2
Comparing model selection techniques for linear regression: LASSO and *Autometrics*

2.1.
Introduction

Several strategies for automatic model selection have been proposed over the years. Two notable approaches are the expanding or specific-to-general methods and the shrinkage or general-to-specific methods. Some examples of specific-to-general methods are stepwise regression, forward selection and the more recent RETINA (Perez-Amaral et al., 2003) and QuickNet (White, 2006). In the general-to-specific (GETS) category the most important methods are based on a model selection strategy developed by the LSE school (‘LSE' approach), revised in *PcGets* (Hendry and Krolzig, 1999, and Krolzig and Hendry, 2001), and more recently in *Autometrics* (Doornik, 2009). Still among shrinkage methods, the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani (1996), and the adaptive LASSO (adaLASSO), proposed by Zou (2006), have received particular attention.

Important work has been done in the comparison between different methodologies. Perez-Amaral, Gallo and White (2005) and Castle (2005) evaluated and compared *PcGets* (general-to-specific approach) and RETINA (specific-to-general approach). The two procedures present different goals: RETINA was developed with the aim of finding a model that has good out-of-sample predictive ability whereas *PcGets* selects a congruent dominant in-sample model, aiming to locate the DGP (Data Generating Process) nested within the GUM (General Unrestricted Model). Ericsson and Kamin (2009) compared and assessed the empirical merits of *PcGets* and *Autometrics*. Castle et al. (2011) considered how to evaluate model selection approaches and compared *Autometrics* to 1-cut approach, which consists in a GETS selection for a constant model in orthogonal variables, where only one single decision is required to select the final model.

Although all the recent literature in this field, no work has been done comparing the *PcGets*, or *Autometrics* (extension of *PcGets*), with the LASSO, or adaLASSO. In this chapter, we compare these methods on linear regression models. In the simulation experiment we compare the predictive power (forecast out-of-sample) and the performance in the correct model selection and estimation (in-sample). The case where the number of candidate variables exceeds the number of observation is considered as well. The different model selection methodologies were compared varying the sample size, the number of relevant variables and the number of candidate variables. Finally we apply both methods to predict the quarterly US GDP on the period from 1959 to 2011.

Chapter 2 is organized as follows. In Section 2.2 and 2.3 we present the variable selection methodologies used in the comparison, algorithms, estimators and settings. Section 2.4 presents the Monte Carlo experiment, the simulation results and the comparison between *Autometrics* (Liberal and Conservative), LASSO and adaLASSO. Section 2.5 presents the application of the methodologies to US GDP forecasting. Finally, Section 2.6 concludes.

2.2. *PcGets* and *Autometrics*

The main pillar of this approach is the concept of GETS modeling: starting from a general dynamic statistical model which captures the main characteristics of the underlying data set, standard testing procedures are used to reduce its complexity by eliminating statistically insignificant variables, checking the validity of the reductions at every stage to ensure the congruence[1] of the selected model.

Hoover and Perez (1999) were the first to evaluate the performance of GETS modeling as a general approach to econometric model building. To analyze this approach systematically, the authors mechanized the decisions in the GETS modeling by coding them in a computer algorithm. The most basic steps that such algorithm follows are:

1. Ascertain that the general statistical model is congruent (well specified).
2. Eliminate a variable (or variables) that satisfies the selection (i.e.,

[1] A congruent model should satisfy: (1) homoscedastic, independent errors; (2) strongly exogenous conditioning variables for the parameters of interest; (3) constant, invariant parameters of interest; (4) theory-consistent, identifiable structures; (5) data admissible formulations on accurate observations. For more details see Hendry and Nielsen (2007).

simplification) criteria.

3. Check that the simplified model remains congruent.
4. Continue steps 2 and 3 until none of the remaining variables can be eliminated.

In order to eliminate the effect of order of variable elimination, i.e., the order in which the variables are eliminated, on the outcome of GETS modeling, Hoover and Perez (1999) considered many reduction paths from an initial general model. When searches lead to different model selections, encompassing tests and/or information criteria can be used to discriminate between these models.

Hendry and Krolzig (1999), and Krolzig and Hendry (2001) proposed improvements on Hoover and Perez's GETS algorithm. They develop and analyze an econometric model selection process, called *PcGets,* present in Ox Package. Using Monte Carlo simulation they studied the probabilities of *PcGets* recovering the data generating process (DGP), and they achieved good results. Campos et al. (2003) established the consistency of *PcGets* procedure. Hendry and Krolzig (2005) discussed how to produce nearly unbiased estimates despite selection.

PcGets opens up econometric analysis to non-expert users, freeing invaluable time for the user to think about the model, and interpret the evidence. The user is required to specify the general unrestricted model based in economic theory, and then let *PcGets*, and the computer, do the rest.

Doornik (2009) introduced a third-generation algorithm, called *Autometrics*, based on the same principles. The new algorithm can also be applied in the general case of more variables than observations. *Autometrics* uses a tree-path search to detect and eliminate statistically insignificant variables, thereby improving on the multi-path of *PcGets*. Such an algorithm does not become stuck in a single-path sequence, where a relevant variable is inadvertently eliminated, retaining other variables as proxies (e.g., as in stepwise regression).

Hendry and Krolzig (2005) advocated that *PcGets* and *Autometrics* can handle perfect collinearity[2]. One of the perfectly collinear variables would be initially excluded from the model, but the multi-path search allows the excluded

[2] Perfect collinearity denotes an exact linear dependence between variables; perfect orthogonality denotes no linear dependencies.

variable to be included in a different path search, with another perfectly singular variable being dropped.

2.2.1. Methodology

GETS methodology embodies an algorithm that automatically selects empirical models from the observed data. The algorithm explores all feasible reduction paths from a very general starting point, eliminating insignificant variables until only the relevant variables are retained. *PcGets* and *Autometrics* has five basic stages: The first stage concerns the formulation of the GUM; the second determines the estimation and testing of the GUM; the third is a pre-search process; the fourth is the multi-path search procedure, in the case of *PcGets*, and tree-path search, in the case of *Autometrics*; and the fifth is the selection of the final model.

We can say that *Autometrics* is an evolution of *PcGets* algorithm, as it is based on the same principles, but it can handle some problems that *PcGets* cannot: outlier detection and more candidate variables than observations. The following description sketches the main stages involved in *PcGets* algorithm: see Krolzig and Hendry (2001) for details.

1. *Formulation of the GUM*

   The first stage of the algorithm requires the user to specify the general unrestricted model (GUM) based on subject-matter theory, institutional knowledge, historical contingencies, data availability and measurement information. The GUM must have stationary regressors. In this "prior specification" the aim of the user is the inclusion of potentially relevant variables, the exclusion of irrelevant effects, and to achieve orthogonality between regressors. The larger the initial regressor set, the more likely adventitious effects will be retained, but the smaller the GUM, the more likely key variables will be omitted. Further, the less orthogonality between variables, the more "confusion" the algorithm faces. Therefore, careful prior analysis remains essential.

2. *Mis-specification tests*

After estimating the GUM appropriately (ordinary least squares), the second stage tests the model for mis-specification. There must be sufficient tests to check the GUM for congruence (which implies that the model matches the evidence in all measured aspects), but not too many to induce a large type-1 error. If a mis-specification test (diagnostic test) is rejected, its significance level is adjusted or the test is excluded from the test battery during simplifications of the GUM. *PcGets* generally tests the following null hypotheses: white-noise errors, conditionally homoscedastic errors, normally distributed errors, unconditionally homoscedastic errors, and constant parameters.

3. *Pre-search reductions*

The next stage in selecting a congruent and parsimonious model involves a pre-search simplification of the GUM. This pre-selection eliminates variables that are 'highly' irrelevant. The reductions are based on F-tests and t-tests of the variables ranked in order of their absolute t-values. The F-test tests for sequentially increasing blocks of omitted variables, using loose significance levels (larger than in the multiple reduction paths). The diagnostic tests are confirmed at every reduction stage to ensure congruence and the failure of a diagnostic test will terminate the reduction at that point. Having eliminated highly insignificant variables, we have a new GUM as the baseline for the remaining stages.

4. *Multiple reduction paths*

The algorithm then implements a multi-path search, commencing from all feasible initial deletion points. The searches repeatedly filter for relevant variables using both t-tests and block F-tests. Again, diagnostic tests are checked at every reduction stage to ensure the congruence of the final model. The path is terminated when all variables remaining are significant, or a diagnostic test fails. The resulting model is the terminal model of that path.

5. *Selection of the final model*

When all paths have been explored and all distinct terminal models have

been found, encompassing can be used to test between them, with only the surviving, usually non-nested, specifications retained. The terminal models are tested against their union to find an undominated encompassing contender; rejected models are removed, and the union of the 'surviving' terminal models becomes a new GUM for another multi-path search iteration; then this entire search process continues and the terminal models are again tested against there union. If more than one model survives the encompassing tests, the set of mutually encompassing and undominated contenders is reported, and a unique final choice is made by the pre-selected information criterion.

The algorithm in *Autometrics* shares these characteristics and stages with the algorithm in *PcGets*. However, *Autometrics* (unlike *PcGets*) uses a tree-search method, with refinements on pre-search simplification and on the objective function. The tree-search procedure follows all feasible paths. For details see Doornik (2009).

In the pre-search stage, *Autometrics* can detect outliers using impulse dummies for all observations in a process known as impulse-indicator saturation (IIS). For more details see Hendry and Krolzig (2004) and Castle et al. (2012).

Allowing for a reasonable lag-length in the GUM or IIS, the researcher can be easily faced with a situation of more candidate variables than observations. In that case, *Autometrics* applies the cross-block algorithm proposed in Hendry and Krolzig (2004), which consist in:

1. dividing the set of variables into subsets (blocks), each of which contains less than half of the observations;
2. applying *Autometrics* model selection to each combination of the blocks (GUMs). The algorithm yields a terminal model for each GUM;
3. taking the union of the terminal models derived from each GUM, forming a new single union model;
4. If the number of variables in this model is less than the number of observations, model selection proceeds from this new union model (new unique GUM), otherwise, restarts the cross-block algorithm with the new set of variables.

For a better visualization of the algorithm described above, take 3 blocks, *A*, *B* and *C*:

$$
\begin{array}{ll}
A \cup B \text{ model selection} \rightarrow & G_1 \\
A \cup B \text{ model selection} \rightarrow & G_2 \\
B \cup C \text{ model selection} \rightarrow & G_3 \\
\hline
\overline{G_1 \cup G_2 \cup G_3} \text{ model selection} \rightarrow S &
\end{array}
$$

In *Autometrics*, the user can set the maximum block size (-1: unlimited). By default, the maximum size is 128 variables.

2.2.2. Algorithm settings

There are some important choices that the modeler should make before running the model selection algorithm. In PcGets, these choices concern the following:

1. Model strategy: Liberal or Conservative. The former seeks a null rejection frequency per candidate variable in a regression of about 5%, whereas the latter is centered on 1%. Hendry and Krolzig (2003) studied the difference between these two strategies. The Liberal strategy minimizes the non-selection probabilities and the Conservative minimizes the non-deletion probabilities.
2. Pre-search variables reduction (including lags): yes or no.
3. Fixed variables: Fixed variables are forced to always be included in regression, whereas free variables may be deleted by the algorithm.
4. Mis-specification tests and their significance levels (default is 0.01). Table 1 shows the diagnostic tests used by default in *PcGets*, recommended in Hendry and Krolzig (2003).
5. Information criteria for the final selection: AIC (Akaike's Information Criterion), BIC (Bayesian Information Criterion) or HQ (Hannan-Quinn Information Criterion)[3].

[3] The information criteria are defined as follows:

$$\text{AIC} = T \times \left(\log\left(\frac{2\pi \text{SSE}}{T}\right) + 1\right) + 2 \times k$$

$$\text{BIC} = T \times \log\left(\frac{\text{SSE}}{T}\right) + k \times \log T$$

$$\text{HQ} = T \times \log\left(\frac{\text{SSE}}{T}\right) + 2 \times k \times \log\log T$$

where SSE is the sum of squared errors, T is the sample size, and k is the number of parameters of the model: see Akaike (1974), Schwarz (1978), and Hannan and Quinn (1979).

TABLE 1. MIS-SPECIFICATION TESTS

There are T observations and k regressors in the model under the null. The values T and k may differ across models and the value m may differ across statistics. By default, *PcGets* sets p=4 and computes two Chow tests at $\tau_1 = (0.5T)/T$ and $\tau_2 = (0.9T)/T$.

Test	Alternative	Statistic	Sources
AR 1-p test	p-th order residual autocorrelation	$\mathrm{F}(p, T-k-p)$	Godfrey (1978), Harvey (1981, p.173)
ARCH 1-p test	p-th order autoregressive conditional heteroscedasticity	$\mathrm{F}(p, T-k-p)$	Engle (1982), Engle, Hendry and Trumbull (1985)
Normality test	skewness and excess kurtosis	$\chi^2(2)$	Jarque and Bera (1980), Doornik and Hansen (1994)
Hetero test	Heteroscedasticity quadratic in regressors x_i^2	$\mathrm{F}(m, T-k-m-1)$	White (1980), Nicholls and Pagan (1983)
Chow (τT)	Predictive falure over a subset of $(1-\tau)T$ obs.	$\mathrm{F}((1-\tau)T, \tau T-k)$	Chow (1960, p.594-595), Hendry (1979)

In *Autometrics*, which is part of the software PcGive version 12 or later, for the model strategy options (choice 1) the user can select a "target size", which means "the proportion of irrelevant variables that survives the simplification process" (Doornik, 2008). The target size values that appear to approximate liberal and conservative strategies in *PcGets* are 5% and 1%, respectively.

For pre-search testing (choice 2), *Autometrics* allows the user to enable variables reduction and lag reduction separately. *Autometrics* calls the fifth choice tie-breaker, and, beyond the information criteria, the user can also set the algorithm to choose the model with minimal number of regressors. By default, BIC is used as tie-breaker. The third and fourth choices above are identical for *PcGets* and *Autometrics*.

Besides these five choices, *Autometrics* also allows the user to enable outlier detection through dummy saturation.

As presented in the last sections, *PcGets* and *Autometrics* are similar methodologies, where the latter is more developed in many aspects. Ericsson and Kamin (2009) showed that, in several instances, *Autometrics* dominates *PcGets* by obtaining a more parsimonious model with a better fit whereas *PcGets* never dominates *Autometrics* in that sense. For all these reasons, we have chosen to use only *Autometrics* algorithm in the comparison exercise with LASSO and adaLASSO.

2.3. LASSO and adaLASSO

Shrinkage methods have become popular in the estimation of large dimensions models. Among these methods, the Least Absolute Shrinkage and Selection Operator (LASSO), proposed by Tibshirani (1996), has received particular attention because of the ability to shrink some parameters to zero, excluding irrelevant regressors. In other words, LASSO has become a popular technique for simultaneous estimation and variable selection for linear models.

LASSO is able to handle more variables than observations and produces sparse models (Zhao and Yu, 2006, Meinshausen and Yu, 2009), which are easy to interpret. Moreover, the entire regularization path of LASSO can be computed efficiently, as shown in Efron et al. (2004), or more recently in Friedman et al. (2010).

Despite all these nice characteristics, Zhao and Yu (2006) noted that the LASSO estimator can only be consistent if the design matrix[4] satisfies a rather strong condition denoted "Irrepresentable Condition", which can be easily violated in the presence of highly correlated variables. Moreover, Zou (2006) noted that the oracle property in the sense of Fan and Li (2001)[5] does not hold for LASSO. To amend these deficiencies, Zou (2006) proposes the adaptive LASSO (adaLASSO).

2.3.1. The LASSO and adaLASSO estimators

The LASSO technique is inspired in ridge regression (Hoerl and Kennard, 1970), which is a standard technique for shrinking coefficients that imposes a ℓ_2-norm penalty on regression coefficients. However, contrarily to the latter, LASSO can set some coefficients to zero, resulting in an easily interpretable model.

Consider model estimation and variable selection in a linear regression framework. Suppose that $\mathbf{y} = (y_1, \ldots, y_T)'$ is the response vector, and $\mathbf{x}_j = (x_{j1}, \ldots, x_{jT})'$, with $j = 1, \ldots, p$, are the predictor variables, possibly containing

[4] Design matrix: matrix of values of explanatory variables.

[5] Oracle property: the method both identifies the correct subset model and the estimates of non-zero parameters have the same asymptotic distribution as the ordinary least squares (OLS) estimator in a regression including only the relevant variables.

lags of $\mathbf{y}$.

The LASSO estimator, introduced by Tibshirani (1996), is given by

$$\hat{\beta}^{LASSO} = arg\min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (1)$$

where $\|.\|$ denotes the standard ℓ^2-norm, and λ is a nonnegative regularization parameter. The second term in (1) is the so-called "ℓ_1 penalty", which is crucial for the success of the LASSO. The LASSO continuously shrinks the coefficients towards 0 as λ increases, and some coefficients are shrunk to exact 0 if λ is sufficiently large.

Zou (2006) showed the LASSO estimator does not enjoy the oracle property, and proposed a simple and effective solution, the adaptive LASSO, or adaLASSO. In LASSO the coefficients are equally penalized in the ℓ_1 penalty. In the adaLASSO each coefficient is assigned with different weights. Zou (2006) showed that if the weights are data-dependent and cleverly chosen, then the adaLASSO can have the oracle property.

The adaLASSO estimator is given by

$$\hat{\beta}^{adaLASSO} = \arg\min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j| \qquad (2)$$

where $\widehat{w}_j = 1/|\hat{\beta}_j^*|^{\gamma}, \gamma > 0$, and $\hat{\beta}_j^*$ is an initial parameter estimate. As the sample size grows, the weights diverge (to infinity) for zero coefficients, whereas, for the non-zero coefficients, the weights converge to a finite constant. Zou (2006) suggests using the ordinary least squares (OLS) estimate of the parameters as the initial parameter estimate $\hat{\beta}_j^*$. However, such estimator is not available when the number of candidate variables is larger than the number of observations. In this case, ridge regression can be used as an initial estimator. Recently, others estimators have been used as pre-estimators. Medeiros and Mendes (2013) showed that the elastic net procedure, proposed by Zou and Hastie (2005), delivers the most robust results using adaLASSO. Therefore, in this work we use the elastic net estimator as the initial parameter estimate in eq. (2).

2.3.2.
Selecting λ and γ

A critical point in the LASSO and adaLASSO literature is the selection of the regularization parameter λ and the weighting parameter γ. Traditionally, one employs cross-validation maximizing some predictive measure. In a time-dependent framework cross-validation becomes more difficult. An alternative approach that has shown good results is using information criteria, such as the Bayesian Information Criterion (BIC). Zou et al. (2007), Wang et al. (2007) and Zhang et al. (2010) study such method.

Zou et al. (2007) showed that the number of nonzero coefficients is an unbiased and consistent estimator of the degrees of freedom of the model, and proposed BIC for the LASSO. Shao (1997) indicated that in a classical linear regression, BIC perform better than cross-validation if the true model has a finite dimension and is among the candidate models. This motivated Wang et al. (2007) to compare LASSO with tuning parameters selected by cross-validation and BIC, and they showed that the LASSO with BIC selector performs better in the identification of the correct model. Finally, Zhang et al. (2010) study a more general criterion (Generalized Information Criterion) and show that the BIC is consistent in selecting the regularization parameter, i.e. enables identification of the true model consistently.

In this work, we will use the BIC as proposed in Wang et al. (2007), based in Zou et al. (2007), in the selection of both parameters λ and γ:

$$\text{BIC} = \log(\hat{\sigma}^2) + \frac{1}{T}\widehat{df}\log(T) \tag{3}$$

where $\hat{\sigma}^2 = \text{var}\,(\mathbf{y} - \hat{\mathbf{y}})$, $\hat{\mathbf{y}}$ is the prediction of $\mathbf{y}$, using the parameters estimates. $\widehat{df}$ is the number of non-zero coefficients in the estimated model, and T is the number of observations. This selection method performs remarkably well in Monte Carlo simulations presented in the next section.

2.4. Simulation

In this section we use a Monte Carlo simulation in order to compare *Autometrics*, LASSO and adaLASSO methodologies. The procedure used to solve LASSO is the *glmnet* package for Matlab, also used for ridge regression and elastic net. The *glmnet* procedure implements a coordinate descent algorithm. For more details, see Friedman et al. (2010).

Our goal is to compare the 'size' and 'power' of the model selection process, namely the probability of inclusion in the final model of variables that do not (do) enter the DGP, i.e. retention frequency of irrelevant variables, and retention frequency of relevant variables.

We also compare each estimator to the oracle estimator, which is the ordinary least squares (OLS) estimator in a regression including only the relevant variables. Finally, we compare the forecasting accuracy of the models selected by each model selection technique. The comparison tables and statistics follow Medeiros and Mendes (2013).

To illustrate our purpose we chose to use a simple statistical model with orthogonal regressors for which the compared methods have already proved to work well and have all asymptotic properties proven. The data generating process (DGP) used is a Gaussian linear regression model, where the strongly exogenous variables are Gaussian white-noise processes:

$$\begin{aligned} y_t &= \sum_{k=1}^{q} \beta_k x_{k,t} + 0.5\,\varepsilon_t, \qquad \varepsilon_t \sim \mathrm{IN}[0,1], \\ \boldsymbol{x}_t &= \boldsymbol{v}_t, \qquad \boldsymbol{v}_t \sim \mathrm{IN}_q\left[0, \boldsymbol{I}_q\right] \text{ for } t = 1, \ldots, T, \end{aligned} \tag{4}$$

where, $\boldsymbol{\beta}$ is a vector of ones of size q and $\boldsymbol{x}_t$ is a vector of q relevant variables.

The GUM is a linear regression model, which includes the intercept, the q relevant variables of the DGP (4), and n-q irrelevant variables, which are also Gaussian white-noise processes. The GUM has n candidate variables and the constant, given by (5).

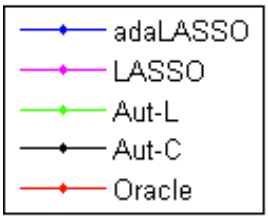
$$y_t = \pi_0 + \sum_{k_r=1}^{q} \pi_{k_r} x_{k_r,t} + \sum_{k_i=1}^{n-q} \pi_{k_i} x_{k_i,t} + u_t, \qquad u_t \sim \mathrm{IN}[0, \sigma^2] \tag{5}$$

where k_r is the index of relevant variables and k_i is the index of irrelevant variables.

We simulate T = 50, 100, 300, 500 observations of DGP (4) for different combinations of candidate (n) and relevant (q) variables. We consider n = 100, 300 and q = 5, 10, 15, 20. The models are estimated by the *Autometrics*, LASSO and adaLASSO methods. The values of the tuning parameters of the LASSO and adaLASSO, λ and γ, are selected by the BIC, as in Section 2.3.2. The parameters settings in *Autometrics* are determined by the Liberal and Conservative strategies, i.e. we compare *Autometrics* with target size of 5% (Liberal) and 1% (Conservative). The remaining *Autometrics*'s settings are defined by default, as showed in Section 2.2.2.

2.4.1. Simulation results

In this section we present the results of the simulation exercise using the different methodologies. We start by analyzing the properties of the estimators for the parameter β_1 in (4), chosen arbitrarily. Figures 1-4 illustrates the distribution of the bias for the Oracle, *Autometrics* Liberal (Aut-L), *Autometrics* Conservative (Aut-C), LASSO and adaLASSO estimators for different sample sizes, number of candidate variables and number of relevant variables, with color lines shown in the color legend:



Color Legend for Figures 1-4

From the several plots, we can say that the bias and variance can vary greatly depending on the number of observations (T) and the number of candidate variables (n). Analyzing the plots along Figures 1 to 4, we notice that in all methodologies, in a general way, the bias and variance decrease with the increasing of T. Only by looking at the distributions we notice that both *Autometrics* (Liberal and Conservative) present the smallest bias and variance and the parameter estimates distribution is very close to the distribution of the Oracle estimator. Analyzing the LASSO and adaLASSO estimators distributions, it is

evident that the latter is closer to the Oracle, but not as much as the *Autometrics* estimators. For T=500, all distributions are close to the Oracle except for the LASSO estimator, which is consistent with the theory presented earlier. For T=50 and q=5, the distribution of each estimator approaches the distribution of the Oracle in an order: the closest is the Aut-C, followed by Aut-L, adaLASSO and LASSO, in this order. Given both cases of n=100 and n=300, we are facing the case $T < n$, and even in that extreme case, the bias and variance are relatively small. However, for the other values of q, the adaLASSO and LASSO distributions present fat-tails caused mainly by some outliers in the estimation, while the *Autometrics* distributions presents a nice behavior for all values of q. When T=300 and T=500, the number of outliers reduces and the adaLASSO distribution gets closer to the Oracle, while the LASSO distribution still presents a greater bias.

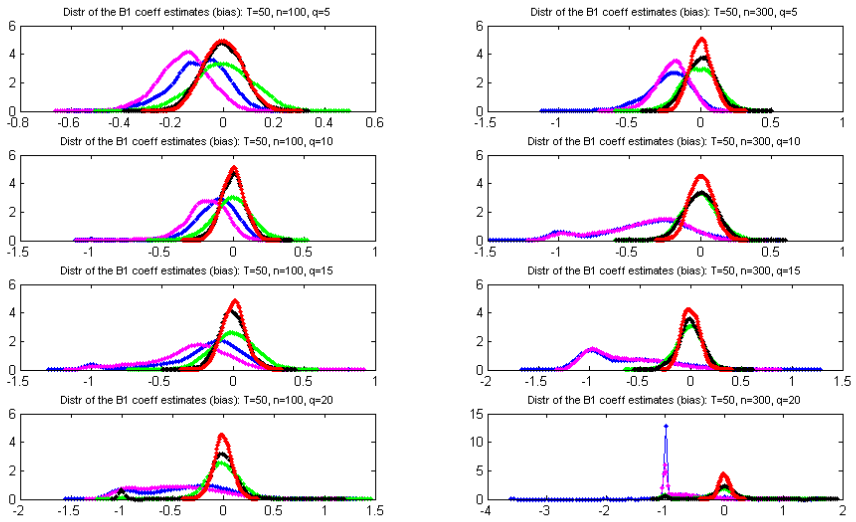


FIGURE 1. Distribution of the bias for the Oracle (red), *Autometrics* Liberal (green), *Autometrics* Conservative (black), LASSO (magenta) and adaLASSO (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 50 observations.

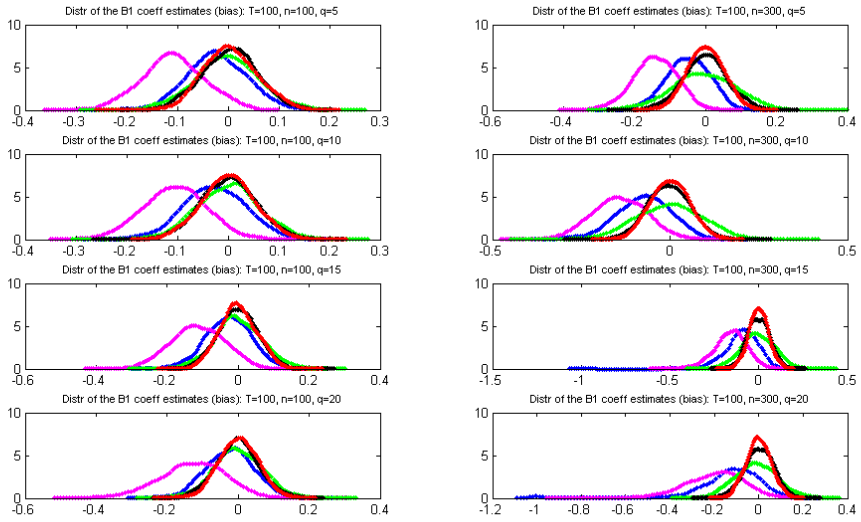


FIGURE 2. Distribution of the bias for the Oracle (red), *Autometrics* Liberal (green), *Autometrics* Conservative (black), LASSO (magenta) and adaLASSO (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 100 observations.

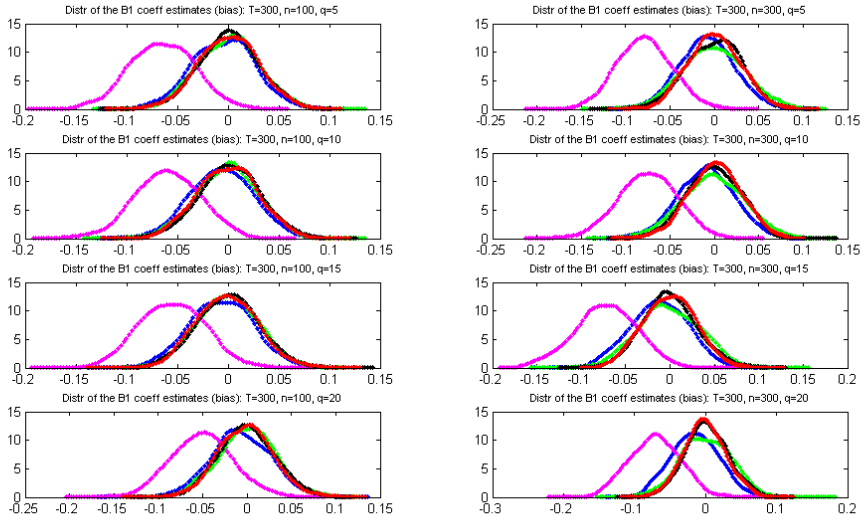


FIGURE 3. Distribution of the bias for the Oracle (red), *Autometrics* Liberal (green), *Autometrics* Conservative (black), LASSO (magenta) and adaLASSO (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 300 observations.

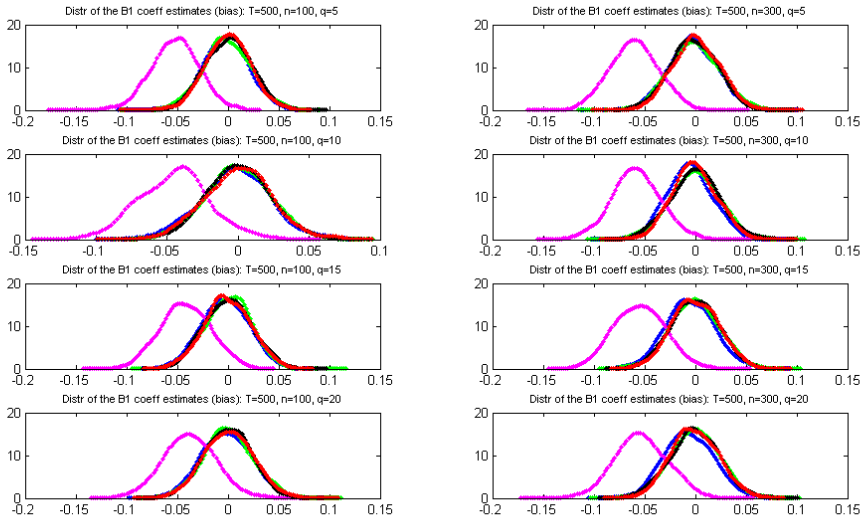


FIGURE 4. Distribution of the bias for the Oracle (red), *Autometrics* Liberal (green), *Autometrics* Conservative (black), LASSO (magenta) and adaLASSO (blue) estimators for the parameter β_1 over 1000 Monte Carlo replications. Different combinations of candidate (n) and relevant (q) variables. The sample size equals 500 observations.

For a descriptive statistics of the parameters estimates, Table 2 shows the average absolute bias and the average mean squared error (MSE) for the *Autometrics* (Liberal), *Autometrics* (Conservative), LASSO and adaLASSO estimators over the Monte Carlo simulations and the candidate variables, i.e.,

$$\text{Bias} = \frac{1}{1000n}\sum_{j=1}^{1000}\left(\sum_{i=1}^{n}\left|\hat{\beta}_i - \beta_i^{\,true}\right|\right) \tag{6}$$

$$\text{MSE} = \frac{1}{1000n}\sum_{j=1}^{1000}\left(\sum_{i=1}^{n}\left(\hat{\beta}_i - \beta_i^{\,true}\right)^2\right) \tag{7}$$

where

$$\beta_i^{\,true} = \begin{cases} 1, & \text{if } 1 \le i \le q \\ 0, & \text{if } q+1 \le i \le n \end{cases} \tag{8}$$

is the vector of size n of "true" values of the parameters of the model.

We observe that both variance (MSE) and bias are very low, especially for the *Autometrics* (Liberal and Conservative) estimators. This can be explained by the large number of zero estimates. The adaLASSO estimator presents better results than the LASSO estimator. Looking at Figures 1-4, we observe that the bias and the MSE decrease with the sample size (T) and increase with the number

of relevant variables (q).

Tables 3-6 present model selection results for each model selection technique. Panel (a) presents the fraction of replications where the correct model has been selected, i.e., all the relevant variables included and all the irrelevant regressors excluded from the final model; 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; and Panel (f) shows the average number of included irrelevant regressors.

In a general analysis the selection performance of all methodologies improves with the sample size (T) and gets worse as the number of relevant variables (q) increases. Analyzing Panel (a), we notice a difference of behavior between *Autometrics* (Liberal and Conservative), and LASSO and adaLASSO methodologies. In LASSO and adaLASSO, T and q have a big influence in correct model selection, while the first two do not show a clear influence. Panel (b) and (c) show better results for both *Autometrics* when T=50. For T=100, T=300 and T=500, the true model is included almost every time and almost all relevant variables are included in the selected model. Analyzing Panel (d), it is clear that the fraction of excluded irrelevant variables is extremely high for all scenarios and methodologies. The number of included variables and, consequentially, the number of included irrelevant variables, increase with the number of candidate variables (n) and decrease with the sample size, as shown in Panel (e) and (f). The *Autometrics* (Conservative) is the methodology that includes fewer variables in the selected model.

TABLE 2. PARAMETER ESTIMATES: DESCRIPTIVE STATISTICS

The table reports for each different sample size, the average absolute bias and the average mean squared error (MSE), for each model selection technique, over all parameter estimates and Monte Carlo simulations. n is the number of candidate variables whereas q is the number of relevant regressors.

	T=50		T=100		T=300		T=500	
$q\backslash n$	100	300	100	300	100	300	100	300
				BIAS - Autometrics (Liberal)				
5	0.028	0.010	0.012	0.017	0.005	0.006	0.004	0.003
10	0.034	0.011	0.014	0.018	0.006	0.006	0.004	0.003
15	0.040	0.012	0.016	0.018	0.007	0.007	0.005	0.003
20	0.053	0.034	0.019	0.019	0.008	0.007	0.006	0.004
				MSE - Autometrics (Liberal)				
5	0.005	0.001	0.001	0.002	0.000	0.000	0.000	0.000
10	0.006	0.002	0.002	0.002	0.000	0.000	0.000	0.000
15	0.008	0.002	0.002	0.002	0.000	0.000	0.000	0.000
20	0.016	0.019	0.002	0.002	0.000	0.000	0.000	0.000
				BIAS - Autometrics (Conservative)				
5	0.006	0.006	0.004	0.004	0.002	0.001	0.002	0.001
10	0.010	0.008	0.006	0.003	0.003	0.002	0.003	0.001
15	0.015	0.010	0.008	0.004	0.005	0.002	0.003	0.002
20	0.039	0.023	0.011	0.005	0.006	0.002	0.004	0.002
				MSE - Autometrics (Conservative)				
5	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.000
10	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
15	0.002	0.002	0.001	0.000	0.000	0.000	0.000	0.000
20	0.021	0.012	0.001	0.001	0.000	0.000	0.000	0.000
				BIAS - LASSO				
5	0.013	0.006	0.007	0.003	0.004	0.001	0.003	0.001
10	0.029	0.024	0.013	0.006	0.007	0.003	0.005	0.002
15	0.072	0.055	0.021	0.011	0.009	0.004	0.007	0.003
20	0.146	0.079	0.032	0.020	0.012	0.005	0.009	0.004
				MSE - LASSO				
5	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000
10	0.007	0.010	0.002	0.001	0.000	0.000	0.000	0.000
15	0.031	0.035	0.003	0.002	0.001	0.000	0.000	0.000
20	0.090	0.058	0.006	0.005	0.001	0.000	0.000	0.000
				BIAS - adaLASSO				
5	0.020	0.012	0.009	0.004	0.004	0.001	0.003	0.001
10	0.035	0.032	0.014	0.008	0.006	0.003	0.005	0.002
15	0.071	0.061	0.018	0.013	0.009	0.004	0.007	0.003
20	0.154	0.084	0.021	0.021	0.012	0.005	0.008	0.003
				MSE - adaLASSO				
5	0.003	0.002	0.001	0.000	0.000	0.000	0.000	0.000
10	0.007	0.014	0.001	0.001	0.000	0.000	0.000	0.000
15	0.028	0.039	0.002	0.002	0.000	0.000	0.000	0.000
20	0.096	0.063	0.002	0.005	0.001	0.000	0.000	0.000

TABLE 3. MODEL SELECTION: DESCRIPTIVE STATISTICS
Autometrics (Liberal)

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.

Autometrics (Liberal)								
	T=50		*T*=100		*T*=300		*T*=500	
*q**n*	100	300	100	300	100	300	100	300
				Panel (a): Correct Sparsity Pattern				
5	0.013	0	0.02	0	0.006	0	0.004	0
10	0.006	0	0.025	0.001	0.011	0	0.005	0
15	0.003	0	0.031	0	0.006	0.001	0.012	0
20	0.004	0	0.04	0	0.011	0	0.01	0
				Panel (b): True Model Included				
5	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1
20	0.952	0.628	1	1	1	1	1	1
				Panel (c): Fraction of Relevant Variables Included				
5	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1
20	0.9884	0.8834	1	1	1	1	1	1
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.820	0.883	0.914	0.768	0.950	0.911	0.949	0.959
10	0.816	0.898	0.924	0.772	0.948	0.914	0.947	0.958
15	0.812	0.913	0.923	0.781	0.946	0.920	0.947	0.959
20	0.823	0.923	0.921	0.792	0.947	0.926	0.945	0.957
				Panel (e): Number of Included Variables				
5	22.086	39.506	13.173	73.548	9.776	31.251	9.871	17.108
10	26.563	39.527	16.813	76.056	14.687	35.008	14.735	22.162
15	31.009	39.678	21.561	77.443	19.569	37.941	19.494	26.6
20	33.962	39.364	26.321	78.275	24.268	40.854	24.384	31.97
				Panel (f): Number of Included Irrelevant Variables				
5	17.086	34.506	8.173	68.548	4.776	26.251	4.871	12.108
10	16.563	29.527	6.813	66.056	4.687	25.008	4.735	12.162
15	16.009	24.678	6.561	62.443	4.569	22.941	4.494	11.6
20	14.194	21.696	6.321	58.275	4.268	20.854	4.384	11.97

TABLE 4. MODEL SELECTION: DESCRIPTIVE STATISTICS
Autometrics (Conservative)

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.

Autometrics (Conservative)								
	T=50		*T*=100		*T*=300		*T*=500	
q\n	100	300	100	300	100	300	100	300
				Panel (a): Correct Sparsity Pattern				
5	0.442	0.109	0.459	0.105	0.354	0.214	0.389	0.083
10	0.387	0.065	0.547	0.184	0.391	0.207	0.384	0.088
15	0.384	0.036	0.52	0.175	0.387	0.167	0.423	0.085
20	0.318	0.033	0.519	0.150	0.39	0.181	0.428	0.074
				Panel (b): True Model Included				
5	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1
20	0.834	0.770	1	1	1	1	1	1
				Panel (c): Fraction of Relevant Variables Included				
5	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1
20	0.9337	0.922	1	1	1	1	1	1
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.987	0.969	0.987	0.976	0.987	0.991	0.989	0.989
10	0.984	0.959	0.990	0.988	0.988	0.990	0.988	0.990
15	0.983	0.953	0.990	0.987	0.987	0.990	0.989	0.989
20	0.981	0.954	0.989	0.986	0.987	0.989	0.989	0.989
				Panel (e): Number of Included Variables				
5	6.273	14.089	6.195	12.215	6.243	7.692	6.087	8.171
10	11.401	22.024	10.859	13.509	11.112	12.778	11.074	13.030
15	16.446	28.253	15.85	18.673	16.092	17.976	15.951	18.165
20	20.197	31.358	20.862	23.985	21.036	22.958	20.862	23.149
				Panel (f): Number of Included Irrelevant Variables				
5	1.273	9.089	1.195	7.215	1.243	2.692	1.087	3.171
10	1.401	12.024	0.859	3.509	1.112	2.778	1.074	3.030
15	1.446	13.253	0.85	3.673	1.092	2.976	0.951	3.165
20	1.523	12.924	0.862	3.985	1.036	2.958	0.862	3.149

TABLE 5. 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*	100	300	100	300	100	300	100	300
				Panel (a): Correct Sparsity Pattern				
5	0.013	0.004	0.098	0.061	0.195	0.156	0.274	0.194
10	0	0	0.007	0.002	0.043	0.027	0.061	0.043
15	0	0	0	0	0.009	0.005	0.015	0.016
20	0	0	0	0	0.004	0	0.007	0.001
				Panel (b): True Model Included				
5	1	0.999	1	1	1	1	1	1
10	0.994	0.666	1	1	1	1	1	1
15	0.751	0.029	1	1	1	1	1	1
20	0.141	0	1	0.99	1	1	1	1
				Panel (c): Fraction of Relevant Variables Included				
5	1	1	1	1	1	1	1	1
10	0.998	0.938	1	1	1	1	1	1
15	0.969	0.698	1	1	1	1	1	1
20	0.849	0.539	1	0.999	1	1	1	1
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.911	0.939	0.961	0.984	0.976	0.992	0.982	0.993
10	0.846	0.906	0.918	0.954	0.950	0.982	0.958	0.985
15	0.797	0.888	0.887	0.923	0.918	0.970	0.931	0.977
20	0.754	0.884	0.868	0.891	0.880	0.955	0.904	0.967
				Panel (e): Number of Included Variables				
5	13.432	22.894	8.682	9.824	7.292	7.387	6.706	7.011
10	23.871	36.743	17.374	23.279	14.483	15.145	13.783	14.432
15	31.781	42.315	24.614	36.808	21.997	23.576	20.860	21.505
20	36.685	43.351	30.538	50.470	29.561	32.514	27.663	29.237
				Panel (f): Number of Included Irrelevant Variables				
5	8.432	17.895	3.682	4.824	2.292	2.387	1.706	2.011
10	13.889	27.359	7.374	13.279	4.483	5.145	3.783	4.432
15	17.248	31.847	9.614	21.808	6.997	8.576	5.860	6.505
20	19.703	32.580	10.538	30.489	9.561	12.514	7.663	9.237

TABLE 6. 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**n*	100	300	100	300	100	300	100	300
				Panel (a): Correct Sparsity Pattern				
5	0.001	0.001	0.015	0.003	0.049	0.035	0.088	0.059
10	0	0	0	0	0.005	0.004	0.013	0.006
15	0	0	0	0	0.001	0	0.001	0.002
20	0	0	0	0	0	0	0.001	0
				Panel (b): True Model Included				
5	1	0.997	1	1	1	1	1	1
10	0.991	0.608	1	1	1	1	1	1
15	0.715	0.029	1	0.999	1	1	1	1
20	0.149	0	1	0.952	1	1	1	1
				Panel (c): Fraction of Relevant Variables Included				
5	1	0.999	1	1	1	1	1	1
10	0.997	0.910	1	1	1	1	1	1
15	0.957	0.659	1	0.999	1	1	1	1
20	0.829	0.513	1	0.980	1	1	1	1
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.857	0.912	0.926	0.967	0.954	0.984	0.964	0.986
10	0.797	0.884	0.878	0.923	0.918	0.969	0.931	0.975
15	0.757	0.884	0.847	0.888	0.875	0.952	0.897	0.963
20	0.738	0.883	0.827	0.859	0.833	0.932	0.863	0.951
				Panel (e): Number of Included Variables				
5	18.612	30.961	12.058	14.676	9.344	9.786	8.446	9.081
10	28.250	42.693	20.945	32.472	17.366	18.995	16.238	17.169
15	34.976	43.071	28.038	46.871	25.584	28.732	23.792	25.448
20	37.516	43.117	33.855	59.140	33.352	38.944	30.955	33.763
				Panel (f): Number of Included Irrelevant Variables				
5	13.612	25.968	7.058	9.676	4.344	4.786	3.446	4.081
10	18.277	33.595	10.945	22.472	7.366	8.995	6.238	7.169
15	20.625	33.183	13.038	31.885	10.584	13.732	8.792	10.448
20	20.936	32.865	13.855	39.539	13.352	18.944	10.955	13.763

Table 7 shows the mean squared error (MSE) for out-of-sample forecasts for *Autometrics* (Liberal and Conservative), LASSO, adaLASSO and oracle models. We consider a total of 100 out-of-sample observations. As expected, all methodologies improve their performance as the sample size increases, and the number of relevant and candidate variables decrease. For *T*=500, all

methodologies has similar performance out-of-sample to the Oracle. In a general way, *Autometrics* (Conservative) presents the lowest MSE (closest to the Oracle).

TABLE 7. FORECASTING: DESCRIPTIVE STATISTICS

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*	100	300	100	300	100	300	100	300
				MSE - Autometrics (Liberal)				
5	0.769	0.688	0.399	0.836	0.278	0.380	0.267	0.292
10	0.901	0.742	0.406	0.882	0.283	0.379	0.270	0.297
15	1.062	0.807	0.438	0.917	0.288	0.377	0.272	0.297
20	1.808	5.996	0.458	0.951	0.291	0.380	0.273	0.303
				MSE - Autometrics (Conservative)				
5	0.340	0.539	0.288	0.406	0.261	0.272	0.258	0.265
10	0.398	0.658	0.297	0.360	0.267	0.277	0.259	0.267
15	0.470	0.799	0.318	0.391	0.271	0.286	0.260	0.272
20	2.339	3.857	0.342	0.431	0.274	0.291	0.265	0.276
				MSE - LASSO				
5	0.464	0.605	0.336	0.378	0.278	0.288	0.266	0.272
10	0.920	3.378	0.419	0.539	0.295	0.319	0.280	0.288
15	3.373	10.939	0.556	0.806	0.313	0.350	0.286	0.304
20	9.388	17.599	0.810	1.680	0.326	0.385	0.295	0.323
				MSE - adaLASSO				
5	0.557	0.962	0.331	0.362	0.272	0.279	0.262	0.267
10	0.962	4.411	0.378	0.524	0.283	0.294	0.271	0.274
15	3.084	12.004	0.424	0.770	0.295	0.313	0.275	0.283
20	10.038	19.115	0.450	1.696	0.306	0.332	0.282	0.294
				MSE - Oracle				
5	0.278	0.278	0.264	0.263	0.255	0.256	0.252	0.253
10	0.314	0.314	0.277	0.278	0.258	0.258	0.257	0.254
15	0.361	0.359	0.296	0.296	0.263	0.263	0.257	0.256
20	0.422	0.420	0.312	0.314	0.268	0.270	0.261	0.261

2.4.2. Comparing the methodologies

To facilitate comparison between the model selection techniques, this section presents the "winner" and the "2nd winner" methodology on each of the statistics presented in Section 2.4.1. The results were obtained comparing the values in Table 2, Tables 3-6 and Table 7, for the *Autometrics* (Liberal), *Autometrics* (Conservative), LASSO and adaLASSO. The tables present the "winner" methodology for each simulated scenario, i.e. for each value of n, T (in the lines) and q (in the columns).

Table 8 presents the "winners" for parameters estimation, i.e. the methodology that provides the lowest average bias and average MSE for the estimator, using the values in Table 2, for each scenario. Both for bias and MSE, the *Autometrics* (Conservative) is the "winner" for almost all scenarios. To have a better visualization of the difference between methodologies, we plot in Figure 5 and Figure 6 the values in Table 2 (Bias and MSE, respectively), for n=100 and n=300. Each surface represents a model selection technique, according to the colors legend. The plot axes are the sample size (T) and the number of relevant variables (q). Figures 5 and 6 illustrate the results in Table 8, and show that both bias and MSE present a large increase when T is small and q is large, especially for LASSO and adaLASSO.

Table 9 and Table 10 give the "winners" for the selection statistics present in Tables 3-6. Table 9 provides the "winners" and "2nd winners" for Panel (a), and Panel (d), (e) and (f) (the three Panels present the same "winners") statistics in Tables 3-6, i.e., the methodology that maximizes Panel (a) and Panel (d) statistics, and minimizes Panel (e) and Panel (f) statistics (using the criterion that the more parsimonious the model selected, the better). Again, *Autometrics* (Conservative) is the "winner" for almost all scenarios.

The four methodologies present similar results for the statistics in Panel (b) and (c) of Tables 3-6, Table 10 provides all winners for each scenario for Panel (b) and (c) (both Panels present the same "winners"). In other words, for one scenario, more than one methodology can present the best result. Therefore, we are analyzing the "worse" methodology (or methodologies) for each simulated scenario. As noted in Tables 3-6, *Autometrics* (Liberal and Conservative) presents a better performance in including the true model only when T=50.

Figures 7-12 provide the plot for all Panels statistics in Tables 3-6. Figures 7, 10, 11 and 12 clearly indicate a superior performance of the *Autometrics* (Conservative) in the statistics of Panel (a), (d), (e) and (f) in Table 9. Figures 8 and 9 illustrate the results of Table 10 for the statistics in Panel (b) and (c).

Table 11 gives the "winner" for the forecasting, i.e. the methodology that presents the lowest MSE for the out-of-sample forecast, as presented in Table 7, for each scenario. *Autometrics* (Conservative) is the "winner" for almost all simulated scenarios, and the adaLASSO is the "2nd winner" in most of scenarios. Figure 13 plots the MSE of the forecast out-of-sample. It is clear that all

methodologies has similar performance out-of-sample, excluding the scenario where T=50 and q=15 or 20, in which LASSO and adaLASSO present a much larger MSE.

TABLE 8. PARAMETER ESTIMATES: WINNER

The table reports for each scenario, the winner and the 2nd winner for the average absolute bias and the average mean squared error (MSE), over all parameter estimates and Monte Carlo simulations. n is the number of candidate variables, q is the number of relevant regressors and T is the sample size.

n	*T**q*	5	10	15	20	5	10	15	20
			Bias - winner				Bias - 2nd winner		
100	50	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	Aut-L	Aut-L
	100	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	Aut-L	Aut-L
	300	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	Aut-L	Aut-L	Aut-L
	500	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	Aut-L	Aut-L	Aut-L
300	50	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	Aut-L	Aut-L	Aut-L
	100	LASSO	Aut-C	Aut-C	Aut-C	adaLASSO	LASSO	LASSO	Aut-L
	300	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	adaLASSO	adaLASSO	adaLASSO
	500	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	adaLASSO	adaLASSO
			MSE - winner				MSE - 2nd winner		
100	50	Aut-C	Aut-C	Aut-C	Aut-L	LASSO	Aut-L	Aut-L	Aut-C
	100	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	adaLASSO	adaLASSO
	300	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	Aut-L	Aut-L	Aut-L
	500	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	Aut-L	Aut-L
300	50	Aut-C	Aut-C	Aut-L	Aut-C	LASSO	Aut-L	Aut-C	Aut-L
	100	adaLASSO	Aut-C	Aut-C	Aut-C	LASSO	adaLASSO	adaLASSO	Aut-L
	300	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	adaLASSO	adaLASSO
	500	adaLASSO	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	adaLASSO

TABLE 9. MODEL SELECTION: WINNER
Panel (a) and Panel (d), (e) and (f)

The table reports for each scenario, the winner and the 2nd winner for Panel (a) statistics and Panel (d), (e) and (f) statistics. Panel (a) presents the fraction of replications where the correct model has been selected. 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.

n	*T**q*	5	10	15	20	5	10	15	20
			Panel (a) - winner				Panel (a) - 2nd winner		
100	50	Aut-C	Aut-C	Aut-C	Aut-C	Aut-L	Aut-L	Aut-L	Aut-L
	100	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	Aut-L	Aut-L	Aut-L
	300	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	LASSO	Aut-L
	500	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	LASSO	Aut-L
300	50	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	Aut-L	Aut-L	Aut-L
	100	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	Aut-L	Aut-L
	300	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	LASSO	Aut-L
	500	LASSO	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	LASSO
			Panel (d), (e) and (f) - winner				Panel (d), (e) and (f) - 2nd winner		
100	50	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	Aut-L	Aut-L
	100	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	Aut-L	Aut-L	Aut-L
	300	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	Aut-L	Aut-L
	500	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	Aut-L	Aut-L
300	50	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	Aut-L	Aut-L
	100	LASSO	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	LASSO
	300	LASSO	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	LASSO
	500	LASSO	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	LASSO	LASSO

TABLE 10. MODEL SELECTION: WINNERS
Panel (b) and (c)

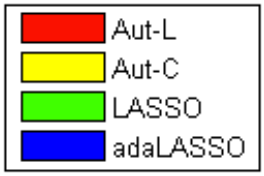
The table reports for each scenario, the winners for Panel (b) and (c) statistics. Panel (b) shows the fraction of replications where the relevant variables are all included. Panel (c) presents the fraction of relevant variables included.

n	*T**q*	5	10	15	20
			Panel (b) and (c) - winners		
100	50	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C	Aut-L/Aut-C	Aut-L
	100	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO
	300	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO
	500	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO
300	50	Aut-L/Aut-C	Aut-L/Aut-C	Aut-L/Aut-C	Aut-C
	100	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO	Aut-L/Aut-C
	300	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO
	500	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO	Aut-L/Aut-C/LASSO/adaLASSO

TABLE 11. FORECASTING: WINNER

The table reports for each scenario, the winner and the 2nd winner for the out-of-sample mean squared error (MSE). *n* is the number of candidate variables, *q* is the number of relevant regressors and *T* is the sample size.

n	*T**q*	5	10	15	20	5	10	15	20
		MSE - winner				MSE - 2nd winner			
100	50	Aut-C	Aut-C	Aut-C	Aut-L	LASSO	Aut-L	Aut-L	Aut-C
	100	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	adaLASSO	adaLASSO
	300	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	Aut-L	Aut-L	Aut-L
	500	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	Aut-L	Aut-L	Aut-L
300	50	Aut-C	Aut-C	Aut-C	Aut-C	LASSO	Aut-L	Aut-L	Aut-L
	100	adaLASSO	Aut-C	Aut-C	Aut-C	LASSO	adaLASSO	adaLASSO	Aut-L
	300	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	adaLASSO	adaLASSO
	500	Aut-C	Aut-C	Aut-C	Aut-C	adaLASSO	adaLASSO	adaLASSO	adaLASSO



Color Legend for Figures 5-13

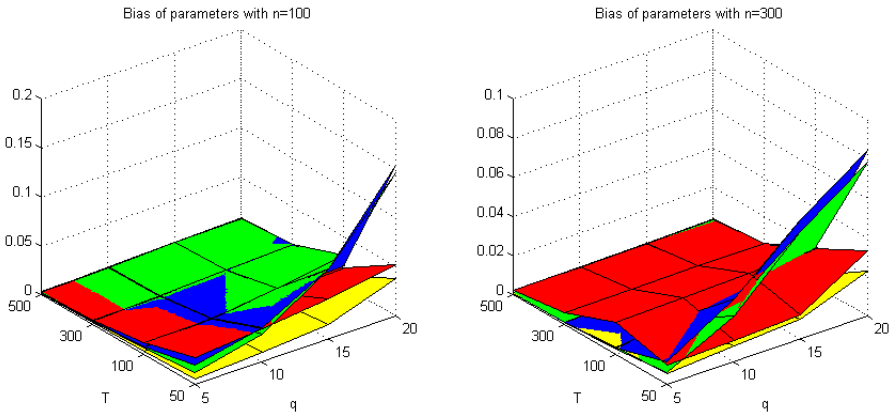


FIGURE 5. Average absolute bias, over all parameter estimates and Monte Carlo simulations, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). *n* is the number of candidate variables, *q* is the number of relevant regressors and *T* is the sample size.

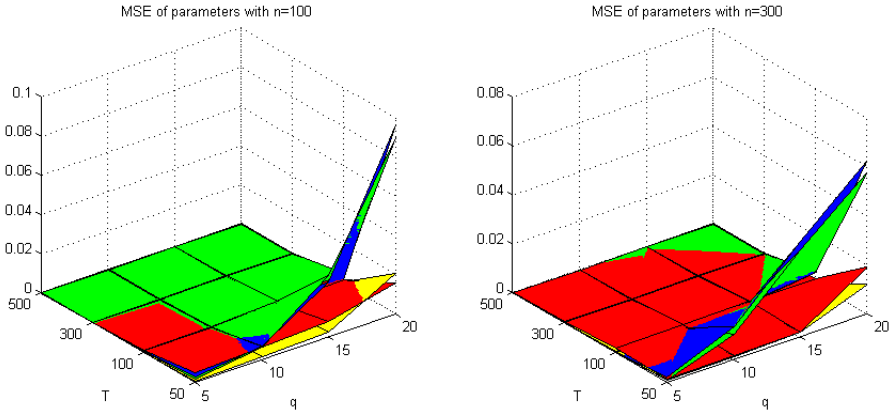


FIGURE 6. Average mean squared error (MSE), over all parameter estimates and Monte Carlo simulations, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). *n* is the number of candidate variables, *q* is the number of relevant regressors and *T* is the sample size.

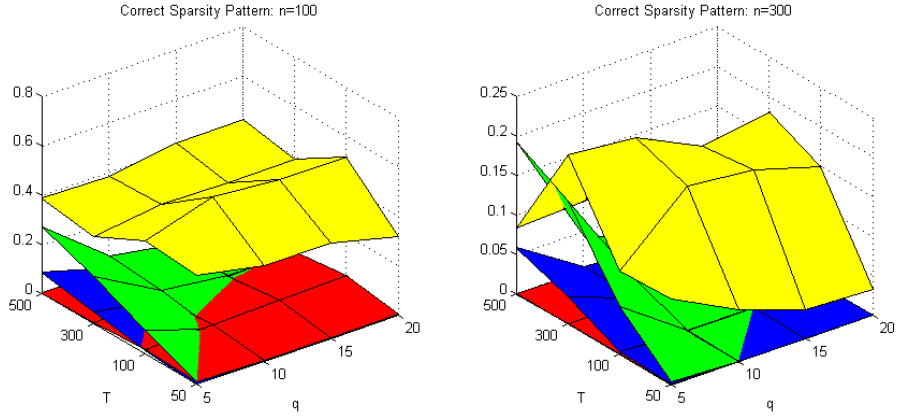


FIGURE 7. Panel (a): fraction of replications where the correct model has been selected, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). *n* is the number of candidate variables, *q* is the number of relevant regressors and *T* is the sample size.

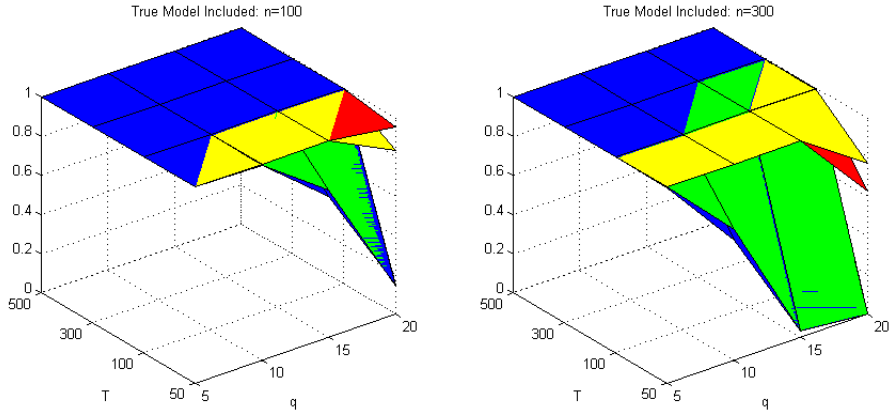


FIGURE 8. Panel (b): fraction of replications where the relevant variables are all included, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). *n* is the number of candidate variables, *q* is the number of relevant regressors and *T* is the sample size.

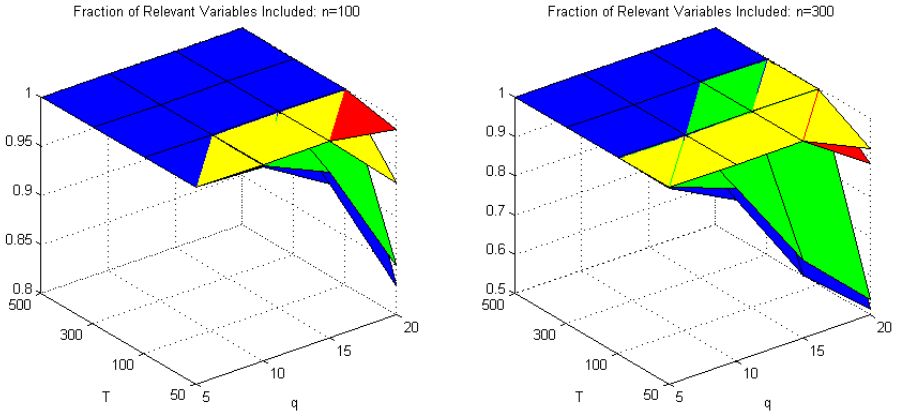


FIGURE 9. Panel (c): fraction of relevant variables included, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). n is the number of candidate variables, q is the number of relevant regressors and T is the sample size.

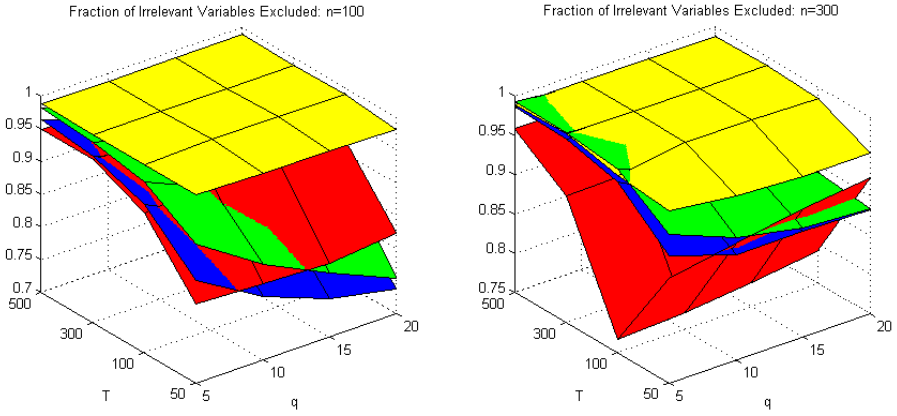


FIGURE 10. Panel (d): fraction of irrelevant variables excluded, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). n is the number of candidate variables, q is the number of relevant regressors and T is the sample size.

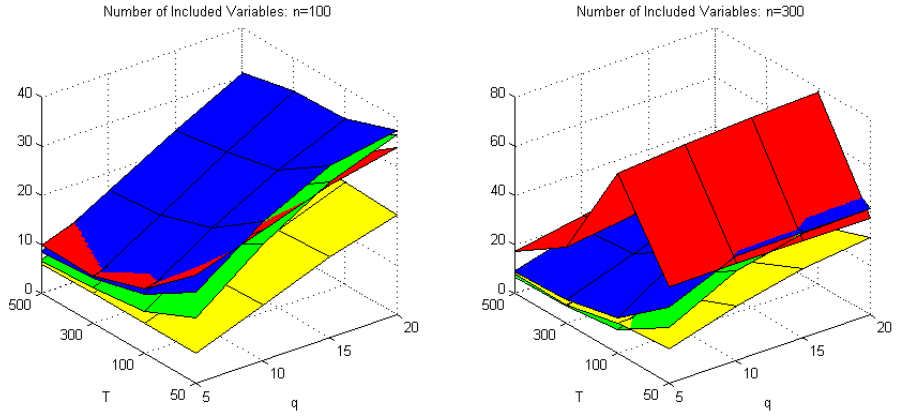


FIGURE 11. Panel (e): average number of included variables, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). n is the number of candidate variables, q is the number of relevant regressors and T is the sample size.

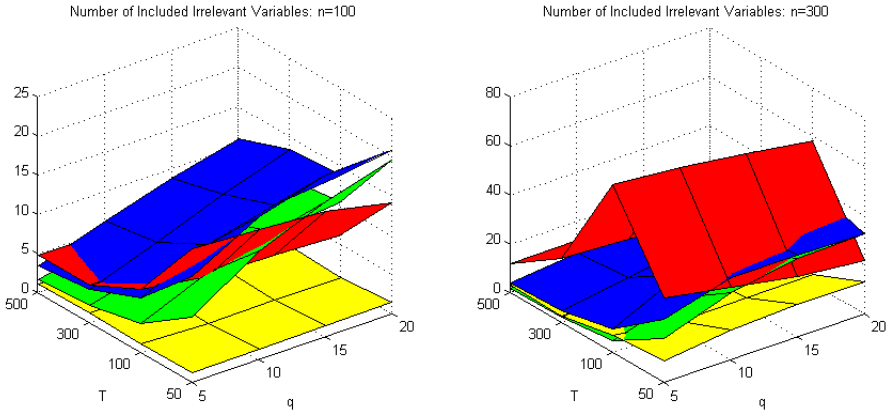


FIGURE 12. Panel (f): average number of included irrelevant regressors, for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). *n* is the number of candidate variables, *q* is the number of relevant regressors and *T* is the sample size.

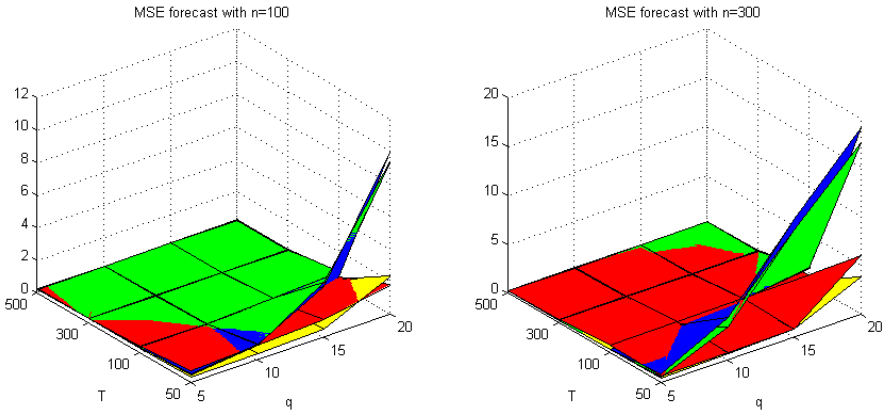


FIGURE 13. Out-of-sample mean squared error (MSE), for Aut-L (red), Aut-C (yellow), LASSO (green) and adaLASSO (blue). *n* is the number of candidate variables, *q* is the number of relevant regressors and *T* is the sample size.

2.5. Application to real data – GDP forecasting

An important problem in economics is the prediction of the future evolution of GDP growth. Forecasts are typically produced either from economic theory based models or from simple linear time series models. A time series model can provide a reasonable benchmark to evaluate the value added of economic theory relative to the pure explanatory power of the past behavior of the variable. The problem faced by economists is that the economic growth theory is not explicit about what variables can predict growth one or more steps ahead. This difficulty has led empirical economists to follow theory loosely and simply "try" various sets of variables relating the potentially important determinants of growth. However, the regressions combining various variables showed that inclusion of one variable

could affect the significance of the others. Since we do not know *a priori* the "true" variables that should be included, we are left with the question: what are the variables that can really explain growth?

An initial answer to this question was given by Levine and Renelt (1992). They used cross-country regressions to identify "robust" empirical relations in the economic growth literature. Moreover, Sala-I-Martin, X. (1997) combined 62 variables in two millions cross-sectional regressions, and succeed to identify some variables strongly related to growth.

In this section, our aim is to use model selection techniques to identify significant variables in the forecast of growth, using time series regression. We compare the results using *Autometrics* (Liberal and Conservative), LASSO, adaLASSO, and some benchmark models.

2.5.1. The model

In order to compare the different model selection techniques in terms of predictive power (forecast out-of-sample), we consider the quarterly US GDP one-step ahead forecasting. Section 2.5.3.1 also study the identification of relevant variables in the linear regression.

Recent empirical literature on economic growth has identified a substantial number of variables that are partially correlated with the rate of economic growth. The basic methodology, presented in the literature for time series forecast of the GDP, consists of running linear regression with explanatory variables, which vary across researchers and papers.

Our main goal is selecting the variables that can explain growth using the techniques compared in this chapter, and performing one-step ahead out-of-sample forecast using the selected model. For that, we define the general unrestricted model (GUM) in equation (9).

$$y_t = \sum_{i=1}^{4} \beta_{0,i}\, y_{t-i} + \sum_{k=1}^{n} \sum_{i=1}^{I} \beta_{k,i}\, x_{k,t-i} + \varepsilon_t, \qquad \varepsilon_t \sim IN[0, \sigma^2] \tag{9}$$

where $x_{k,t-i}$ is the i^{th} lag of the k^{th} explanatory variable, and $I = 1$ or 4.

2.5.2.
The data

The dataset was obtained from the Federal Reserve Bank of Philadelphia and is part of the database called "Real-Time Data Set for Macroeconomists", which consists of vintages of major macroeconomic variables[6]. For the present work we used only the vintage available at the third quarter of 2011, which contains quarterly data from the first quarter of 1959 and ends in the second quarter of 2011, say 210 observations. The dependent variable corresponds to the US real GDP. We consider a total of 64 explanatory variables, related in Table 12.

All variables have been pretested for unit-roots and first-differenced whenever necessary. After taking the first difference and the first four lags of variables, the time series was reduced to 205 observations, and the sample period covers the second quarter of 1960 to the second quarter of 2011. We use 80% of the data for the in-sample specification and estimation of the models, and the final 20% for the out-of-sample forecast, according to Table 13.

TABLE 12 - Data Description

Name	Series
ROUTPUT	GDP – (dependent variable)
RCON	Real Personal Consumption Expenditures: Total
RCONG	Real Personal Consumption Expenditures: Goods
RCONND	Real Personal Consumption Expenditures: Nondurable Goods
RCOND	Real Personal Consumption Expenditures: Durable Goods
RCONS	Real Personal Consumption Expenditures: Services
RCONSHH	Real Household Consumption Expenditures for Services
RCONSNP	Real Final Consumption Expenditures of NPISH
RINVBF	Real Gross Private Domestic Investment: Nonresidential
RINVRESID	Real Gross Private Domestic Investment: Residential
RINVCHI	Real Gross Private Domestic Investment: Change in Private Inventories
REX	Real Exports of Goods and Services
RIMP	Real Imports of Goods and Services
RG	Real Government and Gross Investment: Total
RGF	Real Government and Gross Investment: Federal
RGSL	Real Government and Gross Investment: State and Local

[6] "A real-time data set shows the observations for a variable as those observations were revised over time. The Philadelphia Fed's real-time data set records snapshots, or vintages, of the data as they existed at various points in time in the past, before the data were fully revised. The vintage is an important concept in a real-time data set: It refers to the date on which the data were available to the public". For more details see: *http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/data-files/documentation/gen_doc_NIPA.pdf*

NOUTPUT	Nominal GNP/GDP
NCON	Nominal Personal Consumption Expenditures
NCONG	Nominal Personal Consumption Expenditures: Goods
NCONSHH	Nominal Household Consumption Expenditures for Services
NCONSNP	Nominal Household Consumption Expenditures of NPISH
NCONHH	Nominal Household Consumption Expenditures
WSD	Wage and Salary Disbursements
OLI	Other Labor Income
PROPI	Proprietors' Income
RENTI	Rental Income of Persons
DIV	Dividens
PINTI	Personal Interest Income
TRANR	Transfer Payments
SSCONTRIB	Personal Contributions for Social Insurance
NPI	Nominal Personal Income
PTAX	Personal Tax and Nontax Payments
NDPI	Nominal Disposable Personal Income
PINTPAID	Interest Paid by Consumers
TRANPF	Personal Transfer Payments to Foreigners
NPSAV	Nominal Personal Saving
RATESAVE	Personal Saving Rate, Constructed
NCPROFAT	Nominal Corporate Profits After Tax Without IVA/CCAdj
NCPROFATW	Nominal Corporate Profits After Tax With IVA/CCAdj
PCON	Price Index for Personal Consumption Expenditures, Constructed
PCONG	Price Index for Personal Consumption Expenditures: Goods
PCONSHH	Price Index for Household Consumption Expenditures for Services
PCONSNP	Price Index for Final Consumption Expenditures of NPISH
PCONHH	Price Index for Household Consumption Expenditures
PCONX	Core Price Index for Personal Consumption Expenditures
PIMP	Price Index for Imports of Goods and Services
POP	Civilian Noninstitutional Population, 16+
LFC	Civilian Labor Force, 16+
LFPART	Civilian Participation Rate, 16+, Constructed
RUC	Unemployment Rate
EMPLOY	Thousands of employees, seasonally adjusted
HG	Indexes of Aggregate Weekly Hours: Goods-Producing
OPH	Output Per Hour: Business Sector
ULC	Unit Labor Costs: Business Sector
IPT	Industrial Production Index: Total
IPM	Industrial Production Index: Manufacturing
CUM	Capacity Utilization Rate: Manufacturing
HSTARTS	Housing Starts
BASEBASA	Monetary Base
CPI	Consumer Price Index
M1	M1 Money Stock
M2	M2 Money Stock
TRBASA	Total Reserves
NBRBASA	Nonborrowed Reserves
NBRECBASA	Nonborrowed Reserves Plus Extended Credit

TABLE 13. GDP FORECASTING: TRAINING AND TEST PERIOD

The table reports, the number of observation, the quarter of the beginning and the quarter of the end of each period in the linear regression.

Period Description	Number of Observations	Beginning	End
Total	205	1960 q2	2011 q2
Training (in-sample)	164	1960 q2	2001 q1
Test set (out-of-sample)	41	2001 q2	2011 q2

The data set provided in Table 12 is composed by some highly correlated variables. Some are simply linear combination of others. In order to test the model selection techniques in the presence of collinearity, we used two different sets of explanatory variables. In other words, we compared the methodologies using two different GUMs. In the first GUM we included the 64 explanatory variables and their lags, and in the second GUM we included only the 46 variables that appear in bold in Table 12, taking out highly correlated variables.

In order to compare the methodologies in both cases, where the number of candidate variables exceeds the number of observations ($n>T$), and where the number of candidate variables does not exceeds the sample size ($n<T$), we used two different sets of lags in each GUM. The first set of regressors (GUM_1) is composed by the first 4 lags of the GDP $(y_{t-1}, \dots, y_{t-4})$, and the first lag of the explanatory variables (X_{t-1}). The second set of variables (GUM_2) is composed by the first 4 lags of the GDP $(y_{t-1}, \dots, y_{t-4})$, and the first 4 lags of the explanatory variables $(X_{t-1}, \dots, X_{t-4})$. In other words, we tested variable selection methodologies for four different scenarios, given by eq. (9):

1. GUM_1 with 64 explanatory variables
2. GUM_2 with 64 explanatory variables
3. GUM_1 with 46 explanatory variables
4. GUM_2 with 46 explanatory variables

2.5.3. Results

We compare *Autometrics* (Liberal and Conservative), LASSO, and adaLASSO, with three different benchmark alternatives: a linear regression with all the

regressors included (GUM), a simple first-order AR model (AR(1)), and a fourth-order AR model (AR(4)). The results are presented in Table 14.

In the case where we consider all 64 variables (Panel 1), we observe that both LASSO and adaLASSO models are far superior to the benchmarks and *Autometrics* models for the out-of-sample results. In this case, with the presence of highly correlated candidate variables, *Autometrics* models present negative out-of-sample R^2 (R2_out in Table 14)[7] except in the case of Aut-C_1, where R2_out = 0.169. However, in the case where we only considered 46 variables (Panel 2), the *Autometrics* (Liberal) performs similar to LASSO and adaLASSO models, when $n>T$ (Aut-L_2). Comparing results in Panel 1 and Panel 2, Table 14 shows that *Autometrics* is very sensible to changes in the set of candidate variables and to collinearity between the regressors, while LASSO and adaLASSO models are more stable.

We notice also that the model with all regressors (GUM) presents the worse out-of-sample performance, and, as expected, the largest in-sample R^2, due to the high number of parameters, in all cases. The GUM_2, in Panel 1 and Panel 2, present lower BIC than model selection techniques, due to over parameterization ($n>T$). GUM models present negative out-of-sample R^2 in every case. Negative values of out-of-sample R^2 mean that it is worse using the selected model than simply using the historical mean of the dependent variable as forecast.

Table 14 shows that the best model out-of-sample within *Autometrics* (Liberal and Conservative), LASSO and adaLASSO (largest out-of-sample R^2, and lowest RMSE), is the worse in-sample model (largest BIC) in most of the cases. This fact gives an indication that using the information criteria (in-sample) to select models may not be a good strategy for forecasting, as we can notice in the case of models Aut-L_2 and Aut-C_2 in Panel 1: they present the lowest BIC of four compared techniques and the lowest out-of-sample R^2 and largest out-of-sample RMSE. Finally, in the GDP forecasting application, we conclude that the best model selection technique is the adaLASSO, when $n<T$, and the LASSO, when $n>T$.

[7] The out-of-sample R^2 is the pseudo-R^2, given by:

$$R^2{}_{out} = 1 - \sum_{t \in O} \frac{(\hat{y}_t - y_t)^2}{(y_t - \bar{y})^2}$$

where O is the out-of-sample observations set and $\bar{y}$ is the historical mean of the in-sample set.

TABLE 14. GDP FORECASTING: DESCRIPTIVE STATISTICS

The table reports, for each different set of explanatory variables and different set of lags, the in-sample and out-of-sample R^2, the out-of-sample root mean squared error (RMSE), the Bayesian Information Criterion (BIC), and the number of parameters, for each model selection technique, and benchmark models.

Model	R2_in	R2_out	RMSE_out	BIC	num par
		Benchmark models			
AR(1)	-0.092	0.342	73.867	8.008	1
AR(4)	0.107	0.174	82.799	7.975	4
		Panel 1: Considering 64 explanatory variables			
GUM_1	0.731	-857.087	2668.001	8.792	68
LASSO_1	0.420	0.394	70.886	7.570	3
adaLASSO_1	0.475	0.428	68.891	7.502	5
Aut-L_1	0.586	-248.630	1439.000	7.450	11
Aut-C_1	0.548	0.169	83.044	7.413	7
GUM_2	1.000	-3124.647	5092.022	-27.317	260
LASSO_2	0.459	0.359	72.909	7.592	6
adaLASSO_2	0.521	0.230	79.912	7.513	9
Aut-L_2	0.671	-858.510	2670.200	7.470	19
Aut-C_2	0.569	-774.670	2536.600	7.365	7
		Panel 2: Considering 46 explanatory variables			
GUM_1	0.640	-212.322	1330.265	8.522	50
LASSO_1	0.414	0.376	71.943	7.580	3
adaLASSO_1	0.465	0.413	69.793	7.494	4
Aut-L_1	0.513	0.144	84.263	7.426	5
Aut-C_1	0.513	0.144	84.263	7.426	5
GUM_2	1.000	-32048.468	16305.374	-24.792	188
LASSO_2	0.448	0.358	72.949	7.613	6
adaLASSO_2	0.507	0.284	77.071	7.516	8
Aut-L_2	0.635	0.320	75.103	7.449	15
Aut-C_2	0.528	-684.100	2384	7.426	6

2.5.3.1. Study of relevant variables

In this section we present the relevant variables in the models selected by the model selection techniques. Table 15 and Table 16 show variables per model, the total number of models in which each variable appears, the total number of parameters of each model, and the variables description, for the experiment considering 64 and 46 explanatory variables, respectively. We consider here that models presenting negative out-of-sample R^2 were not able to select a satisfactory model for the US GDP, therefore they are not considered in this section.

Table 15 contains the variables present in models LASSO_1, adaLASSO_1, LASSO_2, adaLASSO_2 and Aut-C_1, considering the GUM with 64 explanatory variables. LASSO_1, adaLASSO_1 and Aut-C_1 were selected from a set of 68 candidate variables (GUM_1), while LASSO_2 and

adaLASSO_2 were selected from a set of 260 candidate variables (GUM_2), as presented in Panel 1 of Table 14.

Analyzing the relevant variables in Table 15, we notice that three regressors are present in all LASSO and adaLASSO models: Real Household Consumption Expenditures for Services; Real Imports of Goods and Services; and Real Gross Private Domestic Investment: Residential. This can be an indication that these three economic variables can explain GDP better than the other candidate variables.

TABLE 15. GDP FORECASTING: RELEVANT VARIABLES – 64 EXPLANATORY VARIABLES

The table shows the relevant variables (regressors) per model (only models with positive out-of-sample R^2 in Panel 1 of Table 14: LASSO_1, adaLASSO_1, LASSO_2, adaLASSO_2 and Aut-C_1), the number of models in which each variable appears, the number of parameters of each model, and the variables description.

Considering 64 explanatory variables							
Variable	LASSO_1	adaLASSO_1	LASSO_2	adaLASSO_2	Aut-C_1	**num models**	Description
EMPLOY (t-1)					X	**1**	Thousands of employees, seasonally adjusted
HSTARTS (t-1)				X		**1**	Housing Starts
NCPROFATW (t-1)					X	**1**	Nominal Corporate Profits After Tax With IVA/CCAdj
NPSAV (t-1)					X	**1**	Nominal Personal Saving
PROPI (t-2)			X	X		**2**	Proprietors' Income
RCOND (t-1)		X		X		**2**	Real Personal Consumption Expenditures: Durable Goods
RCONND (t-2)			X	X		**2**	Real Personal Consumption Expenditures: Nondurable Goods
RCONS (t-1)		X		X	X	**3**	Real Personal Consumption Expenditures: Services
RCONSHH (t-1)	X	X	X	X		**4**	Real Household Consumption Expenditures for Services
RIMP (t-1)	X	X	X	X		**4**	Real Imports of Goods and Services
RINVRESID (t-1)	X	X	X	X	X	**5**	Real Gross Private Domestic Investment: Residential
ROUTPUT (t-1)					X	**1**	Real GNP/GDP
TRANPF (t-1)					X	**1**	Personal Transfer Payments to Foreigners
ULC (t-2)			X	X		**2**	Unit Labor Costs: Business Sector
num par	**3**	**5**	**6**	**9**	**7**		

Table 16 contains the variables present in models LASSO_1, adaLASSO_1, LASSO_2, adaLASSO_2, Aut-L/C_1, and Aut-C_2, considering the GUM with 46 explanatory variables. Aut-L/C_1 represents Aut-L_1 and Aut-C_1 as both methodologies resulted in the same selected model. LASSO_1, adaLASSO_1 and Aut-L/C_1 were selected from a set of 50 candidate variables (GUM_1), while LASSO_2, adaLASSO_2 and Aut-L_2 were selected from a set of 188 candidate variables (GUM_2), as presented in Panel 2 of Table 14.

In Table 16, two regressors are present in all six models: Real Personal Consumption Expenditures: Services; and Real Gross Private Domestic Investment: Residential. And one regressor is present in all LASSO and adaLASSO models: Real Imports of Goods and Services. Once again this can be an indication that these three economic variables can explain GDP better than the others candidate variables.

Tables 15 and 16 show that LASSO_2 and adaLASSO_2, present the same regressors that LASSO_1 and adaLASSO_1, and some more. In other words, LASSO_1 is nested in LASSO_2 and adaLASSO_1 is nested in adaLASSO_2. This indicates that there is some information about the GDP in the 3 lags that does not enter the GUM_1, but enters the GUM_2 ($X_{t-2}, \dots, X_{t-4}$), even if out-of-sample R^2 of LASSO_2 and adaLASSO_2 are lower.

TABLE 16. GDP FORECASTING: RELEVANT VARIABLES – 46 EXPLANATORY VARIABLES

The table shows the relevant variables (regressors) per model (only models with positive out-of-sample R^2 in Panel 2 of Table 14: LASSO_1, adaLASSO_1, LASSO_2, adaLASSO_2, Aut-L/C_1, and Aut-C_2), the number of models in which each variable appears, the number of parameters of each model, and the variables description.

Considering 46 explanatory variables								
Variable	LASSO_1	adaLASSO_1	LASSO_2	adaLASSO_2	Aut-L/C_1	Aut-L_2	**num models**	Description
EMPLOY (t-1)					X	X	**2**	Thousands of employees, seasonally adjusted
EMPLOY (t-3)						X	**1**	Thousands of employees, seasonally adjusted
HG (t-1)						X	**1**	Indexes of Aggregate Weekly Hours: Goods-Producing
HSTARTS (t-1)				X			**1**	Housing Starts
IPT (t-3)						X	**1**	Industrial Production Index: Total
OPH (t-1)						X	**1**	Output Per Hour: Business Sector
PROPI (t-2)			X	X			**2**	Proprietors' Income
RCOND (t-1)		X		X			**2**	Real Personal Consumption Expenditures: Durable Goods
RCONND (t-2)			X	X		X	**3**	Real Personal Consumption Expenditures: Nondurable Goods
RCONS (t-1)	X	X	X	X	X	X	**6**	Real Personal Consumption Expenditures: Services
RCONSNP (t-2)						X	**1**	Real Final Consumption Expenditures of NPISH
RGSL (t-2)						X	**1**	Real Government and Gross Investment: State and Local
RIMP (t-1)	X	X	X	X			**4**	Real Imports of Goods and Services
RIMP (t-2)						X	**1**	Real Imports of Goods and Services
RINVBF (t-2)						X	**1**	Real Gross Private Domestic Investment: Nonresidential
RINVRESID (t-1)	X	X	X	X	X	X	**6**	Real Gross Private Domestic Investment: Residential
ROUTPUT (t-1)					X		**1**	Real GNP/GDP
SSCONTRIB (t-4)						X	**1**	Personal Contributions for Social Insurance
TRANPF (t-1)					X	X	**2**	Personal Transfer Payments to Foreigners
TRANR (t-1)						X	**1**	Transfer Payments
ULC (t-2)			X	X			**2**	Unit Labor Costs: Business Sector
num par	**3**	**4**	**6**	**8**	**5**	**15**		

2.6. Conclusion

In this chapter we evaluated a comparison between model selection techniques. First we explain the *PcGets* and *Autometrics* methodology and algorithms, and the LASSO and adaLASSO techniques and estimators. For the comparison we evaluate a Monte Carlo simulation with a simple linear regression as DGP, with orthogonal regressors, and an application to US GDP forecasting. Two aspects of the performance were considered: the predictive power (forecast out-of-sample) and the performance in the selection of the correct model and estimation (in-sample). The case where the number of candidate variables exceeds the number of observation was considered as well.

In the simulation experiment, we compared the different model selection methodologies in different simulated scenarios, varying the sample size (*T*), the

number of relevant variables (q) and the number of candidate variables (n). In a general way, we can conclude that when T is large (greater than 300) and q is small (smaller than 10), all selection techniques present a similar performance. However, in extreme cases with more candidate variables than observations ($n > T$), and large values of number of relevant variables (q), which increases the variance of the dependent variable y, *Autometrics* (Conservative) presents better results.

It is important to notice that final models selected by the methodologies in this chapter always present less regressors than observations, as expected in a linear regression. Regressions with more variables that observations only appear in the GUM models, before variable selection.

Results in Section 2.4.1 also showed that adaLASSO estimator gets close to the Oracle and consistently chooses the relevant variables as the number of observations increases (model selection consistency). This result provides empirical evidence that the adaLASSO enjoys the oracle properties, while LASSO does not, as shown in Zou (2006).

In the application to US GDP time series we changed to a different setup, with dependent data, i.e. candidate variables are no longer orthogonal. In this case models estimated by LASSO and adaLASSO procedures delivered out-of-sample forecasts significantly superior than *Autometrics* and benchmark models. Results also show that the selected models outperform the general unrestricted models (GUM). This is an important result as the methodologies compared in this chapter are based on the assumption that only a few numbers of candidate variables are in fact relevant to explain the dynamics of the dependent variable (sparse models). In this application, forecasts results suggest that LASSO and adaLASSO procedures are more robust than *Autometrics* algorithm when we have structure dependence in regressors.

In Section 2.5 we don't know the real DGP for the GDP series, so it is impossible to compare selection performance of each methodology. However, *Autometrics* gives the OLS estimate, while LASSO and adaLASSO have a penalty on the OLS estimate. So, if both methods are correctly selecting variables, we expect that *Autometrics* presents lower out-of-sample RMSE. As this is not the case, we have an indicative that LASSO and adaLASSO have better perform in variable selection for the data.

Simulation and application results differ in pointing out the "best" model selection technique. In the simulation, the *Autometrics* (Conservative) showed slightly better results in a general way, while in the application to real data, LASSO and adaLASSO outperformed *Autometrics*. Maybe these results are due to the very simplistic and unrealistic DGP used in the simulation experiment, and give an indication that *Autometrics* underperforms LASSO and adaLASSO with dependent regressors (more realistic scenarium). Medeiros and Mendes (2013) proved that adaLASSO correctly selects the relevant variables and has the oracle property in a time series framework with a very general error term. The comparison of adaLASSO with *Autometrics* in a framework of correlated regressors is an important subject for future research.

Results in Section 2.5 also lead us to an important question: The "best" model in-sample is the "best" model out-of-sample? Under a stationary environment the answer is yes, but, in real world, results can vary according to data and the sample split point, as showed in Hansen and Timmermann (2012). This question has motivated Chapter 4 of this thesis, where an "out-of-sample" error has an important role in the variable selection method.

Based on the main results of simulation experiment and application to dependent data, we can conclude that for a realistic high-dimensional statistical model, LASSO and adaLASSO outperform *Autometrics* selection algorithm, especially when regressors are not orthogonal. Also, LASSO and adaLASSO procedures present an important advantage: computational time. *Autometrics* is based on a clever tree-path search algorithm, however, the number of models to specify increases exponentially with the number of candidate variables, while regularization methods find the optimal solution using much more efficient optimization algorithms. For those main reasons, next chapters of this thesis are based on variable selection using shrinkage methods with penalized regressions.