol{X}_S)^{-1} \boldsymbol{X'}_S \boldsymbol{X'}_j \right|$ is the OLS estimator of the regression of the irrelevant variables over the relevant variables, and so it is not bounded like the correlation between two variables. Therefore, it can easily be greater than $1 - \eta$.

However, this should not be a problem, since we are using variables that are not correlated, so the estimate coefficient of the equation 3.1 should be close to zero. We will show in the next section that, since we are taking the maximum, the distribution is not centered in zero, explaining our results.

## 3.1 The distribution of the maximum correlation

We will describe the distribution of the maximum of the absolute value of the correlations between a sample of a normal distribution with n observations of k-independent variables and another sample of a variable with a normal distribution, independent of all the others. In other words, this is a sample with n observations from a k+1 dimensions multivariate normal, where variance-covariance matrix is the diagonal matrix. This will be done in steps: first, we show the distribution of the correlation between two normal, independent, variables. Next, the distribution of the absolute value of this correlation. And finally the maximum between the n normal variables and the n+1 normal variable.

Assume we have two samples i.i.d. from a normal distribution, each of size $n$. We calculate $r$, the correlation between the two samples. From Kendall and Stuart (1960), we have that $t(r) = r\sqrt{\frac{(n-2)}{1-r^2}}$ has student's t distribution with $n-2$ degrees

of freedom. With that, and knowing that the function $t(r)$ is monotone, we know that the distribution of $r$ is:

$$f_r(r) = \frac{1}{\sqrt{(n-2)}B\left(\frac{1}{2}, \frac{(n-2)}{2}\right)}\left(1+\frac{r^2}{(1-r^2)}\right)^{\frac{(1-n)}{2}}\left(\theta + \frac{r^2(n-2)}{(1-r^2)^2\theta}\right) \tag{3.2}$$

Where B is the Beta Function and $\theta = \sqrt{\frac{(n-2)}{1-r^2}}$

We are interested in $r^* = |r|$. The absolute value function can be broke in two separated parts that are monotone, so we can transform $f(r)$ to $f(r^*)$ by breaking $f(r)$ in two parts and than summing it. If $r > 0$, $f(r) = f(r^*)$. Since r always appears squared, $f(r^*) = f(r)$ and so $f(r^*) = 2f(r)$

For the last step, we know that for a vector $y$ of variables $\mathbf{y} = (y_1, y_2, ..., y_k)$ the maximum y distribution is given by $F_{max}(y) = F_y(y)^k$.. So, the density of the maximum of the absolute value of correlation between n normal independent variables and another normal variable, independent from the n other, is:

$$f_{max}(r^*) = kF_{max}(r^*)^{(k-1)}f_r(r^*) \tag{3.3}$$

Figure 3.1 shows the theoretical density and the histogram of the maximum correlation of 50 variables with a sample size 100. Figure 3.3 sets the number of variables to 50 and shows the distribution with a different number of sample sizes. As expected, when the sample grows, the variance decreases and the mode goes to zero.

Figure 3.2 is the most important graphic of this paper: we fix the sample size in 100 observations and we test the maximum correlation between different number of variables and another variable, that were created to be independent. The greater the number of variables we have, the greater is the mean of the distribution of the maximum of the correlation.

Figure 3.2 also shows that, in a high dimensional setting, the maximum correlation between two variables can be quite high. As a consequence, the irrepresentable condition will be hard to achieve. In a way, independence in a high dimensional setting is an illusion: the variables might even be uncorrelated, but we will have a high sample correlation by pure random variations in the sampling. Therefore, theoretical results that assume the correlation between irrelevant and relevant variables is low are interesting, however they are hardly useful.

Figure 3.1: Theoretical distribution (in red) and Monte Carlo Simulation (1000 replications), with n = 100 and k = 50

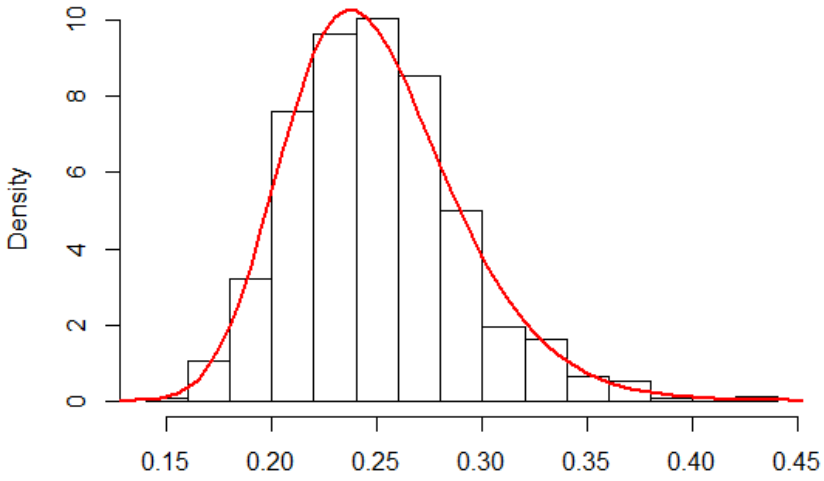



Figure 3.2: Density of the distribution with n = 100 and different number of variables

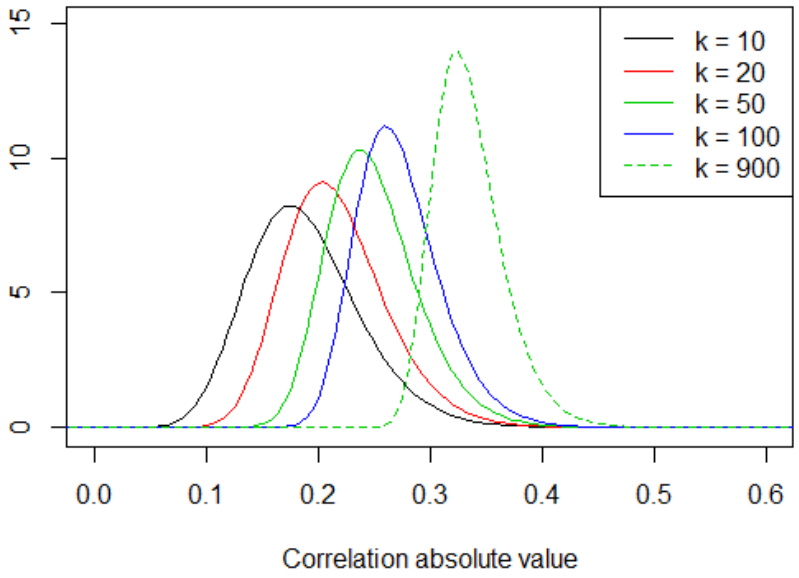
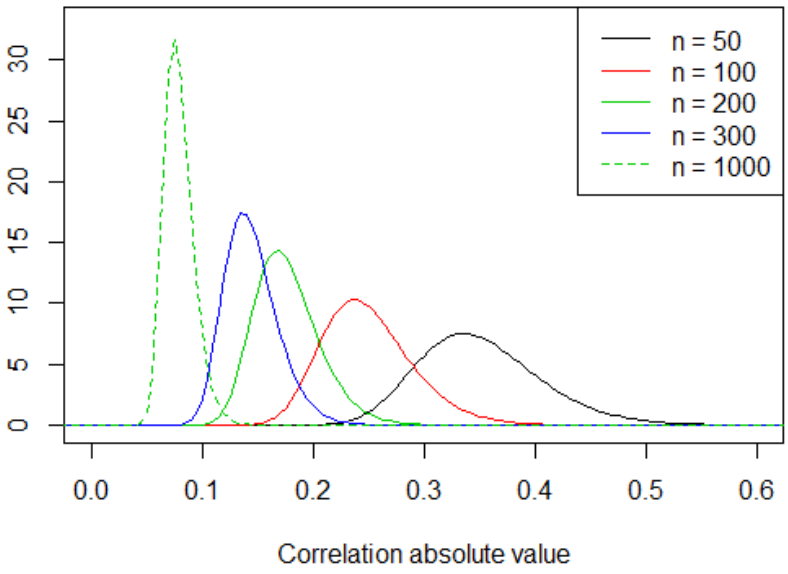


Figure 3.3: Density of the distribution with k = 50 and different sample sizes



# Chapter 4

# Alternative estimators in the LASSO family

The last two chapters showed that LASSO is unable to do model selection. We now search for methods that are able to do variable selection and that have weaker irrepresentable condition. The LASSO generated a number of methods that are closely related to it and that should be able to work with cases that the variables are not uncorrelated. We survey two of them, the elastic net and the adaptive LASSO (adaLASSO). The elastic net is a LASSO that mixes the $\ell_1$ and the $\ell_2$ penalty. Equation 4.1 gives the model to be estimated, and shows that we have a new parameter to select $\alpha$. If $\alpha = 1$, we are back to the usual LASSO; if $\alpha = 0$, we are at the ridge regression.

$$\hat{\beta}(\lambda) = \arg\min \sum_{i=1}^{n} (y_i - \beta' x_i)^2 + \lambda \sum_{j=1}^{p} \alpha|\beta_j| + (1-\alpha)\beta_j^2 \tag{4.1}$$

Hastie, Tibshirani and Wainwright (2015) argues that if we have two covariates that are highly correlated, the LASSO will have a wild behaviour: it will select one variable and leave the other out. The elastic net will select both variables, which makes the model easier to understand. Jia and Yu (2010) give the irrepresentable condition for elastic net. They define the naive elastic net estimator as equation 4.2:

$$\max_{\beta} \left[ \sum_{i=1}^{n} (y_i - X_i\beta)^2 \right] + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2 \tag{4.2}$$

And the naive elastic net selects the same model as the usual elastic net. For the naive elastic net, the irrepresentable condition is given by equation 4.3, where $J_1$ is the set of relevant variables, and $\eta$, as in 3.1 is a constant and $\eta > 0$. Equation 4.3 is really similar to the irrepresentable condition of LASSO, but the expression $\frac{\lambda_2}{n} I$ in the equation such reduce the value of the coefficients and thus the bound $1 - \eta$ should be attained in more cases.

$$\max \left| n^{-1} X'_{J_1^c} X_{J_1} \left( n^{-1} X'_{J_1} X_{J_1} + \frac{\lambda_2}{n} I \right)^{-1} \left( \text{signal}(\beta_{J_1}) + 2\frac{\lambda_2}{\lambda_1} \beta_{J_1} \right) \right| \leq 1 - \eta \tag{4.3}$$

The results for the simulations with elastic net are shown in Table 4.1. The procedure is slow, since we set an value for $\alpha$ and select the best $\lambda$ for that alpha. When we use an information criteria or Cross Validation, selecting the best $\alpha$ is done simply by selecting the smaller value of the information criteria or of the cross validation error, respectively. In the cases in which the penalization parameter is selected by the Average $\lambda$ and Theory Based methods, we select the best $\alpha$ using BIC. In every case, elastic net does not improve the performance.

Table 4.1: Elastic net with n = 100

| | Zeros Right | Non Zeros Right | Correct Model |
|---|---|---|---|
| AIC | 0.44 | 1.00 | 0.00 |
| BIC | 0.83 | 1.00 | 0.01 |
| HQC | 0.67 | 1.00 | 0.00 |
| CV | 0.79 | 1.00 | 0.01 |
| Avg $\lambda$ | 0.94 | 1.00 | 0.13 |
| Theory | 0.92 | 1.00 | 0.04 |
| Fan and Tang (2013)* | 0.94 | 1.00 | 0.12 |

Table 4.2 shows the number of times, in a thousand simulations, that each value of alpha was selected. Most of the times, alpha is set to one - with the notable exception of Cross Validation - and so most of the times, we are actually using LASSO. We also tested spacing alphas by 0.05 instead of 0.1, but the results were unchanged.

Table 4.2: Selection of $\alpha$ in the Elastic Net

| $\alpha$ | AIC | BIC | HQC | CV | Avg $\lambda$ | Theory Based | Fan and Tang (2013)* |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0.2 | 2 | 0 | 1 | 6 | 0 | 0 | 0 |
| 0.3 | 6 | 0 | 1 | 13 | 0 | 0 | 0 |
| 0.4 | 9 | 0 | 0 | 38 | 2 | 0 | 0 |
| 0.5 | 19 | 1 | 4 | 76 | 3 | 0 | 0 |
| 0.6 | 51 | 3 | 14 | 96 | 2 | 2 | 0 |
| 0.7 | 50 | 14 | 23 | 136 | 5 | 21 | 7 |
| 0.8 | 43 | 14 | 33 | 185 | 6 | 70 | 8 |
| 0.9 | 43 | 23 | 35 | 202 | 17 | 286 | 11 |
| 1 | 776 | 945 | 889 | 247 | 965 | 621 | 974 |

The second option is the adaptive LASSO, which is a LASSO in two stages: first,

estimate the model using LASSO. Then, use the coefficients obtained in this first stage (call it $\beta^l$) to say how much they should be penalized. In a way, now every coefficient has its own $\lambda$. To do that, we re-write equation 1.1 as:

$$\hat{\beta}(\lambda) = \arg\min \sum_{i=1}^{n} (y_i - \beta' x_i)^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \tag{4.4}$$

All we are doing is setting an weight for the penalty of each parameter. If we set $\omega_j = 1$ for every j, we are back in equation 1.1. If $\omega_j = 0$, than the parameter won't be penalized at all. On the other hand, setting $\omega_j$ to a value that is high, the $\beta_j$ will be more shrunken toward zero than it would be in the usual LASSO. We can always set $\omega_j$ to an arbitrary value but this gives no guarantee that will be doing something reasonable. What we will do, based on Zou (2006), is to set the weight as a function of the $\beta_j^l$, the coefficients estimated by our first stage LASSO, e.g. $\omega_j = \frac{1}{|\beta_j^l|}$. In this case, if the first stage LASSO has excluded a variable, $\omega_j \to \infty$ and the variable will be excluded, since any $\beta_j^{adaLASSO} \neq 0$, any "budget" will be exhausted. On the other hand, if $\beta_j^l$ is distant of zero, $\omega_j$ is small, and a high value of $\beta_j^{adaLASSO}$ will not spend much of the "budget".

Equation 4.5 from Huang et. al (2008) gives an irrepresentable condition for adaLASSO. Let $\eta_j$ be the weight for the second stage of adaLASSO and $s_1 = (|\eta_j|^{-1}\text{signal}(\beta_{0j}), j \in J_1)'$. Then, for $\kappa < 1$:

$$n^{-1}|x_j' X_{J_1}(n^{-1}X_{J_1}' X_{J_1})^{-1} s_1| \leq \frac{\kappa}{\eta_j} \quad \forall j \notin J_1 \tag{4.5}$$

The $\kappa$ from equation 4.5 does the same job as the $1 - \eta$ from equation 3.1, the irrepresentable condition for LASSO. The $s_1$ and $\eta_j$ under the $\kappa$, that is for the irrelevant variables, help to relax the irrepresentable condition by reducing the $\beta$ of the projection of the relevant variables over the irrelevant variables and the bound of the irrepresentable condition, thus making it easier to attain the bound.

Since adaLASSO is a two-stage procedure with two LASSO estimations, we have to select the shrinkage parameter twice in every fit. This can become quite cumbersome: if we have 2 criterias, there are four ways to select $\lambda$. We can use the first criteria in both stages, the second criteria in both stages, the first criteria in the first stage and the second criteria in the second stage and, last, the second criteria in the first stage and the first criteria in the second stage. If we have 10 criterias, we get 100 ways to select $\lambda$. To keep the amount of computations acceptable, we choose to select the $\lambda$ in the first stage using one of the criterias and *use the same $\lambda$ in the second stage*.

The results for the fixed $\lambda$ approach with a sample size of 100 observations is shown in Table 4.3. We repeat the exercise of breaking the results for every criteria by the maximum absolute value of the correlation between the irrelevant variables and error in table 4.7. Table 4.6 shows the correlation in the quartiles of table 4.7. The results for model selection are much better than for the LASSO, even when the correlation is quite high between irrelevant variables and the error. This improvement is a direct consequence of the adaptive LASSO irrepresentable condition, that is easier to attain. Tables 4.4 and 4.5 show the performance of the adaLASSO when the sample size increases to 500 and 1000 observations, respectively. As expected, the performance is much better when the sample grows.

We also test using the post estimation procedure proposed by Belloni and Chernozhukov (2013). The idea is simple: after doing the LASSO selection, we run an OLS with the selected variables. On a first look, this should only correct the bias and it shouldn't change the variables selected. However, the Belloni, Chernozhukov and Hansen (2010) criteria depends of the variance of the error, that is calculated in an iterative manner. So for every step of iteration, post estimation uses LASSO to select the variables, but calculates the value of the coefficients using OLS. The value of the shrinkage parameter is a function of the variance of the residuals and the variance of the residuals is a function of the coefficients. In this way, using OLS in every iteration changes the shrinkage parameter.[1] Table 4.8 shows the results of using the post estimation scheme. There is a gain of 40% in the selection of the right model in the case of 100 observations. The gain is greater when we have more observations.

Table 4.3: adaLASSO with sample size 100, 1000 simulations, fixed $\lambda$

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.59 | 1.00 | 0.01 |
| BIC | 0.95 | 1.00 | 0.43 |
| HQC | 0.83 | 1.00 | 0.13 |
| CV | 0.93 | 1.00 | 0.30 |
| CV Last Block | 0.76 | 1.00 | 0.12 |
| Avg $\lambda$ | 1.00 | 1.00 | 0.86 |
| Theory Based | 0.99 | 1.00 | 0.82 |
| Fan and Tang (2013)* | 1.00 | 1.00 | 0.92 |

Table 4.4: adaLASSO with sample size 500, 1000 simulation, fixed $\lambda$

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.78 | 1.00 | 0.06 |
| BIC | 0.99 | 1.00 | 0.85 |
| HQC | 0.95 | 1.00 | 0.46 |
| CV | 0.99 | 1.00 | 0.77 |
| CV Last Block | 0.77 | 1.00 | 0.17 |
| Avg $\lambda$ | 1.00 | 1.00 | 1.00 |
| Theory Based | 1.00 | 1.00 | 0.92 |
| Fan and Tang (2013)* | 1.00 | 1.00 | 1.00 |

[1] We thanks Martin Spindler, the maintainer of the `hdm` package, for the explanation

Table 4.5: adaLASSO, sample size = 1000, 1000 simulations, fixed $\lambda$

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.80 | 1.00 | 0.07 |
| BIC | 1.00 | 1.00 | 0.92 |
| HQC | 0.97 | 1.00 | 0.53 |
| CV | 1.00 | 1.00 | 0.90 |
| CV Last Block | 0.77 | 1.00 | 0.18 |
| Avg $\lambda$ | 1.00 | 1.00 | 1.00 |
| Theory Based | 1.00 | 1.00 | 0.92 |
| Fan and Tang (2013)* | 1.00 | 1.00 | 1.00 |

Table 4.6: Max of |Cor(irrelev,res)|, per quartile

| | 4Q | 3Q | 2Q | 1Q |
|---|---|---|---|---|
| adaLASSO | 0.2984 | 0.2605 | 0.2316 | 0.1926 |

## 4.1 How good is adaLASSO?

In this section we show how robust our results are for the adaptive LASSO with fixed $\lambda$, the method that shows some good performance in the easy case. We show variations in two parameters: the variance of the error and the number of irrelevant variables. We begin increasing the variance of the error, that until now was set to one. Table 4.9 shows the results of setting the error variance, $\sigma_\epsilon^2$ and changing the number of observations, $n$.

With $n = 100$, increasing the variance of the error reduces a lot the capacity of every criteria to get the right model. With $\sigma_\epsilon^2 = 3$, some criterias - like BIC and HQC* - start to throw away variables that are relevant. However, a 500 observations sample is enough to avoid this problems and now the criterias do not lose so much of their capacity to choose the right model. A notable exception is the Theory Based criteria, and it is easy to explain why it goes so bad: every other criteria uses the data to select the shrinkage parameter. Even average $\lambda$ relies on the glmnet algorithm, and the glmnet selects the $\lambda$ path based on the data. Therefore, every other criteria is able to correct itself to the fact that the variance of the residual should be higher - which is reflected in a higher variance in the dependent variable. The only $\lambda$ that is completely data independent is the one based on the theory.

Tables 4.10 and 4.11 show the performance of each criteria, but now adding 90 and 150 irrelevant variables, respectively, while keeping the number of observations at 100. BIC gets a lot worse in the 90 irrelevant variables case, but other criterias still have a good performance. With 150 irrelevant variables, a case in which we have more variables than observations, the performance of the Theory Based, Average $\lambda$, HQC* and Fan and Tang (2013) for model selection is good, with the criterias selecting the right model in more the 60% of the simulations.

Tables 4.12 and 4.13 show the cases for 1000 observations. We set 100 relevant

Table 4.7: Maximum absolute correlation between irrelevant variables and residuals: adaptive LASSO

| | Sparsity | | | | Zeros Right | | | | Non Zeros Right | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4Q | 3Q | 2Q | 1Q | 4Q | 3Q | 2Q | 1Q | 4Q | 3Q | 2Q | 1Q |
| AIC | 0.00 | 0.00 | 0.01 | 0.03 | 0.56 | 0.56 | 0.58 | 0.64 | 1.00 | 1.00 | 1.00 | 1.00 |
| BIC | 0.21 | 0.32 | 0.46 | 0.67 | 0.93 | 0.94 | 0.95 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 |
| HQC | 0.03 | 0.06 | 0.12 | 0.28 | 0.79 | 0.80 | 0.83 | 0.88 | 1.00 | 1.00 | 1.00 | 1.00 |
| CV | 0.12 | 0.20 | 0.31 | 0.52 | 0.91 | 0.92 | 0.93 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 |
| CV Last Block | 0.06 | 0.08 | 0.12 | 0.19 | 0.73 | 0.74 | 0.76 | 0.79 | 1.00 | 1.00 | 1.00 | 1.00 |
| Belloni | 0.39 | 0.43 | 0.47 | 0.52 | 0.98 | 0.98 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 |
| Avg $\lambda$ | 0.80 | 0.85 | 0.87 | 0.91 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Theory Based | 0.72 | 0.80 | 0.84 | 0.91 | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| HQC* | 0.85 | 0.92 | 0.94 | 0.97 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 4.8: Results for Belloni with post estimation

| Case | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| Obs = 100 | 0.98 | 0.99 | 0.46 |
| Obs = 500 | 1.00 | 1.00 | 0.98 |
| Obs = 1000 | 1.00 | 1.00 | 0.99 |

variables with $\beta = 1$ and the variance of the error is fixed to 1. The first table shows the results for 400 irrelevant variables and the second for 900 irrelevant variables. In both cases, the criterias still work well, with BIC been the only one that selects the wrong model more than 30% of times.

Tables 4.14 and 4.15 repeat the exercise of Tables 4.12 and 4.13, but now with the variance of the error set to 10. In this way, we keep the signal to noise ration equal to the case of 100 observations and 10 relevant variables. With 400 irrelevant variables, the only criterias that are not selecting the model wrong almost always are the Average $\lambda$ and the Fan and Tang (2013). The case with 900 irrelevant variables is even more dramatic, and the only criteria that is not near zero is the Fan and Tang (2013) proposal.

## 4.2 Forecasting

Another common application for machine learning methods, like the LASSO, is forecasting. There are a number of theorems about the performance of LASSO in forecasting, with conditions that are weaker than the conditions for model selection. One could imagine that, since LASSO is good at forecasting and it is not able to

Table 4.9: Selection of the right model by adaLASSO

| | $n = 100$ | | $n = 500$ | |
|---|---|---|---|---|
| | $\sigma_\epsilon^2 = 2$ | $\sigma_\epsilon^2 = 3$ | $\sigma_\epsilon^2 = 2$ | $\sigma_\epsilon^2 = 3$ |
| AIC | 0.01 | 0.01 | 0.07 | 0.06 |
| BIC | 0.37 | 0.29 | 0.85 | 0.85 |
| HQC | 0.11 | 0.10 | 0.45 | 0.44 |
| CV | 0.27 | 0.21 | 0.75 | 0.79 |
| CV Last Block | 0.11 | 0.08 | 0.18 | 0.17 |
| Avg $\lambda$ | 0.48 | 0.23 | 1.00 | 1.00 |
| Theory Based | 0.34 | 0.08 | 0.40 | 0.09 |
| HQC* | 0.63 | 0.27 | 1.00 | 1.00 |
| Fan and Tang (2013) | 0.55 | 0.41 | 0.89 | 0.90 |

do model selection, adaLASSO should be as good in forecasting as LASSO, since it solves a problem that LASSO is unable to solve. However, there is no theoretical guarantee of this, so in this section we make simulations to understand the performance of each method for forecasting.

We will use the same setup: 60 independent variables in which 10 are relevant and have $\beta = 1$ and the other 50 are irrelevant and have $\beta = 0$. We will use a fixed window that starts with 100 observations and we have a total of 200 observations. We will do a one step ahead forecast, and since we are not using time series data, this does not make any big difference. We show only the results for the criterias that have good performance in the model selection problem, namely BIC, HQC, Cross Validation, Belloni, Theory Based, Average $\lambda$ and Fan and Tang (2013). Table 4.16 shows the MAE and MSE for this simulation, which was repeated a thousand times.

In this situation, adaLASSO is always better than LASSO. However, the difference now is really small, especially compared with the difference observed in the performance of each method in model selection. This results provides no guidance that in other situations - e.g., using time series models, in which the variables are not i.i.d. - the adaLASSO is better than the LASSO.

An interesting point is that some criterias that are not that good at model selection do a better work in forecasting than criterias that are good at model selection. The HQC has a smaller MSE and MAE than Average $\lambda$, although Average $\lambda$ is better than HQC in model selection.

Table 4.10: adaLASSO, n=100 with 90 irrelevant variables

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.05 | 1.00 | 0.00 |
| BIC | 0.20 | 1.00 | 0.07 |
| HQC | 0.06 | 1.00 | 0.00 |
| CV | 0.96 | 1.00 | 0.30 |
| CV Last Block | 0.85 | 1.00 | 0.12 |
| Avg $\lambda$ | 0.99 | 1.00 | 0.55 |
| Theory Based | 0.99 | 1.00 | 0.76 |
| HQC* | 1.00 | 0.97 | 0.86 |
| Fan and Tang (2013) | 0.99 | 1.00 | 0.73 |

Table 4.11: adaLASSO, n=100 with 150 irrelevant variables

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.59 | 1.00 | 0.00 |
| BIC | 0.78 | 1.00 | 0.32 |
| HQC | 0.59 | 1.00 | 0.00 |
| CV | 0.97 | 1.00 | 0.28 |
| CV Last Block | 0.89 | 1.00 | 0.12 |
| Avg $\lambda$ | 1.00 | 0.95 | 0.64 |
| Theory Based | 1.00 | 1.00 | 0.72 |
| HQC* | 1.00 | 0.90 | 0.71 |
| Fan and Tang (2013) | 1.00 | 1.00 | 0.84 |

Table 4.12: 1000 observations, 400 irrelevant variables, $\sigma_\epsilon^2 = 1$

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.53 | 1.00 | 0.00 |
| BIC | 0.98 | 1.00 | 0.12 |
| HQC | 0.90 | 1.00 | 0.00 |
| CV | 0.87 | 1.00 | 0.00 |
| CV Last Block | 0.76 | 1.00 | 0.00 |
| Avg $\lambda$ | 1.00 | 1.00 | 0.77 |
| Theory Based | 1.00 | 1.00 | 0.80 |
| HQC* | 1.00 | 1.00 | 0.89 |
| Fan and Tang (2013) | 0.99 | 1.00 | 0.53 |

Table 4.13: 1000 observations, 900 irrelevant variables, $\sigma_\epsilon^2 = 1$

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.19 | 1.00 | 0.00 |
| BIC | 0.99 | 1.00 | 0.34 |
| HQC | 0.96 | 1.00 | 0.01 |
| CV | 0.92 | 1.00 | 0.00 |
| CV Last Block | 0.86 | 1.00 | 0.00 |
| Avg $\lambda$ | 1.00 | 1.00 | 0.80 |
| Theory Based | 1.00 | 1.00 | 0.73 |
| HQC* | 1.00 | 1.00 | 0.94 |
| Fan and Tang (2013) | 1.00 | 1.00 | 0.74 |

Table 4.14: adaLASSO, n=1000, 400 irrelevant variables, $\sigma_\epsilon^2 = 10$

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.53 | 1.00 | 0.00 |
| BIC | 0.98 | 1.00 | 0.06 |
| HQC | 0.89 | 1.00 | 0.00 |
| CV | 0.86 | 1.00 | 0.00 |
| CV Last Block | 0.76 | 1.00 | 0.00 |
| Avg $\lambda$ | 1.00 | 1.00 | 0.68 |
| Theory Based | 0.76 | 1.00 | 0.00 |
| HQC* | 1.00 | 0.03 | 0.02 |
| Fan and Tang (2013) | 0.99 | 1.00 | 0.37 |

Table 4.15: adaLASSO, n=1000, 900 irrelevant variables,$\sigma_\epsilon^2 = 10$

| | Zeros Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| AIC | 0.01 | 1.00 | 0.00 |
| BIC | 0.99 | 1.00 | 0.17 |
| HQC | 0.01 | 1.00 | 0.00 |
| CV | 0.92 | 1.00 | 0.00 |
| CV Last Block | 0.86 | 1.00 | 0.00 |
| Avg $\lambda$ | 0.99 | 1.00 | 0.09 |
| Theory Based | 0.82 | 1.00 | 0.00 |
| HQC* | 1.00 | 0.00 | 0.00 |
| Fan and Tang (2013) | 1.00 | 1.00 | 0.54 |

Table 4.16: MAE and MSE for LASSO and adaLASSO (with fixed $\lambda$)

| | Lasso, MSE | Lasso, MAE | adaLASSO MSE | adaLASSO MAE |
|---|---|---|---|---|
| BIC | 1.51 | 0.98 | 1.25 | 0.89 |
| HQC | 1.49 | 0.97 | 1.37 | 0.93 |
| CV | 1.57 | 1.00 | 1.23 | 0.88 |
| Belloni | 1.86 | 1.08 | 1.87 | 1.06 |
| Avg $\lambda$ | 1.83 | 1.08 | 1.49 | 0.96 |
| Theory Based | 1.78 | 1.06 | 1.36 | 0.92 |
| HQC* | 1.86 | 1.08 | 1.86 | 1.06 |
| Fan and Tang (2013) | 1.59 | 1.00 | 1.28 | 0.90 |

Table 4.17: MSE and MAE for forecast

# Chapter 5

# Conclusion

Variable selection is a big problem in a world in which we have many candidate variables that might me useful to understand a process or forecast it. What are determinants of inflation? What are the most important variables to explain why we are so rich and they so poor? All this questions boils down to selecting the right set of variable.

Unfortunately, LASSO is unreliable for model selection, even in the situation in which we have independent, gaussian i.i.d variables - in which we have good asymptotical results for model selection. In finite samples with high dimensional data, spurious correlation appears way too often: the variables might be independent, but the correlation is way too high and the independence is illusory. This illusory independence makes LASSO unable to throw all the irrelevant variables out, as shown in chapter 3. Fortunately, adaLASSO is good enough to select the right model in most of cases studied.

There are a number of criterias that are good in selecting the right model when using adaLASSO, and no criteria dominates all the other. Some criterias, as the AIC, are not useful at all. On the other hand, Cross Validation - one of the most widely used methods for selecting the shrinkage parameter - fares relatively well. The Bayesian Information Criteria, Hannan Quinn Criteria, Belloni's method, Fan and Tang (2013), HQC*, Average $\lambda$ and Theory Based methods are reliable, even using the same shrinkage parameter for the second stage. Adaptive LASSO also outperforms the LASSO in forecasting. Nonetheless, the difference in performance is not as large as in the model selection case.

# Bibliography


BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): **Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain**, Econometrica, 80, 2369?2429, Arxiv, 2010.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2010): **Inference for High-Dimensional Sparse Econometric Models,** Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010, III, 245-295, ArXiv, 2011.

BELLONI, A., CHERNOZHUKOV, V..**Least squares after model selection in high-dimensional sparse models.** Bernoulli, 19(2), 521-547, 2013.

HUANG, J., MA, S., ZHANG, C.H.. **Adaptive Lasso for sparse high-dimensional regression models.** Statistica Sinica, p. 1603-1618, 2008.

JIA, J., YU, B.. **On model selection consistency of the elastic net when** $p \gg n$. Statistica Sinica, 595-611, 2010.

KENDALL, M.G., STUART, A. **The Advanced Theory of Statistics**, vol 2., 3rd edition, 1960. p. 475

TIBSHIRANI, R. **Regression shrinkage and selection via the lasso**. Journal of the Royal Statistical Society. Series B (Methodological), 267-288, 1996

HASTIE,T., TIBSHIRANI, R., WAINWRIGHT, M. **Statistical Learning with Sparsity**, 2015.

FAN, Y., TANG, C. Y. **Tuning parameter selection in high dimensional penalized likelihood.** Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(3), 531-552, 2013.

ZOU, Hui. **The adaptive lasso and its oracle properties.** Journal of the American statistical association, v. 101, n. 476, p. 1418-1429, 2006.

# Appendix: Fan Tang

This appendix shows the correct results for the Fan and Tang (2013) criteria, which uses $a = \log(\log(n))\log(p)$. An old version of this monograph showed the wrong values.

Table 5.1: Fan Tang criteria, performance by the quantile of the correlation of residual with the irrelevant variables

| | Max \|Cor(Irrelev, res)\| | Zeros Right | Non Zeros Right | Right Model |
|---|---|---|---|---|
| 4Q | 0.30 | 0.85 | 1.00 | 0.01 |
| 3Q | 0.25 | 0.86 | 1.00 | 0.01 |
| 2Q | 0.23 | 0.88 | 1.00 | 0.02 |
| 1Q | 0.19 | 0.90 | 1.00 | 0.04 |

Table 5.2: Fan and Tang(2011) criteria, applied for LASSO

| Obs | Zeros,Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| 100 | 0.868 | 1.000 | 0.018 |
| 5000 | 0.945 | 1.000 | 0.145 |
| 10000 | 0.946 | 1.000 | 0.155 |

Table 5.3: Fan and Tang(2011) criteria, applied for adaLASSO,fixed $\lambda$

| Obs | Zeros,Right | Non Zeros Right | Sparsity Right |
|---|---|---|---|
| 500 | 0.996 | 1.000 | 0.902 |
| 1000 | 0.998 | 1.000 | 0.937 |

Table 5.4: Fan and Tang(2011) criteria, adaLASSO with fixed $\lambda$, by quartile of Max(|Cor(residual,irrelevant variable)|). n = 100

| | Max(\|Cor(Res,Irrelev)\|) | Zeros Right | Non Zeros Right | Sparsity |
|---|---|---|---|---|
| 4Q | 0.298 | 0.968 | 1.000 | 0.410 |
| 3Q | 0.260 | 0.977 | 1.000 | 0.576 |
| 2Q | 0.231 | 0.985 | 1.000 | 0.707 |
| 1Q | 0.192 | 0.993 | 1.000 | 0.859 |