4 Regularization and variable selection with LASSO and CVaR penalty

4.1. Introduction

Besides interpretability, an important concern in variable selection is the predictive accuracy of selected model. Accurate out-of-sample forecast can be difficult to get if the data set has outliers. When outliers are in the in-sample set, selecting the relevant variables of the model becomes a difficult problem, with the risk of including variables that are not in the “true” model, but can explain the outliers, called in this thesis “spurious variables”. In that case, inclusion of “spurious variables” improves in-sample forecast, but out-of-sample performance will not be satisfactory. In order to avoid this “spurious” selection, we propose an extension of LASSO methodology robust to outliers. The idea is to add a Conditional Value at Risk (CVaR) of “out-of-sample” errors term to the LASSO ℓ_1-penalty, as explained in next section.

CVaR, or expected shortfall, is a risk measure widely used in the recent literature. Known to have better properties than VaR (Value at Risk), CVaR can capture events deep in the tail of the distribution (catastrophic events). Generally speaking, CVaR is the conditional expected value of losses above the VaR. For more details see Alexander and Baptista (2004).

Rockafellar and Uryasev (2000) proposed a technique that calculates the VaR and optimizes CVaR simultaneously, formulating the CVaR as a linear optimization problem. The formulation for the CVaR of a random variable $\boldsymbol{R}$, proposed by the authors, is presented in (31):

$$\mathrm{CVaR}_{\alpha}(\boldsymbol{R}) = \min_{(z,\delta_k)} \left\{ \mathrm{z} + \frac{1}{K(1-\alpha)} \sum_{k=1}^{K} \delta_k \right\}$$

subject to: (31)

$$\mathrm{z} \geq 0, \qquad \forall\, k = 1, \dots, K$$

$$\delta_k \geq R_k - \mathrm{z}, \qquad \forall\, k = 1, \dots, K$$

where z is the VaR_{α} and α is the CVaR confidence level. Common values for α are 0.95 and 0.99.

In this chapter, we use the risk measure CVaR, as presented in (31), in a variable selection problem.

4.2. LASSO-CVaR

We propose a penalized least square criterion based on the LASSO ℓ_1-penalty and CVaR (Conditional Value at Risk) of out-of-sample regression errors. We call this approach LASSO-CVaR. The idea is to select variables controlling the model out-of-sample performance, therefore, we add a CVaR of out-of-sample errors in the penalty term. We believe that LASSO-CVaR method will be capable to identify outliers, not selecting "spurious variables" that would increase the out-of-sample error.

The CVaR_{α} term in the penalty will minimize the expected value of the largest $(1-\alpha)\%$ errors out-of-sample, in other words, the CVaR_{α} will be the conditional expected value of the out-of-sample errors larger than a VaR_{α}value z. In this work, we chose to set α at 75%.

Consider model estimation and variable selection in a linear regression framework. Suppose that $\mathbf{y} = (y_1, \dots, y_{T_{in}})'$ is the response vector, and $\mathbf{x}_j = (x_{j1}, \dots, x_{jT_{in}})'$, with $j = 1, \dots, p$, are the predictor variables. Suppose also $\varepsilon_t = y_t - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_t$ and $\boldsymbol{\varepsilon} = (\varepsilon_{T_{in}+1}, \dots, \varepsilon_{T_{out}})'$ is the "out-of-sample" errors vector. The LASSO-CVaR estimator is given by (32):

$$\widehat{\boldsymbol{\beta}}^{LASSO-CVaR} = arg \min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \gamma \mathrm{CVaR}_{\alpha}(|\boldsymbol{\varepsilon}|) \qquad (32)$$

where $\|.\|$ denotes the standard ℓ_2-norm in the "in-sample" set, λ is the nonnegative regularization parameter of the ℓ_1-penalty (or LASSO penalty) and γ is a nonnegative parameter that gives the weight of CVaR term in the penalty term. More clearly, (32) can be represented as:

$$\widehat{\boldsymbol{\beta}}^{LASSO-CVaR} = arg\min_{\beta} \sum_{t=1}^{T_{in}} \left(y_t - \sum_{j=1}^{p} x_{jt}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \gamma \mathrm{CVaR}_{\alpha}\{|\varepsilon_t|\}_{t=T_{in}+1}^{T_{out}} \tag{33}$$

where $t = 1, \dots, T_{in}, T_{in}+1, \dots, T_{out}$. T_{in} and T_{out} represent the number of "in-sample" and "out-of-sample" observations, respectively. In this context, we are using the quotation marks (" ") for "in-sample" and "out-of-sample" because we want to emphasize these are subsets of observations within the original in-sample set. We will always have the true out-of-sample set unknown at the moment of model estimation. In equation (33) the sets $\{1, \dots, T_{in}\}$ and $\{T_{in}+1, \dots, T_{out}\}$ can be seen as the training and validation sets in neural networks context.

Using the CVaR formulation proposed in Rockafellar and Uryasev (2000) in eq. (31), and taking the vector of explicative variables $\boldsymbol{x}_t = (x_{1t}, \dots, x_{pt})$, we can rewrite (33) as a quadratic optimization problem as follows:

$$\min_{(\boldsymbol{\beta}, \delta_t, \mathrm{z})} \sum_{t=1}^{T_{in}} (y_t - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_t)^2 + \lambda|\boldsymbol{\beta}| + \gamma\left(\mathrm{z} + \frac{1}{T_{out}(1-\alpha)} \sum_{t=T_{in}+1}^{T_{out}} \delta_t \right)$$

subject to: (34)

$$\begin{aligned} &\delta_t \geq 0, && \forall\, t = T_{in}+1, \dots T_{out} \\ &\delta_t \geq |y_t - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_t| - \mathrm{z}, && \forall\, t = T_{in}+1, \dots T_{out} \end{aligned}$$

4.2.1. Theoretical results for the parameter $\boldsymbol{\lambda}$

Let $\mathbf{y} = (y_1, \dots, y_T)'$ be the response vector and $\mathbf{X}$ be the $T{\times}p$ matrix of predictors, with row $\boldsymbol{x}_t = (x_{1t}, \dots, x_{pt})$, and column $\mathbf{x}_j = (x_{j1}, \dots, x_{jT})'$. Consider the original LASSO problem in (35) which is similar to (1):

$$\widehat{\boldsymbol{\beta}}^{LASSO} = arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{35}$$

where $||.||_2$ denotes the ℓ_2-norm and $||.||_1$ denotes the ℓ_1-norm.

Equation (35) represents a convex but not differentiable function due to the ℓ_1-norm. Effectively the function $f(\boldsymbol{\beta}) = |\boldsymbol{\beta}|$ is not differentiable in $\boldsymbol{\beta} = 0$. The optimality condition in this case is that the subgradient of function $f(\boldsymbol{\beta})$ includes the point $\boldsymbol{\beta} = 0$. Therefore, the optimal solution for (35) must satisfies

$$-\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\nu = 0 \tag{36}$$

where ν_j is the jth component of the subgradient of $||\boldsymbol{\beta}||_1$, such that

$$\nu_j \in \begin{cases} \{+1\} & \text{if } \beta_j > 0 \\ \{-1\} & \text{if } \beta_j < 0 \\ [-1,1] & \text{if } \beta_j = 0 \end{cases} \tag{37}$$

We can write (36) as:

$$\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \lambda\nu \tag{38}$$

Setting $\boldsymbol{\beta} = 0$, by the Karush-Kuhn-Tucker (KKT) conditions, we have

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} \in [-\lambda, \lambda] \tag{39}$$

So, for $\boldsymbol{\beta} = 0$ the optimal condition for (35) is

$$-\lambda \leq \mathbf{X}^{\mathrm{T}}\mathbf{y} \leq \lambda \tag{40}$$

or

$$\lambda \geq \max|\mathbf{X}^{\mathrm{T}}\mathbf{y}| \tag{41}$$

Let $\lambda_{\max}$ be the smallest tuning parameter value for which all coefficients in the solution are zero ($\boldsymbol{\beta} = 0$), for the original LASSO in (35) we have

$$\lambda_{\max} = \max|\mathbf{X}^{\mathrm{T}}\mathbf{y}| \tag{42}$$

This is the maximum value in the sequence of λ's used in LASSO and adaLASSO estimation. Section 2.3.2 discuss the issue of selecting the parameter λ.

Analogously to (36), if we derive the LASSO-CVaR in (34) and define $\varepsilon_t = y_t - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_t$, for $t = T_{in} + 1, \dots, T_{out}$, we have

$$-\mathbf{X}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda\nu+\frac{\gamma}{T_{out}(1-\alpha)}\sum_{t\in\{\mathrm{L}(\varepsilon_t)\}}\eta_t=0 \tag{43}$$

where ν_j is defined by (37), $\{\mathrm{L}(\varepsilon_t)\}$ is the set of the $(1-\alpha)\%$ largest ε_t and η_t is the subderivate of ε_t defined in (44).

$$\eta_t\in\begin{cases}\{-\boldsymbol{x}_t\} & \text{if } \varepsilon_t>0\\ \{\boldsymbol{x}_t\} & \text{if } \varepsilon_t<0\\ [-\boldsymbol{x}_t,\boldsymbol{x}_t] & \text{if } \varepsilon_t=0\end{cases} \tag{44}$$

When $\boldsymbol{\beta}=0$ we have $\varepsilon_t=y_t$. To simplify, we assume that $y_t\neq 0$, so from (43), (37) and (44), we have

$$-\lambda\leq\mathbf{X}^{\mathrm{T}}\mathbf{y}+\frac{\gamma}{T_{out}(1-\alpha)}\sum_{t\in\{\mathrm{L}(y_t)\}}\boldsymbol{x}_t\mathrm{sgn}(y_t)\leq\lambda \tag{45}$$

where $\mathrm{sgn}(y_t)$ is the sign function of y_t and $t=T_{in}+1,\dots,T_{out}$.

Let $\overline{\boldsymbol{x}}_{\mathbf{out}}$ be the average of $(\boldsymbol{x}_t\mathrm{sgn}(y_t))$ with $t\in\{\mathrm{L}(y_t)\}$, where $\{\mathrm{L}(y_t)\}$ is the set of the $(1-\alpha)\%$ largest y_t, as presented in (46):

$$\overline{\boldsymbol{x}}_{\mathbf{out}}=\frac{1}{T_{out}(1-\alpha)}\sum_{t\in\{\mathrm{L}(y_t)\}}\boldsymbol{x}_t\mathrm{sgn}(y_t) \tag{46}$$

it follows that

$$\lambda\geq\max|\mathbf{X}^{\mathrm{T}}\mathbf{y}+\gamma\overline{\boldsymbol{x}}_{\mathbf{out}}| \tag{47}$$

Finally, $\lambda_{\max}$ of the LASSO-CVaR is given by eq. (48)

$$\lambda_{\max}^{LASSO-CVaR}=\max|\mathbf{X}^{\mathrm{T}}\mathbf{y}+\gamma\overline{\boldsymbol{x}}_{\mathbf{out}}| \tag{48}$$

4.3. Simulation

The goal of this simulation exercise is to test the robustness of LASSO-CVaR proposed in this chapter. Therefore, we need to generate a data set with outliers.

Consider the following data generating process (DGP1):

$$y_t = \sum_{k=1}^{q} \beta_k x_{k,t} + \left(\beta_{q+1} x_{q+1,t} * \mathrm{I}_{F_{in}}(t)\right) + 0.5\,\varepsilon_t, \qquad \varepsilon_t \sim \mathrm{IN}[0,1],$$

$$\mathrm{I}_{F_{in}}(t) = \begin{cases} 1, & \text{if } t \in F_{in} \\ 0, & \text{if } t \notin F_{in} \end{cases} \tag{49}$$

$$\boldsymbol{x}_t = \boldsymbol{v}_t, \qquad \boldsymbol{v}_t \sim \mathrm{IN}_q\left[0, \boldsymbol{I}_q\right] \quad \text{for } t = 1, \dots, T$$

where $\boldsymbol{\beta}$ is a vector of ones of size q; $\beta_{q+1} = 5$; $\boldsymbol{x}_t$ is a vector of q relevant variables; and F_{in} is a set of 5% of the T_{in} observations. Observations in F_{in}, chosen randomly, suffer the effect of β_{q+1}, called “fake coefficient”.

The term “fake coefficient” is used to give the idea of a coefficient that is not actually relevant in the true model, but still affects the response in some observations. The variable $\mathbf{x}_{q+1}$ is irrelevant in 95% of the “in-sample” set, i.e., $\beta_{q+1} = 5$ in 5% of the “in-sample” observations, and $\beta_{q+1} = 0$ for the rest of the sample. In other words, we are “contaminating” the “in-sample” data with outliers.

The value for β_{q+1} was fixed in order to produce “big outliers” and increase the variance of the in-sample y_t. There was no clear base for the choice of this value, and we could have chosen differently, or test model selection methods for different values of β_{q+1}, likewise the choice of the number of “fake coefficients”. These are issues for future study and simulations.

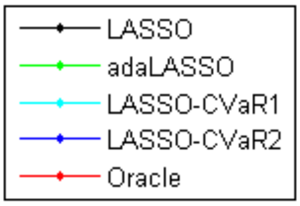
In order to test variable selection and out-of-sample performance of the approach proposed in this chapter, we compare the LASSO-CVaR to the original LASSO and adaLASSO for linear regressions as implemented in Chapter 2. We also compare the methods to the oracle approach that provides a best-case scenario by assuming the true model was known. The true model does not include the “fake coefficient”.

Similarly to simulation in Chapter 2, the comparison takes into account variable selection statistics, properties of estimators and forecast accuracy. We also compare statistics related to the “fake coefficient” and its selection rate.

LASSO-CVaR solves the quadratic optimization problem in (34) using interior point methods. Due to computational time limitations we used the regularization parameter λ chosen by LASSO, using BIC criterion. In the future, we shall discuss how to build a sequence of λ’s from the $\lambda_{\max}^{LASSO-CVaR}$ in (48) and selection criteria.

In this simulation exercise we tested the LASSO-CVaR with $\gamma = 0.25$ and $\gamma = 0.5$. The confidence level α of the CVaR was set at 0.75, and the “in-sample” (T_{in}) and “out-of-sample” (T_{out}) sets are 80% and 20% of the total T observations. We consider a total of 100 out-of-sample observations. For instance, if T = 300, we will have $T_{in} = 240$, $T_{out} = 60$, and 100 more out-of-sample (real out-of-sample) observations. As in Section 2.4, we simulate T = 50, 100, 300, 500 observations of DGP1 (49) for different combinations of candidate (n) and relevant (q) variables. We consider n = 100, 300 and q = 5, 10, 15, 20.

Figures 15-18 illustrate the distribution of the bias for the Oracle, LASSO, adaLASSO, LASSO-CVaR1 (with $\gamma = 0.25$) and LASSO-CVaR2 (with $\gamma = 0.5$) estimators for the parameter β_1, chosen arbitrarily, over 1000 Monte Carlo replications, for different sample sizes, number of candidate variables and number of relevant variables. Figures 19-22 illustrate the distribution of the bias for the “fake coefficient” estimates. In this case it does not make sense to talk about oracle estimator. Color lines of each model are shown in the color legend:



Color Legend for Figures 15-22

As in Section 2.4, from the plots in Figures 15-18, we notice that bias and variance decrease when T increases. As expected, the adaLASSO estimator is the closest to the Oracle, but the LASSO-CVaR estimator presented smaller bias and variance than the LASSO, especially when T increases. LASSO-CVaR2 ($\gamma = 0.5$) presents smaller bias than LASSO-CVaR1 ($\gamma = 0.25$), which is logical as the $\gamma = 0.5$ forces the CVaR term to be smaller than $\gamma = 0.25$, which should reduces de bias and variance of the estimators.

When comparing distributions for the “fake parameter” in Figures 19-22, we notice that when T = 300 and T = 500, the LASSO-CVaR presents much better results than the others. For T = 50 and T = 100, all models present similar bias, but LASSO-CVaR presents smaller variance.

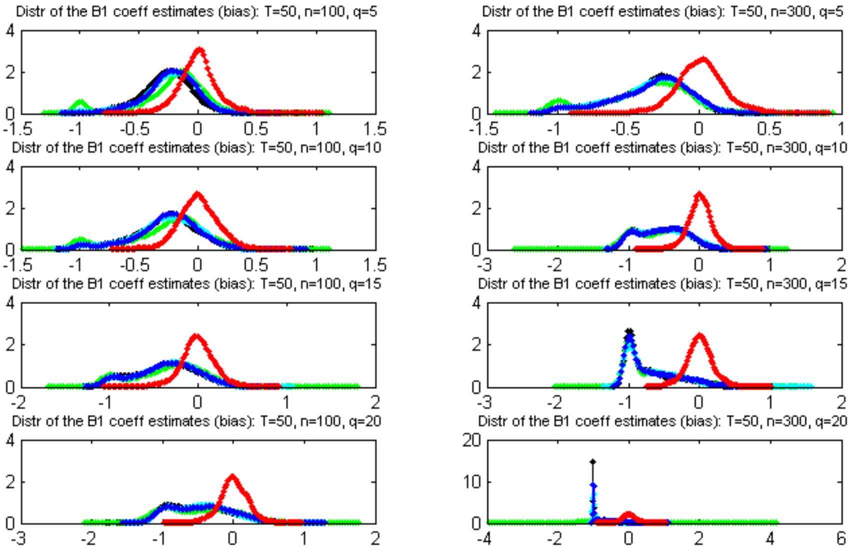


FIGURE 15. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 50 observations.

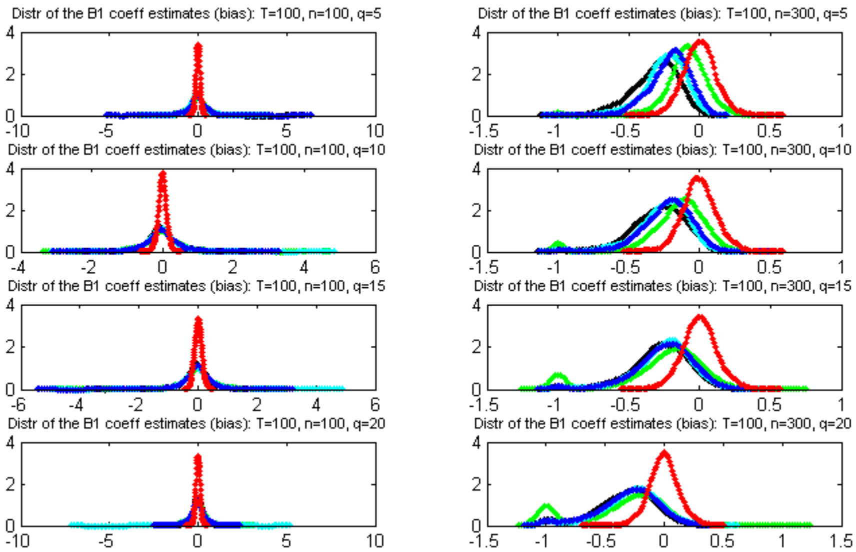


FIGURE 16. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 100 observations.

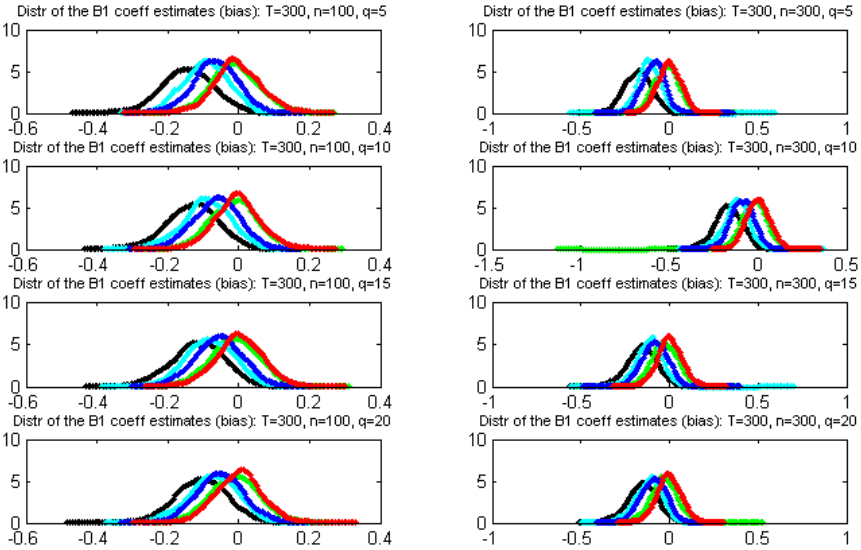


FIGURE 17. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 300 observations.

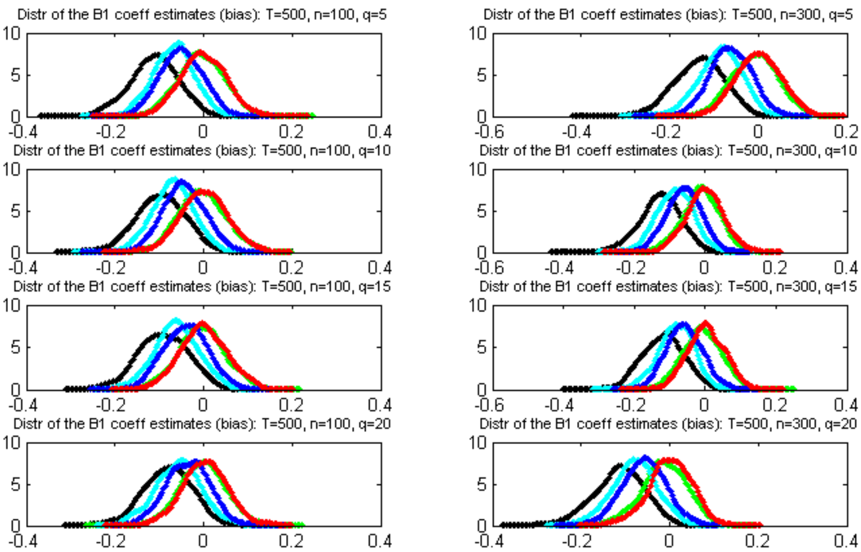


FIGURE 18. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 500 observations.

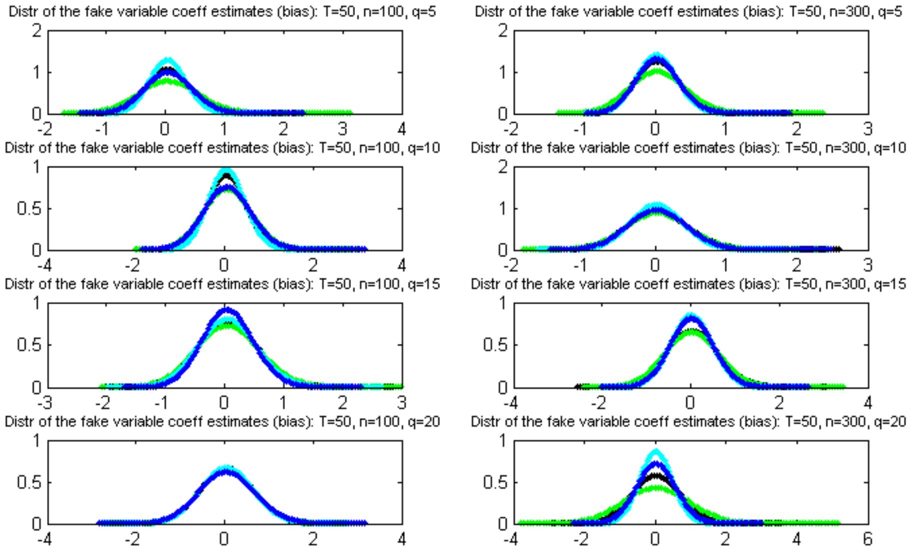


FIGURE 19. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the "fake parameter" over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 50 observations.

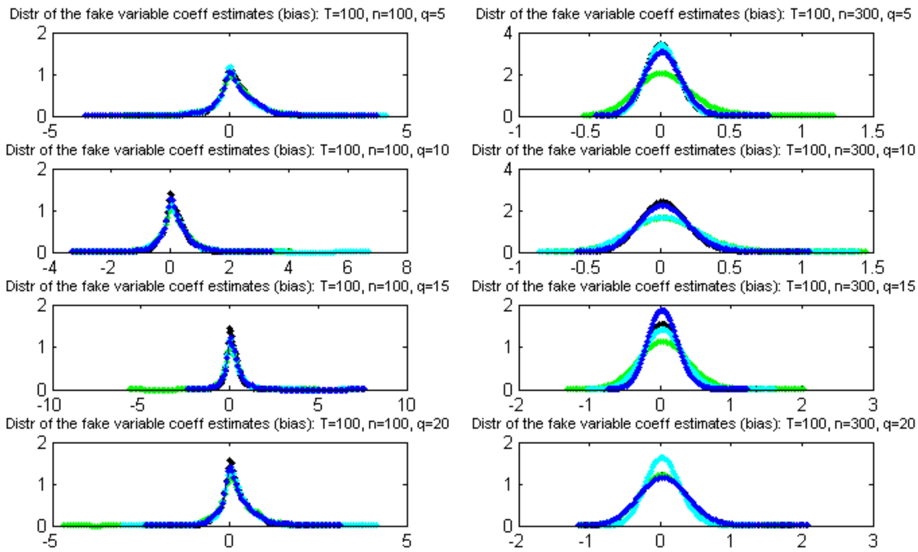


FIGURE 20. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the "fake parameter" over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 100 observations.

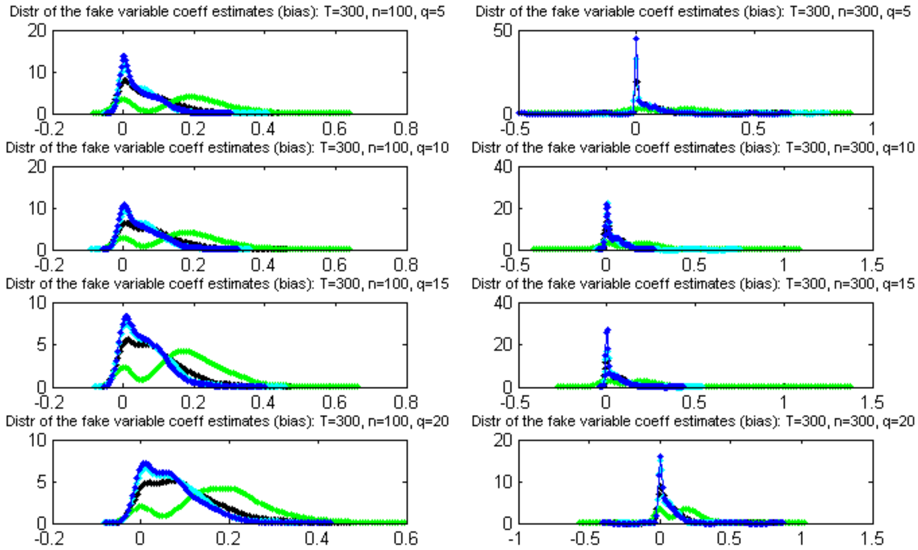


FIGURE 21. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the “fake parameter” over 1000 Monte Carlo replications. Different combinations of candidate (*n*) and relevant (*q*) variables. The sample size equals 300 observations.

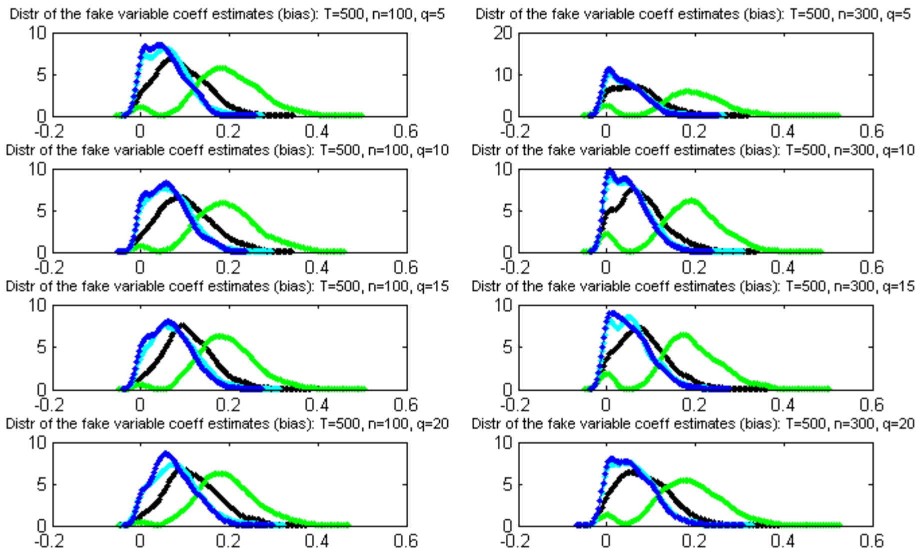


FIGURE 22. Distribution of the bias for the Oracle (red), LASSO (black), adaLASSO (green), LASSO-CVaR1 (cyan) and LASSO-CVaR2 (blue) estimators for the “fake parameter” over 1000 Monte Carlo replications. Different combinations of candidate (*n*) and relevant (*q*) variables. The sample size equals 500 observations.

Tables 45-48 present variable selection results for LASSO, adaLASSO, LASSO-CVaR1 and LASSO-CVaR2, following the format and statistics of Tables 2-5 in Section 2.4.1. Comparing Tables 45 and 46 to Tables 4 and 5, we notice that, in some scenarios, with the presence of outliers, LASSO and adaLASSO included more irrelevant variables (Panel (f)). We attribute this negative result to the “fake coefficient”. The LASSO and adaLASSO methods will select irrelevant variables, including the “fake variable”, trying to explain the effect caused by the “fake coefficient”. We believed the LASSO-CVaR would be able to exclude more irrelevant variables, however Table 47 and 48 show that LASSO-CVaR present worst results than LASSO and adaLASSO concerning variable selection. We notice that LASSO-CVaR1 includes less irrelevant variables (Panel (f)) than LASSO-CVaR2. This can be explained by the increase of parameter γ that forces LASSO-CVaR to decrease the CVaR of the “out-of-sample” errors. In order to do so, the method selects more variables. A possible way of improving this result is reducing γ. This is an important issue for future research.

Table 49 presents the selection rates for the “fake coefficient”. Figures 19-22 show that the LASSO-CVaR minimizes the bias and variance of the “fake coefficient”, however, Table 49 shows the rate of its inclusion is significantly high, even if lower than for LASSO and adaLASSO when T = 500. This combined analysis gives an indication that, differently from LASSO and adaLASSO, even when LASSO-CVaR selects the “fake coefficient”, the method’s estimative for this coefficient is close to zero, like in the true model.

TABLE 45. MODEL SELECTION: DESCRIPTIVE STATISTICS
LASSO

The table reports for each different sample size, several statistics concerning model selection. Panel (a) presents the fraction of replications where the correct model has been selected. Panel (b) shows the fraction of replications where the relevant variables are all included. Panel (c) presents the fraction of relevant variables included. Panel (d) shows the fraction of irrelevant variables excluded. Panel (e) presents the average number of included variables. Panel (f) shows the average number of included irrelevant regressors.

LASSO								
	T=50		*T*=100		*T*=300		*T*=500	
*q**n*	5	10	5	10	5	10	5	10
	Panel (a): Correct Sparsity Pattern							
5	0.011	0.002	0.012	0.064	0.076	0.101	0.027	0.039
10	0	0	0	0.001	0.013	0.015	0.004	0.009
15	0	0	0	0	0.003	0.007	0.001	0.003
20	0	0	0	0	0.001	0	0.001	0.001
	Panel (b): True Model Included							
5	0.948	0.871	0.983	0.995	1	1	1	1
10	0.816	0.295	0.979	0.967	1	1	1	1
15	0.45	0.01	0.952	0.943	1	1	1	1
20	0.064	0	0.943	0.834	1	1	1	1
	Panel (c): Fraction of Relevant Variables Included							
5	0.985	0.966	0.997	0.998	1	1	1	1
10	0.965	0.835	0.998	0.991	1	1	1	1
15	0.914	0.639	0.997	0.983	1	1	1	1
20	0.813	0.508	0.997	0.965	1	1	1	1
	Panel (d): Fraction of Irrelevant Variables Excluded							
5	0.841	0.899	0.123	0.982	0.967	0.985	0.968	0.989
10	0.773	0.888	0.109	0.936	0.938	0.976	0.943	0.980
15	0.749	0.883	0.081	0.881	0.905	0.964	0.917	0.971
20	0.735	0.881	0.085	0.850	0.869	0.949	0.890	0.962
	Panel (e): Number of Included Variables							
5	20.072	34.612	88.319	10.365	8.128	9.489	7.998	8.128
10	30.113	40.751	90.174	28.336	15.591	16.868	15.168	15.718
15	35.041	42.785	93.083	48.702	23.06	25.344	22.083	23.126
20	37.439	43.499	93.116	61.289	30.494	34.248	28.839	30.681
	Panel (f): Number of Included Irrelevant Variables							
5	15.149	29.783	83.336	5.376	3.128	4.489	2.998	3.128
10	20.466	32.405	80.195	18.422	5.591	6.868	5.168	5.718
15	21.333	33.204	78.131	33.963	8.060	10.344	7.083	8.126
20	21.170	33.344	73.181	41.990	10.494	14.248	8.839	10.681

TABLE 46. MODEL SELECTION: DESCRIPTIVE STATISTICS
adaLASSO

The table reports for each different sample size, several statistics concerning model selection. Panel (a) presents the fraction of replications where the correct model has been selected. Panel (b) shows the fraction of replications where the relevant variables are all included. Panel (c) presents the fraction of relevant variables included. Panel (d) shows the fraction of irrelevant variables excluded. Panel (e) presents the average number of included variables. Panel (f) shows the average number of included irrelevant regressors.

adaLASSO								
	T=50		T=100		T=300		T=500	
$q\backslash n$	5	10	5	10	5	10	5	10
			Panel (a): Correct Sparsity Pattern					
5	0.003	0	0.004	0.016	0.023	0.05	0.006	0.013
10	0	0	0	0	0.002	0.002	0.001	0.003
15	0	0	0	0	0	0.001	0.002	0.001
20	0	0	0	0	0	0	0	0
			Panel (b): True Model Included					
5	0.878	0.817	0.957	0.978	1	1	1	1
10	0.794	0.282	0.937	0.923	1	1	1	1
15	0.444	0.012	0.917	0.848	1	1	1	1
20	0.069	0.001	0.915	0.718	1	1	1	1
			Panel (c): Fraction of Relevant Variables Included					
5	0.921	0.911	0.991	0.990	1	1	1	1
10	0.922	0.782	0.992	0.961	1	1	1	1
15	0.863	0.602	0.994	0.902	1	1	1	1
20	0.742	0.483	0.994	0.826	1	1	1	1
			Panel (d): Fraction of Irrelevant Variables Excluded					
5	0.838	0.895	0.105	0.977	0.952	0.967	0.953	0.984
10	0.759	0.881	0.088	0.924	0.912	0.960	0.920	0.972
15	0.743	0.882	0.073	0.873	0.870	0.948	0.888	0.961
20	0.747	0.881	0.080	0.858	0.825	0.922	0.853	0.947
			Panel (e): Number of Included Variables					
5	20.023	35.596	89.952	11.671	9.561	14.696	9.452	9.677
10	30.892	42.268	91.971	31.616	17.888	21.727	17.164	18.055
15	34.825	42.779	93.714	49.612	26.068	29.798	24.515	26.218
20	35.088	43.109	93.492	56.207	33.961	41.748	31.721	34.734
			Panel (f): Number of Included Irrelevant Variables					
5	15.417	31.042	84.997	6.721	4.561	9.696	4.452	4.677
10	21.674	34.451	82.047	22.003	7.888	11.727	7.164	8.055
15	21.873	33.752	78.809	36.086	11.068	14.798	9.515	11.218
20	20.257	33.448	73.603	39.683	13.961	21.748	11.721	14.734

TABLE 47. MODEL SELECTION: DESCRIPTIVE STATISTICS
LASSO-CVaR1

The table reports for each different sample size, several statistics concerning model selection. Panel (a) presents the fraction of replications where the correct model has been selected. Panel (b) shows the fraction of replications where the relevant variables are all included. Panel (c) presents the fraction of relevant variables included. Panel (d) shows the fraction of irrelevant variables excluded. Panel (e) presents the average number of included variables. Panel (f) shows the average number of included irrelevant regressors.

LASSO-CVaR1								
	T=50		*T*=100		*T*=300		*T*=500	
*q**n*	5	10	5	10	5	10	5	10
				Panel (a): Correct Sparsity Pattern				
5	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
				Panel (b): True Model Included				
5	0.958	0.894	0.975	0.988	1	1	1	1
10	0.848	0.336	0.95	0.977	1	1	1	1
15	0.481	0.013	0.934	0.952	1	1	1	1
20	0.097	0	0.921	0.863	1	1	1	1
				Panel (c): Fraction of Relevant Variables Included				
5	0.987	0.972	0.995	0.996	1	1	1	1
10	0.971	0.850	0.995	0.992	1	1	1	1
15	0.919	0.656	0.995	0.986	1	1	1	1
20	0.835	0.538	0.996	0.970	1	1	1	1
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.761	0.881	0.125	0.942	0.875	0.931	0.875	0.942
10	0.715	0.878	0.097	0.891	0.826	0.919	0.835	0.923
15	0.706	0.875	0.075	0.838	0.778	0.901	0.791	0.908
20	0.701	0.874	0.069	0.816	0.736	0.885	0.754	0.891
				Panel (e): Number of Included Variables				
5	27.596	39.840	88.070	22.161	16.905	25.278	16.905	22.041
10	35.368	43.865	91.218	41.495	25.692	33.440	24.835	32.345
15	38.770	45.386	93.587	60.961	33.845	43.165	32.745	41.268
20	40.632	45.996	94.403	70.963	41.099	52.280	39.690	50.480
				Panel (f): Number of Included Irrelevant Variables				
5	22.663	34.978	83.095	17.181	11.905	20.278	11.905	17.041
10	25.657	35.362	81.268	31.577	15.692	23.44	14.835	22.345
15	24.983	35.547	78.655	46.170	18.845	28.165	17.745	26.268
20	23.940	35.244	74.485	51.555	21.099	32.280	19.690	30.480

TABLE 48. MODEL SELECTION: DESCRIPTIVE STATISTICS
LASSO-CVaR2

The table reports for each different sample size, several statistics concerning model selection. Panel (a) presents the fraction of replications where the correct model has been selected. Panel (b) shows the fraction of replications where the relevant variables are all included. Panel (c) presents the fraction of relevant variables included. Panel (d) shows the fraction of irrelevant variables excluded. Panel (e) presents the average number of included variables. Panel (f) shows the average number of included irrelevant regressors.

LASSO-CVaR2								
	T=50		T=100		T=300		T=500	
$q\backslash n$	5	10	5	10	5	10	5	10
				Panel (a): Correct Sparsity Pattern				
5	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
				Panel (b): True Model Included				
5	0.972	0.881	0.978	0.998	1	1	1	1
10	0.852	0.332	0.963	0.979	1	1	1	1
15	0.517	0.012	0.936	0.957	1	1	1	1
20	0.110	0	0.915	0.856	1	1	1	1
				Panel (c): Fraction of Relevant Variables Included				
5	0.993	0.970	0.996	0.999	1	1	1	1
10	0.974	0.846	0.996	0.996	1	1	1	1
15	0.932	0.655	0.995	0.985	1	1	1	1
20	0.843	0.523	0.995	0.970	1	1	1	1
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.747	0.882	0.103	0.900	0.715	0.841	0.704	0.818
10	0.711	0.878	0.092	0.867	0.663	0.819	0.648	0.793
15	0.702	0.875	0.069	0.831	0.613	0.799	0.607	0.775
20	0.697	0.873	0.074	0.814	0.568	0.779	0.571	0.758
				Panel (e): Number of Included Variables				
5	29.036	39.644	90.229	34.455	32.110	51.785	33.159	58.814
10	35.716	43.783	91.727	48.488	40.373	62.420	41.689	70.135
15	39.289	45.517	94.033	63.063	47.859	72.427	48.389	79.246
20	41.074	45.902	93.970	71.470	54.571	81.752	54.308	87.729
				Panel (f): Number of Included Irrelevant Variables				
5	24.073	34.793	85.251	29.459	27.110	46.785	28.159	53.814
10	25.981	35.326	81.764	38.532	30.373	52.42	31.689	60.135
15	25.315	35.689	79.102	48.285	32.859	57.427	33.389	64.246
20	24.207	35.446	74.063	52.066	34.571	61.752	34.308	67.729

TABLE 49. SELECTION OF "FAKE COEFFICIENT": SELECTION RATE
DGP1

The table reports for each different sample size, number of candidate variables (n) and number of relevant variables (q), the selection rate for the "fake coefficient" over 1000 Monte Carlo replications.

	T=50		T=100		T=300		T=500	
q/n	5	10	5	10	5	10	5	10
				LASSO				
5	28%	18%	89%	12%	72%	55%	93%	87%
10	31%	17%	91%	21%	81%	61%	96%	90%
15	29%	17%	92%	30%	83%	67%	97%	91%
20	29%	14%	92%	25%	86%	65%	98%	91%
				adaLASSO				
5	25%	18%	90%	12%	77%	58%	95%	90%
10	30%	17%	93%	20%	84%	64%	97%	92%
15	29%	17%	93%	25%	86%	68%	98%	92%
20	28%	13%	92%	21%	89%	67%	98%	91%
				LASSO-CVaR1				
5	32%	19%	90%	20%	70%	58%	87%	81%
10	34%	16%	92%	27%	75%	62%	91%	85%
15	34%	14%	93%	32%	80%	66%	93%	85%
20	35%	15%	94%	29%	82%	66%	93%	86%
				LASSO-CVaR2				
5	37%	20%	92%	24%	66%	53%	77%	71%
10	36%	17%	92%	27%	68%	58%	82%	74%
15	35%	18%	94%	35%	72%	62%	83%	75%
20	31%	15%	93%	30%	77%	58%	86%	78%

Finally, Table 50 presents the mean squared error (MSE) for out-of-sample forecast for LASSO, adaLASSO, LASSO-CVaR1, LASSO-CVaR2 and oracle models. We consider a total of 100 out-of-sample observations. As expected, all methodologies improved their performance as the sample size increases, and the number of relevant and candidate variables decrease. LASSO and adaLASSO presented larger MSE than in Table 6, and LASSO-CVaR1 and LASSO-CVaR2 presented smaller MSE than the others (closer to the Oracle). This result can indicates that LASSO and adaLASSO are less capable than LASSO-CVaR of identifying outliers in the "in-sample" data. As expected, when γ increases (LASSO-CVaR2), the MSE out-of-sample decreases.

TABLE 50. FORECASTING: DESCRIPTIVE STATISTICS
DGP1

The table reports for each different sample size, the out-of-sample mean squared error (MSE) for each model selection technique. *n* is the number of candidate variables whereas *q* is the number of relevant regressors.

	T=50		*T*=100		*T*=300		*T*=500	
q/n	5	10	5	10	5	10	5	10
				MSE - Oracle				
5	0.391	0.401	0.320	0.319	0.272	0.271	0.262	0.262
10	0.568	0.571	0.392	0.388	0.295	0.295	0.275	0.277
15	0.840	0.801	0.478	0.479	0.317	0.319	0.291	0.288
20	1.125	1.180	0.563	0.568	0.340	0.343	0.301	0.305
				MSE - LASSO				
5	1.422	1.997	30.728	0.920	0.393	0.552	0.336	0.364
10	2.874	6.001	23.204	1.703	0.492	0.643	0.395	0.455
15	6.104	12.286	23.027	2.734	0.568	0.847	0.446	0.536
20	11.010	18.573	17.206	4.646	0.630	0.933	0.476	0.619
				MSE - adaLASSO				
5	1.925	2.815	41.243	0.722	0.396	0.911	0.349	0.367
10	3.915	7.291	37.725	1.785	0.458	0.800	0.385	0.416
15	7.399	13.778	36.909	3.950	0.510	0.776	0.417	0.452
20	13.546	19.799	33.633	7.114	0.560	0.996	0.439	0.500
				MSE - LASSO-CVaR1				
5	1.363	1.917	41.192	0.787	0.331	0.704	0.298	0.312
10	2.698	5.716	37.364	1.530	0.396	0.677	0.334	0.361
15	5.855	12.028	33.495	2.540	0.460	0.647	0.363	0.415
20	10.582	18.174	28.748	4.411	0.506	0.760	0.390	0.466
				MSE - LASSO-CVaR2				
5	1.363	1.966	45.588	0.706	0.339	0.498	0.302	0.326
10	2.717	5.802	33.811	1.449	0.395	0.498	0.331	0.370
15	5.718	12.092	32.668	2.547	0.439	0.672	0.357	0.407
20	10.433	18.384	23.955	4.429	0.484	0.669	0.375	0.453

Comparing to LASSO and adaLASSO, the LASSO-CVaR method presented vantages concerning forecasting accuracy and disadvantages in variable selection of the true model, when the data presents outliers. One may argue that outliers would hardly be only in the "in-sample" set, and there is an old discussion on how to split the "in-sample" and "out-of-sample" set. With this motivation, we evaluated a new simulation exercise, "contaminating" also the "out-of-sample" data with outliers.

Consider now another data generating process (DGP2):

$$y_t = \sum_{k=1}^{q} \beta_k x_{k,t} + \left(\beta_{q+1} x_{q+1,t} * \mathrm{I}_{F_{in}}(t)\right) + \left(\beta_{q+1} x_{q+1,t} * \mathrm{I}_{F_{out}}(t)\right) + 0.5\ \varepsilon_t$$

$$\varepsilon_t \sim \mathrm{IN}[0,1]$$

$$\mathrm{I}_{F_{in}}(t) = \begin{cases} 1, & \text{if } t \in F_{in} \\ 0, & \text{if } t \notin F_{in} \end{cases} \tag{50}$$

$$\mathrm{I}_{F_{out}}(t) = \begin{cases} 1, & \text{if } t \in F_{out} \\ 0, & \text{if } t \notin F_{out} \end{cases}$$

$$\boldsymbol{x}_t = \boldsymbol{v}_t, \qquad \boldsymbol{v}_t \sim \mathrm{IN}_q\left[0, \boldsymbol{I}_q\right] \quad \text{for } t = 1, \dots, T$$

where $\boldsymbol{\beta}$ is a vector of ones of size q, $\boldsymbol{x}_t$ is a vector of q relevant variables, the "fake coefficient" $\beta_{q+1} = 5$, F_{in} is the set of 5% of T_{in} observations and F_{out} is the set of 5% of T_{out} observations. Now, the "fake coefficient" is active also in the "out-of-sample" set.

We omitted the plots of the distribution of the bias for the parameter β_1 and the "fake coefficient" because they are very similar to Figures 15-22. Variable selection statistics also are very close to the case of DGP1, so tables are omitted as well. However, with the presence of outliers in the "out-of-sample" data, the "fake coefficient" selection rate has increased for all methods as presented in Table 51. When *T* increases the "fake coefficient" selection rate is almost 100% for all methods. This can be explained by the increasing of F_{out} that increases proportionally with *T*, increasing the number of outliers.

Table 52 reports the MSE for out-of-sample forecast. Comparing to Table 50, we notice that the MSE increased with outliers in the "out-of-sample" set, however LASSO-CVaR still present better results than LASSO and adaLASSO in forecasting accuracy for most of the scenarios.

TABLE 51. SELECTION OF "FAKE COEFFICIENT": SELECTION RATE
DGP2

The table reports for each different sample size, number of candidate variables (n) and number of relevant variables (q), the selection rate for the "fake coefficient" over 1000 Monte Carlo replications.

	T=50		T=100		T=300		T=500	
q/n	5	10	5	10	5	10	5	10
				LASSO				
5	39%	26%	93%	16%	86%	72%	99%	97%
10	40%	21%	93%	26%	89%	77%	100%	96%
15	35%	16%	94%	34%	92%	78%	100%	97%
20	35%	17%	94%	32%	93%	78%	99%	98%
				adaLASSO				
5	32%	24%	93%	14%	90%	75%	99%	98%
10	37%	19%	93%	23%	92%	80%	100%	96%
15	32%	15%	94%	26%	94%	79%	99%	97%
20	31%	16%	94%	23%	95%	79%	100%	97%
				LASSO-CVaR1				
5	42%	30%	92%	22%	82%	67%	95%	89%
10	44%	23%	93%	31%	86%	74%	97%	91%
15	44%	19%	96%	37%	88%	75%	97%	94%
20	37%	16%	94%	34%	90%	77%	98%	95%
				LASSO-CVaR2				
5	47%	29%	94%	31%	84%	73%	94%	91%
10	45%	22%	94%	36%	83%	77%	95%	92%
15	42%	18%	94%	38%	88%	79%	95%	91%
20	38%	17%	94%	34%	91%	79%	97%	94%

TABLE 52. FORECASTING: DESCRIPTIVE STATISTICS
DGP2

The table reports for each different sample size, the out-of-sample mean squared error (MSE) for each model selection technique. n is the number of candidate variables whereas q is the number of relevant regressors.

	T=50		T=100		T=300		T=500	
q/n	5	10	5	10	5	10	5	10
				MSE - Oracle				
5	0.456	0.454	0.333	0.331	0.277	0.277	0.265	0.264
10	0.694	0.718	0.422	0.422	0.302	0.304	0.281	0.280
15	1.050	1.079	0.520	0.538	0.331	0.331	0.296	0.299
20	1.460	1.503	0.645	0.641	0.360	0.362	0.313	0.311
				MSE - LASSO				
5	2.056	2.593	40.396	1.048	0.420	0.981	0.363	0.391
10	3.759	6.804	31.166	1.995	0.542	0.794	0.434	0.494
15	7.293	13.003	28.559	3.304	0.624	1.097	0.491	0.598
20	11.793	18.963	24.969	5.733	0.719	1.153	0.540	0.677
				MSE - adaLASSO				
5	2.678	3.469	51.710	0.854	0.435	1.388	0.385	0.410
10	5.279	8.042	48.695	2.203	0.510	1.279	0.427	0.456
15	9.431	14.435	44.637	5.019	0.579	1.591	0.460	0.509
20	14.458	20.247	44.142	8.884	0.645	1.293	0.493	0.550
				MSE - LASSO-CVaR1				
5	2.283	2.844	58.685	0.914	0.356	0.852	0.312	0.331
10	4.187	6.888	49.706	1.857	0.435	0.691	0.358	0.401
15	7.510	12.836	39.364	3.398	0.517	0.993	0.395	0.470
20	11.951	18.863	37.837	5.794	0.577	1.089	0.430	0.533
				MSE - LASSO-CVaR2				
5	2.409	2.732	62.583	0.960	0.365	1.011	0.320	0.344
10	4.184	6.840	46.086	2.052	0.439	0.680	0.357	0.398
15	7.464	12.935	40.923	3.503	0.506	1.010	0.392	0.462
20	11.767	18.694	33.983	5.861	0.582	0.993	0.429	0.515

4.4. Conclusion

In this chapter we introduce an extension of LASSO with a second regularization term. For this, we use the risk measure, widely used in optimization literature, CVaR (Conditional Value at Risk).

Analyzing the results in Section 4.3, we conclude that the LASSO-CVaR has presented good features when the focus is the predictive accuracy of selected models, showing better results in out-of-sample forecasting. However, our goal is the specification of the model selecting the relevant variables of the true model simultaneously. We want to be able to interpret the model and to understand the effects of the explanatory variables on the response. In that matter, LASSO-CVaR presented worst results than the original LASSO.

Nevertheless, these are the first results for LASSO-CVaR, and many details have to be adjusted. We identified a promising field in the blend of variable selection methods based in shrinkage and the risk measure CVaR. The CVaR term showed to be useful when the data set has outliers and made the LASSO-CVaR a robust method of estimation and variable selection.

However, there are several remain questions to address. In future research we will propose an adaptive version of the LASSO-CVaR (adaLASSO-CVaR). It is important to study more carefully the parameter γ, its importance and selection criteria. Likewise, it would be interesting to estimate several LASSO-CVaR models using a sequence of λ's, as in the original LASSO.