3
Variable selection for STR/STAR models

3.1.
Introduction

The main literature on model selection and regularization methods concern linear regressions. There is a relatively small number of papers on model selection involving interactions or non-linearity. Choi et al. (2010) extended the LASSO method for identifying interaction terms that can be included in the model only if the corresponding main terms are also includes in the model, automatically enforcing the heredity constraint.

Tateishi et al. (2010) applied LASSO regularization to non-linear regression models constructed with Gaussian basis functions. The authors proposed a weighted penalty to select the number of basis functions, but do not perform actual variable selection. Also based on basis functions, Ravikumar et al. (2009) proposed a generalization of the LASSO to non-linear basic additive models, using a penalty on the ℓ_2-norm of the main effects. Radchenko and James (2010) introduced a new approach based on a penalized least square criterion for non-linear problems. They use a preselected finite orthonormal basis functions with respect to Lebesgue measure on the unit square that they assume to be a good approximation for the true non-linear model.

In this chapter, we introduce a variable selection methodology for smooth transition regressive (STR) and autoregressive (STAR) models based on LASSO regularization. We present a direct approach and a stepwise approach, with and without the heredity constraint, as explained in next sections.

3.2.
STR - LASSO

Our goal here is to fit the non-linear model (10) and to find out which terms, especially which non-linear terms, have an important effect on the response. The model is composed potentially by p linear regressors (linear main effects),

PUC-Rio - Certificação Digital Nº 0821483/CA

$(n_c * p)$ non-linear main effects, $(n_c * p * p)$ interactions of first order and $(n_c * p * p * ((n_c * p) + 1)/2)$ interactions of second order. Let $\mathrm{h}_j(x_{k,i})$ be a non-linear function where $\boldsymbol{x}_i \in \mathbb{R}^p$, and the parameter vector $\psi = (\boldsymbol{\beta}, \boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}, \boldsymbol{\alpha_2})' \in \mathbb{R}^r$, where r is the total number of parameters, we define the non-linear model:

$$y_i = \sum_{k=1}^{p} \beta_k \, x_{k,i} + \sum_{j=1}^{n_c} \sum_{k=1}^{p} \alpha_{0,j,k} \, \mathrm{h}_j(x_{k,i}) + \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{l=1}^{p} \alpha_{1,j,k,l} \mathrm{h}_j(x_{k,i}) x_{l,i}$$
$$+ \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{a=j}^{n_c} \sum_{b=1}^{p} \sum_{l=1}^{p} \alpha_{2,j,k,a,b,l} \mathrm{h}_j(x_{k,i}) \mathrm{h}_a(x_{b,i}) x_{l,i} + \varepsilon_i, \qquad (10)$$
$$\varepsilon_i \sim \mathrm{IN}[0, \sigma^2]$$
$$i = 1, \ldots T$$

The number of coefficient to be estimated ($\boldsymbol{\beta}, \boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$) can increase exponentially when p is large. However, we assume that only a small fraction of the main effects and interaction terms are present in the true model. In order to fit the model (10) we assume that the non-linear function $\mathrm{h}_j(.)$ is the logistic function expressed in (11). The choice of the logistic function is usual in neural networks, and we know that STR model is a particular case of a neural network with a single hidden layer, which is a function universal approximator. The logistic function is also of easy interpretation, which is important, as we want to be able to interpret the variables selected.

$$\mathrm{h}_j(x_{k,i}) = G(x_{k,i}; \gamma, c_j) = \frac{1}{1 + e^{-\gamma(x_{k,i} - c_j)}} \qquad (11)$$

The logistic function $G(.;.,.)$ represents smooth transitions, where the slope parameter $\gamma \in \mathbb{R}^+$ controls the smoothness of the function. The parameter $c_j \in \mathbb{R}, j = 1, \ldots n_c$, is called the location parameter, where the variable x_k is the transition variable.

Substituting (11) in (10) we have a STR (Smooth Transition Regressive) model, class of non-linear models proposed in Chan and Tong (1986).

PUC-Rio - Certificação Digital Nº 0821483/CA

$$
\begin{aligned}
y_i = \sum_{k=1}^{p} \beta_k \, x_{k,i} + \sum_{j=1}^{n_c} \sum_{k=1}^{p} \alpha_{0,j,k} \, G\left(x_{k,i}; \gamma, c_j\right) \\
+ \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{l=1}^{p} \alpha_{1,j,k,l} G\left(x_{k,i}; \gamma, c_j\right) x_{l,i} \\
+ \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{a=j}^{n_c} \sum_{b=1}^{p} \sum_{l=1}^{p} \alpha_{2,j,k,a,b,l} G\left(x_{k,i}; \gamma, c_j\right) G\left(x_{b,i}; \gamma, c_a\right) x_{l,i} + \varepsilon_i, \\
\varepsilon_i \sim \mathrm{IN}[0, \sigma^2] \\
i = 1, \ldots T
\end{aligned}
\tag{12}
$$

It is important to notice that when the slope parameter approaches infinity, the model approaches a TR (Threshold Regression) model, where

$$
G\left(x_{k,i}; c_j\right) = \begin{cases} 1, & x_{k,i} \leq c_j \\ 0, & x_{k,i} > c_j \end{cases} \tag{13}
$$

Figure 14 shows the logistic function for different values of γ, with $c_j = 0$.

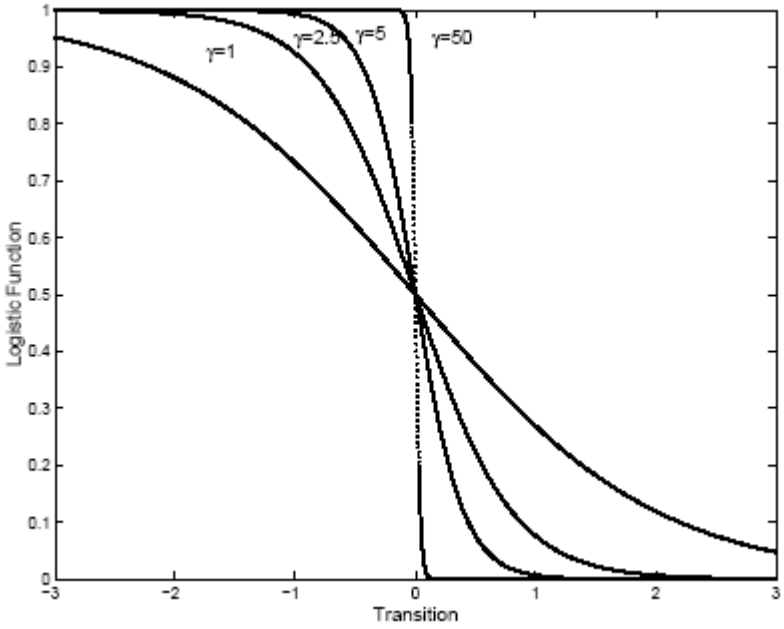

FIGURE 14. Logistic function with fixed parameters

In order to simplify the notation, we can write eq. (12) as

PUC-Rio - Certificação Digital Nº 0821483/CA

$$\mathrm{Y} = \sum_{k=1}^{p} \mathrm{G}_k + \sum_{j=1}^{n_c}\sum_{k=1}^{p} \mathrm{G}_{jk} + \sum_{j=1}^{n_c}\sum_{k=1}^{p}\sum_{l=1}^{p} \mathrm{G}_{jkl} + \sum_{j=1}^{n_c}\sum_{k=1}^{p}\sum_{a=j}^{n_c}\sum_{b=1}^{p}\sum_{l=1}^{p} \mathrm{G}_{jkabl} + \epsilon \qquad (14)$$

where $\mathbf{y} = (y_1, \dots, y_T)'$, $\mathbf{x}_k = (x_{k,1}, \dots, x_{k,T})'$, $\epsilon = (\varepsilon_1, \dots, \varepsilon_T)'$ and

$\mathrm{G}_k = \beta_k \mathbf{x}_k$ → linear main effects

$\mathrm{G}_{jk} = \alpha_{0,j,k} G(\mathbf{x}_k; \gamma, c_j)$ → non-linear main effects

$\mathrm{G}_{jkl} = \alpha_{1,j,k,l} G(\mathbf{x}_k; \gamma, c_j)\mathbf{x}_l$ → interactions of first order

$\mathrm{G}_{jkabl} = \alpha_{2,j,k,a,b,l} G(\mathbf{x}_k; \gamma, c_j) G(\mathbf{x}_b; \gamma, c_a)\mathbf{x}_l$ → interactions of second order

Notice that, even when p is small, the number of candidate variables in (14) will easily be larger than the number of observations ($n > T$). Our general approach for fitting (14) is to minimize the following penalized regression criterion,

$$\|\mathbf{y} - \mathbf{G}\|^2 + P(\mathbf{G})$$

where

$$\mathbf{G} = \sum_{k=1}^{p} \mathrm{G}_k + \sum_{j=1}^{n_c}\sum_{k=1}^{p} \mathrm{G}_{jk} + \sum_{j=1}^{n_c}\sum_{k=1}^{p}\sum_{l=1}^{p} \mathrm{G}_{jkl} + \sum_{j=1}^{n_c}\sum_{k=1}^{p}\sum_{a=j}^{n_c}\sum_{b=1}^{p}\sum_{l=1}^{p} \mathrm{G}_{jkabl} \qquad (15)$$

Penalty function $P(\mathbf{G})$, given by (16), is similar to the penalty function of the LASSO, presented in (1). We also showed in Section 2.4 that LASSO methodology presented good results in the case where the number of candidate variables is larger than the number of observations. Therefore we can apply LASSO to fit the STR model in (14). Furthermore, as we show in Section 2.4, the adaLASSO outperforms the LASSO in many situations and enjoy the oracle property (see Chapter 2). So, in order to penalize different parameters differently, we also use the adaLASSO penalty $P'(\mathbf{G})$ given in (17), which is analogous to penalty term in eq. (2).

$$P(\mathbf{G}) = \lambda \left(\sum_{k=1}^{p} |\beta_k| + \sum_{j=1}^{n_c}\sum_{k=1}^{p} |\alpha_{0,j,k}| + \sum_{j=1}^{n_c}\sum_{k=1}^{p}\sum_{l=1}^{p} |\alpha_{1,j,k,l}| + \sum_{j=1}^{n_c}\sum_{k=1}^{p}\sum_{a=j}^{n_c}\sum_{b=1}^{p}\sum_{l=1}^{p} |\alpha_{2,j,k,a,b,l}| \right) \qquad (16)$$

PUC-Rio - Certificação Digital Nº 0821483/CA

$$P'(\mathbf{G}) = \lambda \left(\begin{array}{l} \sum_{k=1}^{p} \widehat{w}_k |\beta_k| + \sum_{j=1}^{n_c} \sum_{k=1}^{p} \widehat{w}_{0,j,k} |\alpha_{0,j,k}| + \\ \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{l=1}^{p} \widehat{w}_{1,j,k,l} |\alpha_{1,j,k,l}| + \\ \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{a=j}^{n_c} \sum_{b=1}^{p} \sum_{l=1}^{p} \widehat{w}_{2,j,k,a,b,l} |\alpha_{2,j,k,a,b,l}| \end{array} \right) \quad (17)$$

where

$\widehat{w}_k = \frac{1}{|\hat{\beta}_k^*|^{\tau}}$, $\widehat{w}_{0,j,k} = \frac{1}{|\alpha_{0,j,k}^*|^{\tau}}$, $\widehat{w}_{1,j,k,l} = \frac{1}{|\alpha_{1,j,k,l}^*|^{\tau}}$, $\widehat{w}_{2,j,k,a,b,l} = \frac{1}{|\alpha_{2,j,k,a,b,l}^*|^{\tau}}$,

$\tau > 0$; $\hat{\beta}_k^*, \alpha_{0,j,k}^*, \alpha_{1,j,k,l}^*$ and $\alpha_{2,j,k,a,b,l}^*$ are the elastic net estimates, and τ is selected by BIC criterion, as explained in Section 2.3.1 and 2.3.2.

Section 2.5 showed that LASSO and adaLASSO can handle correlated candidate variables, so we expect that the inclusion of non-linear effects and interaction will not be a problem for variable selection and model specification. However, a possible disadvantage of LASSO and adaLASSO is that $P(\mathbf{G})$ and $P'(\mathbf{G})$ treat main effects and interaction similarly. Choi et al. (2010) argued that when interaction terms exists, there is a natural hierarchy among the variables, that is, an interaction term can be included in the model only if both of the corresponding main terms are also included in the model. They claim that if an interaction term is selected but not the corresponding main terms, the model becomes difficult to interpret in practice.

Likewise Radchenko and James (2010) argued that adding an interaction when the corresponding main effects are not present results in two new predictors, which in practice is equivalent to adding two main effects, and that interaction terms are more difficult to interpret than main effects. Therefore, given similar predictive ability, they argued that it is preferable to add a main effect ahead of an interaction.

Motivated by this idea, we propose a variable selection methodology based on LASSO/adaLASSO structured on stages and group of variables. We call it Group Stepwise LASSO/adaLASSO (GS-LASSO/GS-adaLASSO).

PUC-Rio - Certificação Digital Nº 0821483/CA

3.2.1. Group Stepwise LASSO

The goal of Group Stepwise LASSO is to fit the non-linear model (12) and select relevant variables, enforcing the heredity constraint. In other words, an interaction term can only be added to the model, if both corresponding main effect (linear and non-linear) are also included. This strategy significantly reduces the complexity of the high-dimensional data, making the problem more tractable. To simplify, we consider only interactions of first order.

Algorithm:

1. Linear main effects:

   Solve the penalized regression criterion (15), with $\mathbf{G} = \sum_{k=1}^{p} \mathrm{G}_k$:

$$\widehat{\boldsymbol{\beta}}^1 = arg\min_{\beta} \left\| \mathbf{y} - \sum_{k=1}^{p} \beta_k \mathbf{x}_k \right\|^2 + \lambda \sum_{k=1}^{p} |\beta_k| \tag{18}$$

2. Non-linear main effects:

   Let k_{r1} and p_1 be the index and the number of variables selected ($\hat{\beta}^1 \neq 0$) in step 1, respectively. Set the selected variables fixed and solve the penalized regression criterion (15), with $\mathbf{G} = \sum_{j=1}^{n_c} \sum_{k=1}^{p} \mathrm{G}_{jk}$:

$$(\widehat{\boldsymbol{\beta}}^2, \widehat{\boldsymbol{\alpha}_{\mathbf{0}}}^2) = arg\min_{\beta,\alpha_0} \left\| \mathbf{y} - \sum_{k_{r1}=1}^{p_1} \beta_{k_{r1}} \mathbf{x}_{k_{r1}} - \sum_{j=1}^{n_c} \sum_{k=1}^{p} \alpha_{0,j,k} G\left(\mathbf{x}_k; \gamma, c_j\right) \right\|^2 + \lambda \sum_{j=1}^{n_c} \sum_{k=1}^{p} |\alpha_{0,j,k}| \tag{19}$$

3. Interactions of first order:

   Let jk_{r2} and p_2 be the index and the number of variables selected ($\hat{\alpha}_0{}^2 \neq 0$) in step 2, respectively. Set the selected variables in step 1 and step 2 fixed and solve the penalized regression criterion (15), with $\mathbf{G} = \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{l=1}^{p} \mathrm{G}_{jkl}$, using only the p_1 linear main effects and the p_2 non-linear main effects selected in step 1 and 2:

PUC-Rio - Certificação Digital Nº 0821483/CA

$$(\widehat{\boldsymbol{\beta}}^3, \widehat{\boldsymbol{\alpha}_0}^3, \widehat{\boldsymbol{\alpha}_1}^3) = arg \min_{\beta,\alpha_0,\alpha_1} \left\| \mathbf{y} - \sum_{k_{r1}=1}^{p_1} \beta_{k_{r1}} \mathbf{x}_{k_{r1}} - \sum_{jk_{r2}=1}^{p_2} \alpha_{0,jk_{r2}} G_{jk_{r2}} - \sum_{jk_{r2}=1}^{p_2} \sum_{k_{r1}=1}^{p_1} \alpha_{1,jk_{r2},k_{r1}} G_{jk_{r2}} \mathbf{x}_{k_{r1}} \right\|^2 + \lambda \left(\sum_{jk_{r2}=1}^{p_2} \sum_{k_{r1}=1}^{p_1} |\alpha_{1,jk_{r2},k_{r1}}| \right) \quad (20)$$

4. Post-selection (optional):

   The vector $(\widehat{\boldsymbol{\beta}}^3, \widehat{\boldsymbol{\alpha}_0}^3, \widehat{\boldsymbol{\alpha}_1}^3)$ is the penalized final estimated vector. Let $jk_{r2}k_{r1}$ and p_3 be the index and the number of variables selected ($\hat{\alpha}_1^3 \neq 0$) in step 3, respectively. Evaluate OLS estimation, with the variables selected in step 1, 2 and 3 fixed.

$$(\widehat{\boldsymbol{\beta}}^{post}, \widehat{\boldsymbol{\alpha}_0}^{post}, \widehat{\boldsymbol{\alpha}_1}^{post}) = arg \min_{\beta,\alpha_0,\alpha_1} \left\| \mathbf{y} - \sum_{k_{r1}=1}^{p_1} \beta_{k_{r1}} \mathbf{x}_{k_{r1}} - \sum_{jk_{r2}=1}^{p_2} \alpha_{0,jk_{r2}} G_{jk_{r2}} - \sum_{jk_{r2}k_{r1}=1}^{p_3} \alpha_{1,jk_{r2}k_{r1}} G_{jk_{r2}k_{r1}} \right\|^2 \quad (21)$$

3.3. Simulation exercises

In our simulation study we tested the variable selection methodology for STR models presented in Section 3.2 and 3.2.1.

As in Section 2.4, we aim to evaluate the 'size' and 'power' of the model selection process, and the forecast accuracy of the selected models. We also want to compare LASSO applied to the whole set of candidate variables (main effects and interactions) at one time, and the Group Stepwise LASSO approach. Therefore, we tested the variable selection methodology for STR models in five non-linear simulation exercises over 1000 Monte Carlo replications.

For the simulation exercises, we considered different scenarios: varying the set of candidate variables, where linear main effect, non-linear main effects, interactions of first order and interactions of second order are represent as X1, X2, X3 and X4, respectively; the simulations can present an heredity structure for the DGP or not; we can use the direct non-linear methodology (LASSO) or the Group Stepwise methodology (GS-LASSO); and relevant variables present in the DGP

PUC-Rio - Certificação Digital Nº 0821483/CA

may not be in the set of candidate variables (true model not available). It is important to note that when we refer to the STR – LASSO, it includes the adaLASSO estimator as well. Table 17 presents a summary of the simulation scenarios.

TABLE 17. SIMULATION EXERCISES: SCENARIOS

The table reports the different scenarios for each simulation exercise, concerning the set of candidate variables, the structure for the DGP, the methodology applied, and the availability of the true model variables.

Simulation	Candidate Variables				Structure for the DGP		Methodology		True model
	X1	X2	X3	X4	heredity	non-heredity	LASSO	GS-LASSO	not available
1	✓	✓	✓		✓		✓	✓	
2	✓	✓	✓			✓	✓	✓	
3	✓	✓	✓	✓		✓	✓		
4	✓	✓	✓	✓	-	-	✓		✓
5	✓	✓	✓		✓		✓	✓	

3.3.1. Simulation 1

For the first simulation, we consider only linear main effects, non-linear main effects and interactions of first order on the data generating process (DGP), as presented in (22):

$$y_i = \sum_{k=1}^{q_1} \beta_k \, x_{k,i} + \sum_{jk=1}^{q_2} \alpha_{0,jk} G_{jk,i} + \sum_{jkl=1}^{q_3} \alpha_{1,jkl} G^1_{jkl,i} + 0.5 \, \varepsilon_i,$$
$$\varepsilon_i \sim \text{IN}[0,1]$$
$$\boldsymbol{x}_i \sim \text{IN}_{q_1}\left[0, \boldsymbol{I}_{q_1}\right] \quad (22)$$
$$i = 1, \dots T$$

where, $\boldsymbol{\beta}, \boldsymbol{\alpha_0}$ and $\boldsymbol{\alpha_1}$ are vectors of ones of size q_1, q_2 and q_3, respectively; $\boldsymbol{x}_i$ is a vector of q_1 relevant linear main effects, $\boldsymbol{G}_i$ is a vector of q_2 relevant non-linear main effects, and $\boldsymbol{G}^1_i$ is a vector of q_3 interactions of first order, as shown in (23):

$$\boldsymbol{G}_{j,k} = G\left(\mathbf{x}_k; \gamma, c_j\right)$$
$$\boldsymbol{G}^1_{jkl} = G\left(\mathbf{x}_k; \gamma, c_j\right)\mathbf{x}_l \quad (23)$$

where $G\left(\mathbf{x}_k; \gamma, c_j\right)$ is the logistic function according to (11).

As candidate variables, we consider n linear main effects, which generates a total of $(n + (n_c * n) + (n_c * n * n)$ candidate variables, where n_c is the

PUC-Rio - Certificação Digital Nº 0821483/CA

number of different percentiles used for the location parameter c_j in the logistic function. In all simulations exercises we considered $n_c = 5$, the percentiles of 20%, 35%, 50%, 65% and 80% of the transition variable for the values of c_j, and the slope parameter $\gamma = 10$. We tested setting γ with other values, as 2.5 and 5, but the results were similar, so we chose to fix this parameter to reduce algorithm and results complexity. We simulate T = 50, 100, 300, 500 observations of DGP (22) for different combinations of candidate (n) and relevant ($q = q_1 + q_2 + q_3$) variables. We consider n = 5 (total of 155 candidate variables) and n = 10 (total of 560 candidate variables) and q = 5, 10, 15, 20. The relevant q variables are available among the candidate variables. Simulation 1 assumes the heredity structure for the true model (DGP). 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.

Tables 18-21 present variable selection results for the direct LASSO and adaLASSO applied to the STR model, the GS-LASSO and the GS-adaLASSO, following the format and statistics of tables in Section 2.4.1: 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; and Panel (f) shows the average number of included irrelevant regressors.

As in the case of linear regression, the selection performance of all methodologies improves with the sample size (T) and gets worse as the number of relevant variables (q) increases. In general, the GS-LASSO presents slightly better results than the LASSO, but the adaLASSO shows the better performance then the GS-adaLASSO.

Table 22 shows the mean squared error (MSE) for one-step ahead out-of-sample forecasts for the direct LASSO and adaLASSO applied to the STR model, the GS-LASSO, the GS-adaLASSO, the post GS-LASSO, the post GS-adaLASSO, and the oracle model (OLS estimator in a regression including only the relevant variables). The post GS-LASSO and post GS-adaLASSO models have their parameters estimated in post-selection, as described in equation (21). We consider a total of 100 out-of-sample observations.

PUC-Rio - Certificação Digital Nº 0821483/CA

As expected, all methodologies improve their performance as the sample size increases, and the number of relevant and candidate variables decrease. The direct LASSO and adaLASSO show slight better results then the others. The post GS-LASSO and post GS-adaLASSO did not improve the forecast results. The adaLASSO presented the best general result in simulation 1.

TABLE 18. MODEL SELECTION: DESCRIPTIVE STATISTICS
LASSO – Simulation 1

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	0	0	0	0	0	0.002	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	0.771	0.393	0.994	0.959	1	1	1	1
10	0.090	0.003	0.809	0.406	1	0.998	1	1
15	0.001	0	0.059	0.014	0.653	0.436	0.833	0.499
20	0	0	0	0	0.218	0.089	0.29	0.121
				Panel (c): Fraction of Relevant Variables Included				
5	0.947	0.824	0.999	0.992	1	1	1	1
10	0.730	0.504	0.976	0.897	1	1	1	1
15	0.606	0.343	0.874	0.733	0.975	0.962	0.989	0.967
20	0.483	0.219	0.745	0.529	0.951	0.926	0.961	0.934
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.930	0.970	0.940	0.976	0.943	0.983	0.946	0.987
10	0.882	0.955	0.889	0.954	0.882	0.969	0.887	0.977
15	0.889	0.949	0.894	0.940	0.865	0.962	0.868	0.973
20	0.882	0.945	0.877	0.926	0.836	0.941	0.839	0.953
				Panel (e): Number of Included Variables				
5	15.189	20.834	13.932	18.06	13.493	14.241	13.083	12.417
10	24.435	30.049	25.8	34.384	27.072	26.878	26.386	22.453
15	24.611	32.817	27.897	43.641	33.479	34.933	33.331	29.105
20	25.642	33.89	31.437	50.732	41.157	50.418	41.02	43.814
				Panel (f): Number of Included Irrelevant Variables				
5	10.456	16.712	8.939	13.101	8.493	9.241	8.083	7.417
10	17.139	25.013	16.04	25.413	17.072	16.88	16.386	12.453
15	15.524	27.669	14.789	32.649	18.86	20.498	18.503	14.606
20	15.982	29.517	16.545	40.152	22.129	31.904	21.794	25.129

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 19. MODEL SELECTION: DESCRIPTIVE STATISTICS
adaLASSO – Simulation 1

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*	5	10	5	10	5	10	5	10
				Panel (a): Correct Sparsity Pattern				
5	0.009	0.001	0.084	0.06	0.165	0.222	0.188	0.538
10	0	0	0	0	0.002	0.038	0.003	0.253
15	0	0	0	0	0	0	0	0.001
20	0	0	0	0	0	0	0	0
				Panel (b): True Model Included				
5	0.524	0.209	0.964	0.897	1	1	1	1
10	0.025	0	0.631	0.259	1	0.997	1	0.999
15	0	0	0.001	0.004	0.08	0.019	0.125	0.005
20	0	0	0	0	0.001	0.003	0	0
				Panel (c): Fraction of Relevant Variables Included				
5	0.854	0.697	0.992	0.976	1	1	1	1
10	0.595	0.392	0.937	0.819	1	1	1	1
15	0.462	0.256	0.768	0.609	0.896	0.933	0.905	0.933
20	0.350	0.168	0.608	0.432	0.879	0.892	0.886	0.905
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.962	0.981	0.979	0.992	0.984	0.997	0.986	0.999
10	0.918	0.967	0.940	0.973	0.945	0.993	0.951	0.997
15	0.921	0.962	0.932	0.958	0.902	0.987	0.906	0.994
20	0.915	0.958	0.921	0.944	0.874	0.970	0.878	0.982
				Panel (e): Number of Included Variables				
5	10.027	14.221	8.125	9.541	7.474	6.706	7.128	5.675
10	17.903	22.289	18.011	23.232	17.991	13.847	17.057	11.499
15	17.991	24.735	21.003	32.106	27.131	21.041	26.804	17.521
20	18.448	26.188	22.788	38.767	34.585	33.9	34.243	27.966
				Panel (f): Number of Included Irrelevant Variables				
5	5.755	10.736	3.167	4.659	2.474	1.706	2.128	0.675
10	11.949	18.368	8.64	15.038	7.991	3.854	7.057	1.501
15	11.058	20.9	9.485	22.977	13.69	7.04	13.229	3.53
20	11.439	22.826	10.625	30.137	17.012	16.066	16.532	9.876

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 20. MODEL SELECTION: DESCRIPTIVE STATISTICS
GS-LASSO – Simulation 1

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.

GS-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	0.001	0	0	0	0	0	0
10	0.012	0	0.062	0	0.126	0	0.171	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	0.519	0.292	0.866	0.654	1	1	1	1
10	0.477	0.012	0.843	0.013	1	0.006	1	0.006
15	0.005	0	0.097	0.064	0.562	0.586	0.724	0.713
20	0	0	0.002	0.005	0.079	0.398	0.208	0.673
				Panel (c): Fraction of Relevant Variables Included				
5	0.715	0.589	0.920	0.792	1	1	1	1
10	0.826	0.663	0.950	0.747	1	0.801	1	0.801
15	0.649	0.623	0.794	0.717	0.966	0.970	0.979	0.981
20	0.428	0.464	0.497	0.473	0.699	0.892	0.824	0.976
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.970	0.988	0.979	0.993	0.984	0.995	0.985	0.995
10	0.952	0.977	0.974	0.989	0.982	0.990	0.984	0.992
15	0.895	0.965	0.900	0.980	0.878	0.972	0.880	0.973
20	0.922	0.959	0.914	0.978	0.869	0.954	0.859	0.951
				Panel (e): Number of Included Variables				
5	8.04	9.684	7.791	7.617	7.406	8.021	7.203	7.556
10	15.246	19.41	13.201	13.372	12.674	13.308	12.3	12.627
15	24.423	28.382	25.928	21.793	31.512	29.840	31.444	29.595
20	19.101	31.393	21.533	21.186	31.662	42.782	35.533	45.718
				Panel (f): Number of Included Irrelevant Variables				
5	4.463	6.741	3.193	3.655	2.406	3.021	2.203	2.556
10	6.982	12.784	3.703	5.902	2.674	5.296	2.300	4.615
15	14.690	19.039	14.016	11.041	17.028	15.295	16.759	14.883
20	10.542	22.113	11.598	11.732	17.687	24.942	19.044	26.197

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 21. MODEL SELECTION: DESCRIPTIVE STATISTICS
GS-adaLASSO – Simulation 1

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.

GS-adaLASSO								
	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	0.002	0	0	0	0	0	0
10	0.016	0	0.1	0	0.185	0	0.239	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	0.545	0.334	0.866	0.684	1	1	1	1
10	0.464	0.011	0.841	0.013	1	0.005	1	0.003
15	0.005	0	0.09	0.073	0.387	0.54	0.422	0.496
20	0	0	0.001	0.001	0.018	0.319	0.071	0.476
				Panel (c): Fraction of Relevant Variables Included				
5	0.732	0.616	0.920	0.810	1	1	1	1
10	0.828	0.669	0.950	0.753	1	0.801	1	0.801
15	0.634	0.631	0.788	0.720	0.948	0.966	0.952	0.966
20	0.416	0.462	0.481	0.477	0.672	0.886	0.786	0.964
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.974	0.987	0.982	0.994	0.987	0.995	0.988	0.996
10	0.954	0.976	0.978	0.990	0.985	0.992	0.988	0.993
15	0.900	0.965	0.904	0.981	0.889	0.975	0.891	0.976
20	0.923	0.957	0.919	0.978	0.876	0.957	0.867	0.956
				Panel (e): Number of Included Variables				
5	7.576	10.141	7.367	7.278	6.954	7.515	6.793	7.113
10	15.012	19.738	12.702	13.090	12.152	12.609	11.742	12.012
15	23.534	28.536	25.273	21.101	29.817	28.363	29.477	27.811
20	18.665	32.246	20.581	21.569	30.202	40.876	33.724	43.006
				Panel (f): Number of Included Irrelevant Variables				
5	3.918	7.062	2.769	3.226	1.954	2.515	1.793	2.113
10	6.732	13.045	3.205	5.56	2.152	4.6	1.742	4.006
15	14.031	19.07	13.455	10.294	15.599	13.876	15.204	13.316
20	10.348	23.001	10.96	12.036	16.764	23.163	18.001	23.736

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 22. FORECASTING: DESCRIPTIVE STATISTICS
Simulation 1

The table reports for each different sample size, the one-step ahead mean squared error (MSE) for each model selection technique. n is the number of candidate linear main effects 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	0.278	0.281	0.264	0.265	0.254	0.254	0.252	0.252
10	0.322	0.322	0.279	0.279	0.259	0.260	0.255	0.253
15	0.376	0.372	0.295	0.297	0.264	0.265	0.257	0.257
20	0.453	0.438	0.315	0.315	0.267	0.268	0.261	0.260
				MSE - LASSO				
5	0.505	0.669	0.348	0.395	0.276	0.298	0.267	0.285
10	1.530	2.369	0.600	0.833	0.296	0.408	0.277	0.373
15	3.114	5.246	1.236	1.975	0.302	0.653	0.279	0.569
20	5.249	9.786	2.119	4.461	0.312	1.249	0.287	0.999
				MSE - adaLASSO				
5	0.468	0.639	0.300	0.326	0.262	0.263	0.257	0.255
10	1.235	2.044	0.405	0.555	0.278	0.274	0.266	0.257
15	2.480	4.648	0.654	1.208	0.292	0.294	0.274	0.272
20	4.169	9.107	1.179	2.972	0.303	0.340	0.281	0.288
				MSE - GS-LASSO				
5	1.034	1.471	0.456	0.741	0.259	0.261	0.257	0.256
10	1.425	3.052	0.501	1.238	0.268	0.762	0.261	0.742
15	2.257	4.032	0.921	1.680	0.309	0.346	0.283	0.288
20	5.211	9.546	2.443	4.403	0.836	0.820	0.515	0.347
				MSE - GS-adaLASSO				
5	0.968	1.484	0.444	0.691	0.258	0.258	0.256	0.254
10	1.541	3.045	0.487	1.178	0.263	0.756	0.258	0.740
15	2.729	4.422	0.935	1.639	0.293	0.324	0.273	0.275
20	5.973	10.567	2.499	4.523	0.842	0.761	0.525	0.321
				MSE - post GS-LASSO				
5	0.962	1.582	0.449	0.732	0.258	0.259	0.256	0.254
10	1.492	54.853	0.489	1.224	0.264	0.752	0.258	0.736
15	2.490	5.152	0.913	1.661	0.294	0.323	0.274	0.275
20	5.482	13.175	2.416	4.519	0.811	0.766	0.503	0.323
				MSE - post GS-adaLASSO				
5	0.943	1.507	0.443	0.690	0.258	0.258	0.256	0.254
10	1.439	6.678	0.486	1.175	0.263	0.756	0.258	0.740
15	2.573	4.457	0.929	1.632	0.292	0.324	0.273	0.275
20	5.656	11.845	2.485	4.534	0.842	0.761	0.525	0.321

PUC-Rio - Certificação Digital Nº 0821483/CA

3.3.2. Simulation 2

The second simulation is similar to the first, but now the heredity structure does not hold for the true model (DGP). The results for model selection statistics are presented in Tables 23-26, and forecasting results in Table 27.

The LASSO and adaLASSO presented a superior performance comparing to GS-LASSO and GS-adaLASSO both in variable selection as in forecast

accuracy. When *T* increases the adaLASSO presents a one-step ahead MSE close to the oracle.

The difference of performance between the direct and the Group Stepwise methodologies is much higher in simulation 2. This is not a surprising result since the Group Stepwise algorithm can only select models with the heredity structure, and the DGP in simulation 2 does not assume the heredity structure.

We can also notice that the LASSO and adaLASSO present better results in simulation 1. This may indicate that the direct methodology performs better under the heredity constrain.

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 23. MODEL SELECTION: DESCRIPTIVE STATISTICS
LASSO – Simulation 2

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	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	0.356	0.148	0.761	0.706	0.969	0.988	0.996	0.998
10	0.011	0	0.144	0.004	0.839	0.049	0.967	0.047
15	0.001	0	0.007	0.012	0.730	0.897	0.879	0.978
20	0	0	0	0	0.035	0	0.063	0
				Panel (c): Fraction of Relevant Variables Included				
5	0.821	0.682	0.950	0.935	0.994	0.998	1	1
10	0.669	0.319	0.876	0.723	0.984	0.896	0.997	0.903
15	0.605	0.280	0.810	0.693	0.981	0.993	1	0.999
20	0.465	0.246	0.698	0.538	0.929	0.886	1	0.898
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.920	0.965	0.931	0.973	0.932	0.981	0.936	0.986
10	0.894	0.949	0.903	0.944	0.887	0.962	0.890	0.972
15	0.891	0.944	0.897	0.928	0.857	0.951	0.861	0.964
20	0.871	0.948	0.866	0.930	0.822	0.944	0.825	0.958
				Panel (e): Number of Included Variables				
5	16.165	22.736	15.115	19.917	15.153	15.363	14.532	13.023
10	22.125	31.164	22.795	38.162	26.268	29.655	25.988	24.670
15	24.366	34.783	26.541	49.561	34.697	41.664	34.316	34.609
20	26.706	32.989	31.992	48.759	42.548	47.760	42.442	40.388
				Panel (f): Number of Included Irrelevant Variables				
5	12.061	19.324	10.366	15.243	10.184	10.375	9.536	8.025
10	15.433	27.971	14.038	30.933	16.432	20.697	16.021	15.644
15	15.296	30.584	14.392	39.159	19.977	26.768	19.437	19.631
20	17.399	28.066	18.037	37.999	23.973	30.050	23.605	22.424

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 24. MODEL SELECTION: DESCRIPTIVE STATISTICS
adaLASSO – Simulation 2

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*	5	10	5	10	5	10	5	10
				Panel (a): Correct Sparsity Pattern				
5	0.002	0	0.007	0.003	0.028	0.056	0.018	0.133
10	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0.004
20	0	0	0	0	0	0	0	0
				Panel (b): True Model Included				
5	0.073	0.025	0.156	0.144	0.228	0.217	0.248	0.244
10	0	0	0	0	0.191	0.000	0.360	0
15	0	0	0	0.003	0.287	0.369	0.398	0.281
20	0	0	0	0	0.006	0.000	0.023	0
				Panel (c): Fraction of Relevant Variables Included				
5	0.646	0.499	0.764	0.748	0.819	0.798	0.836	0.810
10	0.497	0.242	0.748	0.640	0.884	0.840	0.918	0.840
15	0.481	0.217	0.709	0.591	0.945	0.948	0.958	0.942
20	0.338	0.193	0.587	0.462	0.872	0.852	0.888	0.867
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.950	0.976	0.967	0.987	0.970	0.994	0.971	0.997
10	0.926	0.961	0.943	0.964	0.930	0.987	0.931	0.991
15	0.926	0.957	0.940	0.948	0.898	0.981	0.905	0.990
20	0.911	0.960	0.916	0.948	0.860	0.972	0.866	0.983
				Panel (e): Number of Included Variables				
5	10.696	15.563	8.835	10.840	8.564	7.089	8.550	5.930
10	15.640	23.678	15.705	26.130	18.964	15.816	19.133	13.089
15	17.572	26.925	19.034	37.268	28.495	24.677	27.699	19.558
20	18.834	25.680	23.105	37.586	36.307	32.425	35.804	26.397
				Panel (f): Number of Included Irrelevant Variables				
5	7.468	13.069	5.013	7.098	4.469	3.098	4.371	1.880
10	10.668	21.260	8.226	19.733	10.122	7.415	9.949	4.692
15	10.356	23.669	8.402	28.404	14.316	10.454	13.332	5.422
20	12.071	21.816	11.372	28.345	18.862	15.377	18.050	9.050

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 25. MODEL SELECTION: DESCRIPTIVE STATISTICS
GS-LASSO – Simulation 2

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.

GS-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	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	0.003	0	0.001	0	0.005	0	0.001	0
10	0	0	0.003	0	0.048	0	0.104	0
15	0	0	0	0	0.016	0.011	0.021	0.013
20	0	0	0	0	0.001	0	0	0
				Panel (c): Fraction of Relevant Variables Included				
5	0.609	0.593	0.604	0.602	0.668	0.636	0.706	0.679
10	0.686	0.633	0.700	0.644	0.801	0.696	0.844	0.718
15	0.617	0.659	0.670	0.740	0.820	0.823	0.856	0.847
20	0.422	0.452	0.500	0.463	0.638	0.664	0.672	0.692
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.864	0.945	0.925	0.976	0.879	0.967	0.871	0.962
10	0.876	0.958	0.923	0.979	0.843	0.973	0.831	0.964
15	0.866	0.951	0.893	0.967	0.837	0.966	0.824	0.957
20	0.876	0.960	0.854	0.979	0.811	0.969	0.801	0.968
				Panel (e): Number of Included Variables				
5	23.457	33.702	14.308	16.390	21.455	21.310	22.881	24.529
10	24.895	29.402	18.181	18.119	30.771	21.868	33.006	26.798
15	28.014	36.692	25.064	29.087	35.162	30.828	37.492	36.199
20	25.180	30.427	29.681	20.598	38.203	29.918	40.316	31.244
				Panel (f): Number of Included Irrelevant Variables				
5	20.414	30.736	11.288	13.381	18.116	18.128	19.352	21.134
10	18.031	23.068	11.181	11.678	22.766	14.907	24.565	19.621
15	18.755	26.801	15.014	17.993	22.863	18.485	24.651	23.494
20	16.732	21.393	19.674	11.337	25.453	16.645	26.881	17.406

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 26. MODEL SELECTION: DESCRIPTIVE STATISTICS
GS-adaLASSO – Simulation 2

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.

GS-adaLASSO								
	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	0.003	0	0.001	0	0.001	0	0.000	0
10	0	0	0.002	0	0.021	0	0.028	0
15	0	0	0	0	0.008	0.008	0.014	0.009
20	0	0	0	0	0.001	0	0	0
				Panel (c): Fraction of Relevant Variables Included				
5	0.607	0.595	0.604	0.602	0.643	0.624	0.667	0.651
10	0.684	0.634	0.698	0.640	0.779	0.691	0.804	0.712
15	0.611	0.661	0.668	0.741	0.815	0.820	0.849	0.840
20	0.419	0.439	0.492	0.458	0.623	0.659	0.656	0.689
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.878	0.948	0.927	0.978	0.886	0.968	0.883	0.965
10	0.880	0.958	0.926	0.980	0.849	0.974	0.842	0.965
15	0.869	0.950	0.894	0.967	0.844	0.967	0.833	0.959
20	0.879	0.958	0.861	0.979	0.820	0.971	0.811	0.969
				Panel (e): Number of Included Variables				
5	21.294	31.691	13.929	15.377	20.364	20.658	20.842	22.481
10	24.289	29.678	17.782	17.347	29.660	21.380	30.913	26.174
15	27.524	37.111	24.894	28.948	34.019	30.091	36.057	34.911
20	24.675	31.353	28.589	20.565	36.720	28.938	38.655	30.364
				Panel (f): Number of Included Irrelevant Variables				
5	18.261	28.718	10.907	12.369	17.150	17.540	17.505	19.228
10	17.446	23.336	10.800	10.943	21.868	14.471	22.874	19.053
15	18.361	27.191	14.875	17.828	21.797	17.789	23.329	22.311
20	16.293	22.569	18.751	11.412	24.262	15.764	25.529	16.576

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 27. FORECASTING: DESCRIPTIVE STATISTICS
Simulation 2

The table reports for each different sample size, the one-step ahead mean squared error (MSE) for each model selection technique. n is the number of candidate linear main effects 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	0.278	0.281	0.263	0.264	0.253	0.255	0.252	0.252
10	0.322	0.317	0.279	0.279	0.259	0.260	0.255	0.253
15	0.377	0.363	0.295	0.296	0.263	0.264	0.257	0.257
20	0.447	0.440	0.315	0.315	0.267	0.268	0.260	0.259
				MSE - LASSO				
5	0.524	0.725	0.365	0.415	0.278	0.307	0.269	0.295
10	1.328	2.651	0.601	0.995	0.296	0.462	0.277	0.419
15	2.505	5.652	1.014	2.024	0.304	0.574	0.280	0.481
20	4.591	9.242	1.732	4.094	0.316	1.209	0.288	0.976
				MSE - adaLASSO				
5	0.480	0.687	0.316	0.348	0.269	0.271	0.262	0.262
10	1.077	2.306	0.421	0.631	0.288	0.291	0.272	0.271
15	1.933	5.242	0.647	1.284	0.293	0.300	0.273	0.273
20	3.546	8.675	0.999	2.578	0.309	0.441	0.283	0.374
				MSE - GS-LASSO				
5	1.044	1.232	0.437	0.481	0.313	0.341	0.285	0.295
10	0.986	2.746	0.531	0.948	0.341	0.595	0.297	0.522
15	1.948	4.184	0.994	1.299	0.517	0.507	0.434	0.424
20	4.992	14.350	2.128	4.557	0.967	1.882	0.818	1.630
				MSE - GS-adaLASSO				
5	1.254	1.508	0.442	0.486	0.301	0.326	0.277	0.283
10	1.272	3.479	0.546	0.927	0.320	0.589	0.283	0.511
15	2.495	5.034	1.019	1.314	0.499	0.489	0.423	0.406
20	6.173	11.541	2.218	4.813	0.957	1.840	0.810	1.597
				MSE - post GS-LASSO				
5	1.603	6.410	0.444	0.493	0.302	0.331	0.276	0.283
10	1.244	3.488	0.543	0.955	0.324	0.586	0.282	0.508
15	2.314	4.808	1.006	1.321	0.498	0.486	0.420	0.403
20	5.706	39.709	2.173	4.677	0.937	1.813	0.798	1.585
				MSE - post GS-adaLASSO				
5	1.199	1.540	0.440	0.485	0.300	0.326	0.277	0.283
10	1.232	3.378	0.543	0.923	0.320	0.589	0.283	0.511
15	2.418	5.170	1.012	1.314	0.499	0.489	0.423	0.406
20	5.681	11.670	2.198	4.811	0.956	1.840	0.809	1.597

3.3.3. Simulation 3

Simulation 3 is similar to simulation 2, but in (22) we also consider interactions of second order (DGP2), as presented in (24):

PUC-Rio - Certificação Digital Nº 0821483/CA

$$y_i = \sum_{k=1}^{q_1} \beta_k \, x_{k,i} + \sum_{jk=1}^{q_2} \alpha_{0,jk} G_{jk,i} + \sum_{jkl=1}^{q_3} \alpha_{1,jkl} G^1_{jkl,i} + \sum_{jkabl=1}^{q_4} \alpha_{2,jkabl} G^2_{jkabl,i} + 0.5 \, \varepsilon_i, \tag{24}$$

$$\varepsilon_i \sim \text{IN}[0,1]$$

$$\boldsymbol{x}_i \sim \text{IN}_{q_1}[0, \boldsymbol{I}_{q_1}]$$

$$i = 1, \ldots T$$

where, $\boldsymbol{\beta}, \boldsymbol{\alpha_0}$, $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ are vectors of ones of size q_1, q_2, q_3 and q_4 respectively. $\boldsymbol{x}_i$ is a vector of q_1 relevant linear main effects, $\boldsymbol{G}_i$ is a vector of q_2 relevant non-linear main effects, $\boldsymbol{G}_i^1$ is a vector of q_3 interactions of first order, as shown in (23), and $\boldsymbol{G}_i^2$ is a vector of q_4 interactions of second order, as shown in (25):

$$\boldsymbol{G}^2_{jkabl} = G\big(\mathbf{x}_k; \gamma, c_j\big) G(\mathbf{x}_b; \gamma, c_a) \mathbf{x}_l \tag{25}$$

where $G\big(\mathbf{x}_k; \gamma, c_j\big)$ is the logistic function according to (11).

The number of candidate variables now increases to $\Big(n + (n_c * n) + (n_c * n * n) + \big(n_c * n * n * \big((n_c * n) + 1\big)/2\big)\Big)$. When $n = 5$ the total of candidate variables is 1780, and when $n = 10$ this number goes to 13310 candidate variables. Since the Group Stepwise methodology embodies only interactions of first order, and as simulation 1 and 2 have shown that the direct methodology presents similar or better results, in simulation 3 we tested only the direct LASSO and adaLASSO applied to the STR model and compared the results with the oracle model. Model selection statistics results are shown in Tables 28 and 29. Table 30 presents the forecasting results.

With the inclusion of interactions of second order, we got worse results than in simulation 2. However, especially for the adaLASSO, the methodology gets satisfactory rates of including the relevant variables, and excluding the irrelevant variables (Panel (c) and (d)) when *T* increases. Also for adaLASSO, when *T*=300 and *T*=500, the one-step ahead MSE is close to the oracle.

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 28. MODEL SELECTION: DESCRIPTIVE STATISTICS
LASSO - DGP2 – Simulation 3

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 - GDP2								
	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	0.105	0.001	0.662	0.063	0.986	0.548	1	0.669
10	0	0	0.073	0	0.899	0.222	0.979	0.388
15	0	0	0	0	0	0.06	0	0.323
20	0	0	0	0	0	0.056	0	0.498
				Panel (c): Fraction of Relevant Variables Included				
5	0.647	0.251	0.917	0.623	0.997	0.907	1.000	0.933
10	0.436	0.153	0.790	0.460	0.990	0.890	0.998	0.927
15	0.171	0.065	0.365	0.228	0.796	0.856	0.866	0.934
20	0.200	0.047	0.427	0.159	0.846	0.857	0.916	0.968
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.991	0.997	0.991	0.997	0.993	0.998	0.994	0.998
10	0.986	0.997	0.984	0.996	0.986	0.996	0.987	0.996
15	0.984	0.997	0.978	0.995	0.975	0.994	0.975	0.995
20	0.984	0.997	0.978	0.995	0.972	0.993	0.972	0.994
				Panel (e): Number of Included Variables				
5	19.764	36.914	19.733	41.632	16.602	35.326	15.228	33.012
10	28.897	39.324	35.921	58.182	35.173	63.703	32.867	57.692
15	30.525	40.961	43.444	67.576	56.482	94.357	56.604	86.192
20	32.508	40.865	47.854	67.597	65.6	106.304	66.951	96.703
				Panel (f): Number of Included Irrelevant Variables				
5	16.531	35.66	15.146	38.515	11.616	30.789	10.228	28.347
10	24.541	37.799	28.025	53.586	25.277	54.806	22.888	48.425
15	27.959	39.993	37.962	64.152	44.541	81.511	43.619	72.189
20	28.505	39.918	39.306	64.423	48.673	89.158	48.634	77.339

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 29. MODEL SELECTION: DESCRIPTIVE STATISTICS
adaLASSO - DGP2 – Simulation 3

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 - GP2								
	T=50		*T*=100		*T*=300		*T*=500	
*q**n*	5	10	5	10	5	10	5	10
				Panel (a): Correct Sparsity Pattern				
5	0.002	0	0.002	0	0.037	0.003	0.116	0.003
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	0.032	0	0.174	0.054	0.358	0.533	0.452	0.642
10	0	0	0.013	0	0.389	0.103	0.555	0.17
15	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0.009
				Panel (c): Fraction of Relevant Variables Included				
5	0.515	0.186	0.741	0.583	0.841	0.896	0.875	0.924
10	0.324	0.120	0.665	0.393	0.919	0.858	0.947	0.888
15	0.122	0.045	0.269	0.183	0.614	0.731	0.713	0.793
20	0.153	0.032	0.333	0.120	0.725	0.739	0.822	0.858
				Panel (d): Fraction of Irrelevant Variables Excluded				
5	0.994	0.998	0.996	0.998	0.998	0.999	0.999	1
10	0.991	0.998	0.991	0.997	0.994	0.998	0.995	0.999
15	0.989	0.998	0.986	0.997	0.985	0.997	0.986	0.998
20	0.989	0.998	0.985	0.996	0.984	0.996	0.985	0.997
				Panel (e): Number of Included Variables				
5	12.924	26.568	11.048	25.264	7.957	13.508	6.892	11.725
10	19.580	28.463	23.337	41.767	19.545	34.195	17.501	27.896
15	20.602	29.775	28.760	49.022	35.590	54.920	35.439	44.880
20	22.266	29.611	32.591	50.376	42.797	70.325	42.828	57.075
				Panel (f): Number of Included Irrelevant Variables				
5	10.350	25.640	7.342	22.351	3.750	9.030	2.517	7.105
10	16.340	27.267	16.690	37.838	10.356	25.613	8.029	19.019
15	18.770	29.096	24.724	46.279	26.381	43.949	24.746	32.992
20	19.202	28.962	25.930	47.970	28.289	55.543	26.383	39.923

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 30. FORECASTING: DESCRIPTIVE STATISTICS
Simulation 3

The table reports for each different sample size, the one-step ahead mean squared error (MSE) for each model selection technique. *n* is the number of candidate linear main effects 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	0.280	0.281	0.262	0.263	0.255	0.256	0.252	0.253
10	0.321	0.327	0.280	0.279	0.258	0.260	0.253	0.255
15	0.387	0.373	0.299	0.297	0.263	0.263	0.256	0.259
20	0.514	0.451	0.322	0.316	0.271	0.268	0.261	0.259
				MSE - LASSO - DGP2				
5	0.909	8.284	0.425	0.723	0.323	0.358	0.313	0.310
10	25.877	8.375	0.963	1.974	0.432	0.491	0.383	0.386
15	12.449	37.408	2.791	4.477	1.070	0.837	0.842	0.553
20	135.398	105.996	4.190	8.126	1.292	1.613	0.952	0.931
				MSE - adaLASSO - DGP2				
5	1.119	7.647	0.359	0.815	0.267	0.292	0.258	0.270
10	57.411	7.524	0.700	1.718	0.284	0.334	0.264	0.283
15	13.665	236.835	2.252	3.978	0.581	0.411	0.397	0.304
20	144.451	229.679	3.200	7.216	0.585	0.620	0.385	0.321

PUC-Rio - Certificação Digital Nº 0821483/CA

3.3.4. Simulation 4

In simulation 4 we contemplate the case where we don't know the structure of the true non-linear model, but we know the linear main effects. In other words, the relevant variables are not available in the candidate variables set. Our goal is to test if the methodology can identify some non-linear effects capable to explain the dependent variable.

As DGP, we use a wind-power forecasting model known to be highly non-linear, and we want to test if the STR model generated by the linear main effects can identify linear and non-linear effects that can explain the response. The DGP3 for the fourth simulation is expressed in equation (26), where v represents the wind speed in m/s and θ represents the wind direction in degrees.

$$\begin{gathered} y_i = \beta_o + \beta_1(v_i \cos\theta_i)^3 + \beta_2(v_i \sin\theta_i)^3 + 0.5\,\varepsilon_i,\ \ \varepsilon_i \sim \text{IN}[0,1] \\ v_i \sim U[0,10] \\ \theta_i \sim U[0,360] \\ i = 1, \dots T \end{gathered} \tag{26}$$

As candidate variables we consider v and θ as linear main effects $(n = 2)$, the $(n_c * n)$ non-linear main effects, the $(n_c * n * n)$ interactions of first order,

and the $\left(n_c * n * n * \left((n_c * n) + 1\right)/2\right)$ interactions of second order, which gives a total of 142 candidate variables. As we consider interactions of second order in the set of candidate variables, we only apply the direct methodology in this simulation.

Table 31 shows the MSE and the R^2 out-of-sample for T = 50, 100, 300, 500, for the LASSO, adaLASSO and oracle models. As expected, when the sample size increases the forecasting results improve, and when T = 500, the MSE and R^2 are really close to the oracle. This shows that the methodology identifies some linear and non-linear effects that explain great part of the dependent variable.

Table 32 shows the variables that were included in more that 90% of the 1000 replications for each model and sample size. The scenarios that do not appear in the table did not include any variable in more than 90% of the replications. In Table 32, $\mathrm{g}_j(x_1)$ represents the logistic function as in eq. (27):

$$\mathrm{g}_j(x_1) = G\left(x_{1,i}; \gamma, c_j\right) = \frac{1}{1 + e^{-\gamma(x_{1,i} - c_j)}} \tag{27}$$

Table 33 shows the mean of variables included in each scenario per group of variables. X1 consists in the linear main effects, X2 is formed by the non-linear main effects, X3 is composed by the interactions of first order, and X4 by the interactions of second order.

TABLE 31. FORECASTING: DESCRIPTIVE STATISTICS
Simulation 4

The table reports the one-step ahead mean squared error (MSE) and R^2 for each model selection technique.

	T=50		*T*=100		*T*=300		*T*=500	
Model	MSE	R2	MSE	R2	MSE	R2	MSE	R2
LASSO	1.085	59%	0.723	74%	0.392	86%	0.359	87%
adaLASSO	1.091	56%	0.665	76%	0.380	86%	0.346	87%
Oracle	0.263	90%	0.256	91%	0.253	91%	0.251	91%

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 32. RELEVANT VARIABLES – 90%
Simulation 4

The table shows the relevant variables (regressors) per model, the number of models in which each variable appears, the number of parameters of each model, and the variables description.

Variables that appear more than 90%							
Variable	LASSO T=500	adaLASSO T=500	LASSO T=300	adaLASSO T=300	LASSO T=100	**num models**	Description
3	X	X	X	X		**4**	g1(v)
11	X					**1**	g5(v)
19	X	X				**2**	g2(θ)*v
35	X					**1**	g1(v)*g1(θ)*v
39	X	X	X			**3**	g1(v)*g2(θ)*v
57	X		X			**2**	g1(θ)*g2(θ)*v
97	X	X	X	X		**4**	g2(θ)*g5(v)*v
138					X	**1**	g5(v)*g5(v)*θ
141	X	X				**2**	g5(θ)*g5(θ)*v
num var	**8**	**5**	**4**	**2**	**1**		

TABLE 33. RELEVANT VARIABLES
Simulation 4

The table shows the mean of variables included per group of variables for LASSO and adaLASSO, and the total of candidate variables per group.

	num.	*T*=50		*T*=100		*T*=300		*T*=500	
Group var	candidates	LASSO	adaLASSO	LASSO	adaLASSO	LASSO	adaLASSO	LASSO	adaLASSO
X1	2	0.256	0.201	0.410	0.254	0.622	0.522	0.694	0.692
X2	10	2.233	2.086	3.152	2.597	5.016	4.143	5.071	4.461
X3	20	1.120	1.334	1.329	1.321	3.060	2.728	3.559	3.371
X4	110	5.237	3.705	6.062	4.224	15.726	11.704	16.785	13.339
Total	**142**	**8.846**	**7.326**	**10.953**	**8.396**	**24.424**	**19.097**	**26.109**	**21.863**

3.3.5. Simulation 5

Simulation 5 is similar to simulation 1, but now we want to test the performance of both variable selection methodologies (LASSO and GS-LASSO) for a larger set of linear main effects candidates. We consider 50 linear main effects ($n = 50$), $(n_c * n)$ non-linear main effects and $(n_c * n * n)$ interactions of first order, which gives a total of 12800 candidate variables. Since we want to compare LASSO and Group Stepwise LASSO, and the latter only takes into account interactions of first order, we do not consider interactions of second order.

Results for model selection statistics for the LASSO and GS-LASSO are presented in Table 34. Table 35 shows the model selection statistics for the adaLASSO and GS-adaLASSO. The forecasting results are presented in Table 36.

For selection statistics, LASSO and GS-LASSO presented similar results, excepting for Panel (e) and (f), where the LASSO includes much more variables than the GS-LASSO. The adaLASSO presented slight better results than the GS-adaLASSO. In forecast accuracy, adaLASSO showed the best results, especially when T = 300 and 500, where the MSE is really close to the oracle.

PUC-Rio - Certificação Digital Nº 0821483/CA

Table 37 shows the computation time, or CPU time, of variable selection and model estimation using the direct methodology and the Group Stepwise algorithm. The time corresponds to 1000 Monte Carlo replications of simulation 5 on an Intel Core i7-3960X, 3.3 GHz with 64 GB of RAM using glmnet package, from Friedman et al. (2010), on Matlab R2011b.

There is a significant difference of computation time of both methodologies. In some cases, Group Stepwise LASSO can take only 15% of the time of direct LASSO. So, as selection and forecasting results are quite similar for both methodologies, it can be better to use the Group Stepwise LASSO for STR models.

TABLE 34. MODEL SELECTION: DESCRIPTIVE STATISTICS
LASSO and GS-LASSO – Simulation 5

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				GS-LASSO			
	T=50	*T*=100	*T*=300	*T*=500	*T*=50	*T*=100	*T*=300	*T*=500
*q**n*	50	50	50	50	50	50	50	50
	Panel (a): Correct Sparsity Pattern				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				Panel (b): True Model Included			
5	0	0.095	0.993	1	0	0.293	0.992	1
10	0.000	0	0.516	0.995	0	0.004	0.007	0.006
15	0	0	0.006	0.137	0	0	0.118	0.450
20	0	0	0	0.013	0	0	0.014	0.156
	Panel (c): Fraction of Relevant Variables Included				Panel (c): Fraction of Relevant Variables Included			
5	0.529	0.703	0.999	1	0.400	0.578	0.995	1
10	0.153	0.479	0.935	1	0.496	0.661	0.801	0.801
15	0.083	0.302	0.775	0.940	0.461	0.555	0.797	0.911
20	0.053	0.175	0.660	0.892	0.344	0.436	0.597	0.843
	Panel (d): Fraction of Irrelevant Variables Excluded				Panel (d): Fraction of Irrelevant Variables Excluded			
5	0.998	0.998	0.998	0.999	0.996	1	1	1
10	0.997	0.995	0.995	0.996	0.997	0.999	0.999	1
15	0.997	0.995	0.994	0.996	0.997	0.997	0.999	0.999
20	0.997	0.996	0.993	0.994	0.997	0.996	0.999	0.998
	Panel (e): Number of Included Variables				Panel (e): Number of Included Variables			
5	34.333	33.410	25.163	22.617	49.634	7.783	8.642	8.133
10	42.754	65.450	75.132	61.082	49.493	24.331	15.170	14.352
15	41.731	65.899	90.870	70.770	48.911	42.659	26.534	30.499
20	41.912	60.998	101.748	91.819	49.110	64.873	30.848	42.616
	Panel (f): Number of Included Irrelevant Variables				Panel (f): Number of Included Irrelevant Variables			
5	31.689	29.893	20.170	17.617	47.636	4.892	3.666	3.133
10	41.228	60.658	65.785	51.087	44.536	17.723	7.158	6.340
15	40.487	61.364	79.241	56.666	41.993	34.338	14.575	16.841
20	40.845	57.490	88.545	73.983	42.237	56.162	18.906	25.755

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 35. MODEL SELECTION: DESCRIPTIVE STATISTICS
adaLASSO and GS-adaLASSO – Simulation 5

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				GS-adaLASSO			
	T=50	*T*=100	*T*=300	*T*=500	*T*=50	*T*=100	*T*=300	*T*=500
*q**n*	50	50	50	50	50	50	50	50
	Panel (a): Correct Sparsity Pattern				Panel (a): Correct Sparsity Pattern			
5	0	0.004	0.108	0.228	0	0	0	0
10	0	0	0	0.001	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				Panel (b): True Model Included			
5	0	0.103	1	1	0	0.347	0.994	1
10	0	0	0.680	0.997	0	0.005	0.003	0.006
15	0	0	0	0	0	0	0.115	0.423
20	0	0	0	0	0	0	0.009	0.145
	Panel (c): Fraction of Relevant Variables Included				Panel (c): Fraction of Relevant Variables Included			
5	0.432	0.662	1	1	0.398	0.613	0.996	1
10	0.114	0.434	0.959	1	0.492	0.674	0.800	0.801
15	0.061	0.266	0.727	0.928	0.458	0.565	0.796	0.906
20	0.041	0.167	0.611	0.861	0.340	0.445	0.611	0.843
	Panel (d): Fraction of Irrelevant Variables Excluded				Panel (d): Fraction of Irrelevant Variables Excluded			
5	0.998	0.999	1	1	0.996	1	1	1
10	0.997	0.996	0.998	0.999	0.997	0.998	1	1
15	0.998	0.996	0.997	0.999	0.997	0.997	0.999	0.999
20	0.997	0.996	0.995	0.998	0.997	0.995	0.999	0.998
	Panel (e): Number of Included Variables				Panel (e): Number of Included Variables			
5	25.156	19.967	7.937	6.900	48.692	8.218	7.936	7.434
10	33.369	50.309	29.249	19.969	48.583	26.185	14.056	13.349
15	32.402	51.724	54.843	26.288	48.177	44.246	25.165	28.583
20	32.815	55.367	73.815	47.564	48.401	69.491	30.581	40.929
	Panel (f): Number of Included Irrelevant Variables				Panel (f): Number of Included Irrelevant Variables			
5	22.997	16.655	2.937	1.900	46.703	5.154	2.954	2.434
10	32.232	45.969	19.660	9.972	43.661	19.441	6.052	5.337
15	31.493	47.741	43.931	12.373	41.302	35.774	13.225	14.991
20	32.003	52.024	61.592	30.349	41.592	60.599	18.370	24.060

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 36. FORECASTING: DESCRIPTIVE STATISTICS
Simulation 5

The table reports for each different sample size, the one-step ahead mean squared error (MSE) for each model selection technique. n is the number of candidate linear main effects whereas q is the number of relevant regressors.

	T=50	T=100	T=300	T=500
q/n	50	50	50	50
		MSE - Oracle		
5	0.278	0.264	0.253	0.253
10	0.321	0.279	0.259	0.255
15	0.366	0.297	0.266	0.257
20	0.433	0.315	0.266	0.260
		MSE - LASSO		
5	1.244	0.552	0.345	0.307
10	6.645	1.839	0.558	0.412
15	11.567	5.007	1.169	0.760
20	18.031	11.501	2.284	1.434
		MSE - adaLASSO		
5	1.274	0.459	0.272	0.260
10	6.943	1.474	0.325	0.281
15	12.224	4.051	0.496	0.293
20	19.334	9.564	0.794	0.357
		MSE - GS-LASSO		
5	43.615	1.329	0.276	0.258
10	69.396	2.631	0.945	0.922
15	70.091	4.219	0.917	0.454
20	152.819	10.163	2.887	0.911
		MSE - GS-adaLASSO		
5	41.626	1.242	0.268	0.255
10	72.777	2.621	0.937	0.914
15	69.649	4.234	0.868	0.438
20	139.263	10.577	2.622	0.852
		MSE - post GS-LASSO		
5	5.06E+03	1.325	0.272	0.256
10	1.52E+06	2.649	0.930	0.913
15	9.25E+03	5.687	0.874	0.429
20	2.72E+04	41.343	2.765	0.847
		MSE - post GS-adaLASSO		
5	2.67E+03	1.245	0.268	0.255
10	4.65E+04	2.627	0.937	0.914
15	7.28E+03	4.375	0.868	0.438
20	2.19E+04	10.985	2.621	0.852

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 37. COMPUTATION TIME
Simulation 5

The table reports for each different sample size, for each q number of relevant regressors, and 50 candidate linear main effects, the computational time of variable selection and parameter estimation for each model selection technique, in seconds and minutes, and the fraction.

	T=50		T=100		T=300		T=500	
q/time	sec	min	sec	min	sec	min	sec	min
				Time - LASSO and adaLASSO				
5	924.72	15.41	1168.62	19.48	1463.84	24.40	1940.04	32.33
10	1060.38	17.67	1822.10	30.37	2725.42	45.42	2970.25	49.50
15	1041.33	17.36	1782.65	29.71	3323.74	55.40	3257.33	54.29
20	1066.63	17.78	1858.60	30.98	4047.85	67.46	4404.48	73.41
				Time - GS-LASSO and GS-adaLASSO				
5	279.12	4.65	228.36	3.81	799.74	13.33	758.60	12.64
10	270.64	4.51	265.00	4.42	803.03	13.38	773.96	12.90
15	234.74	3.91	313.14	5.22	720.70	12.01	783.66	13.06
20	253.00	4.22	419.19	6.99	847.09	14.12	881.92	14.70
				Time fraction				
5	30%	30%	20%	20%	55%	55%	39%	39%
10	26%	26%	15%	15%	29%	29%	26%	26%
15	23%	23%	18%	18%	22%	22%	24%	24%
20	24%	24%	23%	23%	21%	21%	20%	20%

3.4. STAR – LASSO

In this section we extend the methodology for variable selection for STR models to STAR (Smooth Transition AutoRegressive) models. Following the same idea in Section 3.2, we define the STAR model in (28) with p linear autoregressive variables as linear main effects (lags of y_t), $(n_c * p)$ non-linear main effects and $(n_c * p * p)$ interactions of first order:

$$y_t = \sum_{k=1}^{p} \beta_k \, y_{t-k} + \sum_{j=1}^{n_c} \sum_{k=1}^{p} \alpha_{0,j,k} \, G\left(y_{t-k}; \gamma, c_j\right) + \sum_{j=1}^{n_c} \sum_{k=1}^{p} \sum_{l=1}^{p} \alpha_{1,j,k,l} G\left(y_{t-k}; \gamma, c_j\right) y_{t-l} + \varepsilon_t, \tag{28}$$

$$\varepsilon_t \sim \text{IN}[0, \sigma^2]$$

$$t = 1, \dots T$$

where

$$G\left(y_{t-k}; \gamma, c_j\right) = \frac{1}{1 + e^{-\gamma (y_{t-k} - c_j)}} \tag{29}$$

PUC-Rio - Certificação Digital Nº 0821483/CA

is the logistic function, with the slope parameter γ, the location parameter $c_j, j = 1, \dots n_c$, and the transition variable y_{t-k}. To simplify, we do not consider external variables in this model.

As in Section 3.2, our general approach for fitting (28) is to minimize the following penalized regression criterion in equation (15), where now

$G_k = \beta_k \mathbf{y}_{t-k}$ → linear main effects

$G_{jk} = \alpha_{0,j,k} G(\mathbf{y}_{t-k}; \gamma, c_j)$ → non-linear main effects

$G_{jkl} = \alpha_{1,j,k,l} G(\mathbf{y}_{t-k}; \gamma, c_j)\mathbf{y}_{t-l}$ → interactions of first order

$G_{jkabl} = 0$ → interactions of second order

The penalty function $P(\mathbf{G})$ is given by (16) for LASSO estimator and $P'(\mathbf{G})$ is given by (17), for adaLASSO estimator. The Group Stepwise LASSO for the STAR model follows analogous to Section 3.2.1.

PUC-Rio - Certificação Digital Nº 0821483/CA

3.4.1. Simulation 6

In order to test LASSO and GS-LASSO for STAR models, we evaluate one simulation exercise. This simulation is similar to simulation 1, presented in Section 3.3.1, as we can see in Table 38.

TABLE 38. SIMULATION EXERCISE: SCENARIO

The table reports the scenario for the simulation exercise, concerning the set of candidate variables, the structure for the DGP, the methodology applied, and the availability of the true model variables.

Simulation	Candidate Variables				Structure for the DGP		Methodology		True model
	X1	X2	X3	X4	heredity	non-heredity	LASSO	GS-LASSO	not available
6	✓	✓	✓		✓		✓	✓	

As candidate linear main effects we consider $n = 6$ and $n = 12$, representing the 6 and 12 firsts lags of the response, respectively, and we consider 5 relevant variables ($q = 5$). As in simulation 1, we considered $n_c = 5$, the percentiles of 20%, 35%, 50%, 65% and 80% of the transition variable for the values of c_j, and the slope parameter $\gamma = 10$.

The data generating process (DGP) is given by (30):

$$\begin{gathered} y_t = 0.7\, y_{t-1} - 0.3\, y_{t-3} + 0.5\, \mathrm{g}_3(y_{t-2}) + 0.4\, \mathrm{g}_3(y_{t-2}) y_{t-1} - \\ 0.2\, \mathrm{g}_3(y_{t-2}) y_{t-3} + 0.5\, \varepsilon_t, \\ \varepsilon_t \sim \mathrm{IN}[0,1] \\ t = 1, \ldots T \end{gathered} \tag{30}$$

where $\mathrm{g}_3(y_{t-2})$ is the logistic function using the percentile of 50% of the error as c_j in (27). To generate non-explosive series, the coefficients were chosen such that the absolute values of the eigenvalues of the characteristic polynomial were less than 1. In order to find the eigenvalues we need to take into account the extreme regions of $\mathrm{g}_3(y_{t-2})$, i.e. when $\mathrm{g}_3(y_{t-2}) = 0$ and $\mathrm{g}_3(y_{t-2}) = 1$.

We simulate 1000 Monte Carlo replications of T = 50, 100, 300, 500 observations of DGP (30). When $n = 6$ we have a total of 216 candidate variables, and with $n = 12$, this number increases to 792.

Variable selection statistics results and forecasting results for both methodologies applied to the STAR model are presented in Table 39 and 40, respectively.

For the STAR model, the GS-LASSO and GS-adaLASSO present the best results when it comes to variable selection and forecast accuracy. The out-of-sample MSE is close to the oracle and variable selection succeeds in selecting a parsimonious model.

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 39. MODEL SELECTION: DESCRIPTIVE STATISTICS
Simulation 6

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.

	T=50		*T*=100		*T*=300		*T*=500	
Model \ *n*	5	10	5	10	5	10	5	10
				Panel (a): Correct Sparsity Pattern				
LASSO	0	0	0	0	0	0	0	0
adaLASSO	0	0	0	0	0	0	0	0
GS-LASSO	0	0	0.001	0	0.010	0.002	0.014	0.005
GS-adaLASSO	0	0	0	0.001	0.008	0.004	0.012	0.015
				Panel (b): True Model Included				
LASSO	0	0	0	0	0.043	0.001	0.101	0.006
adaLASSO	0	0	0	0	0.011	0	0.031	0
GS-LASSO	0	0.001	0.004	0.002	0.072	0.057	0.194	0.185
GS-adaLASSO	0.001	0.001	0.005	0.003	0.067	0.050	0.189	0.230
				Panel (c): Fraction of Relevant Variables Included				
LASSO	0.223	0.166	0.306	0.271	0.550	0.367	0.652	0.421
adaLASSO	0.191	0.155	0.232	0.207	0.425	0	0.561	0.287
GS-LASSO	0.376	0.354	0.417	0.398	0.529	0.520	0.671	0.659
GS-adaLASSO	0.366	0.355	0.421	0.395	0.534	0.522	0.667	0.669
				Panel (d): Fraction of Irrelevant Variables Excluded				
LASSO	0.968	0.986	0.965	0.988	0.948	0.985	0.944	0.982
adaLASSO	0.980	0.990	0.980	0.993	0.970	0.992	0.966	0.992
GS-LASSO	0.988	0.992	0.991	0.996	0.989	0.996	0.983	0.995
GS-adaLASSO	0.990	0.991	0.992	0.997	0.989	0.997	0.984	0.996
				Panel (e): Number of Included Variables				
LASSO	7.948	11.681	8.848	10.808	13.801	13.999	15.114	15.959
adaLASSO	5.074	8.737	5.315	6.413	8.457	7.285	10.014	7.881
GS-LASSO	4.338	7.823	3.920	4.805	5.056	5.844	6.869	7.291
GS-adaLASSO	3.988	8.703	3.786	4.144	4.927	5.192	6.663	6.513
				Panel (f): Number of Included Irrelevant Variables				
LASSO	6.832	10.849	7.318	9.451	11.052	12.165	11.855	13.856
adaLASSO	4.121	7.961	4.157	5.380	6.333	6.009	7.207	6.446
GS-LASSO	2.456	6.053	1.833	2.813	2.409	3.244	3.515	3.997
GS-adaLASSO	2.160	6.926	1.682	2.167	2.256	2.581	3.327	3.167

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 40. FORECASTING: DESCRIPTIVE STATISTICS
Simulation 6

The table reports for each different sample size, the one-step ahead mean squared error (MSE) for each model selection technique. *n* is the number of candidate linear main effects.

MSE out-of-sample								
	T=50		*T*=100		*T*=300		*T*=500	
Model*n*	5	10	5	10	5	10	5	10
Oracle	0.563	0.320	0.264	0.269	0.255	0.253	0.252	0.253
LASSO	0.418	0.546	0.345	0.400	0.282	0.311	0.267	0.295
adaLASSO	0.405	0.634	0.322	0.370	0.278	0.294	0.264	0.283
GS-LASSO	0.346	0.445	0.295	0.308	0.274	0.278	0.263	0.267
GS-adaLASSO	0.354	0.519	0.296	0.307	0.272	0.277	0.261	0.265
post GS-LASSO	3.764	0.586	0.311	0.320	0.285	0.288	0.267	0.272
post GS-adaLASSO	0.355	0.561	0.310	0.320	0.281	0.286	0.264	0.268

3.5. Application to genetic data

With recent evolution on genetics sciences the set of genetic information available is increasing every day, and the study of high-dimensional statistical models becomes an important matter. In this scenario, variable selection is an important subject, and some methodologies have already been applied to genetic data. See Tian et al. (2012) and Tian and Suárez-Fariñas (2013) for some examples of regularization in genomic data.

3.5.1. The data

In this application we use the data for the largest study in psoriasis, published by Suárez-Fariñas et al. (2012)[8]. Psoriasis is a common chronic inflammatory skin disease, which the cause is not fully understood.

The database includes 85 patients with several clinical variables, and 54675 genomic expressions for each patient. The goal here is to select the genes that are correlated with psoriasis, i.e., the relevant genes to explain de disease, being able to predict the severity of the disease for each patient out-of-sample.

We used 80% of the data set (68 observations) for the in-sample model

[8] We want to thank Mayte Suárez-Fariñas from the Laboratory of Investigative Dermatology, Rockefeller University, New York, USA, for providing the data and all the help with the genetic language.

PUC-Rio - Certificação Digital Nº 0821483/CA

specification and 20% (17 observations) for the out-of-sample forecast. The dependent variable used in the fitted models was the clinical index of disease (disease severity) and the explanatory variables are the 54675 genomic expressions. The data set is composed by quantitative variables.

3.5.2. The models

We want to use genomic data to explain the psoriasis. This is a high dimensional problem, where we have 85 independent observations and 54675 explanatory variables. This would be an impossible model to fit and select the relevant variables using classical methods based on statistical tests and information criteria. The *Autometrics* would also spend too much computational time, with all the paths and tests in the algorithm. Therefore, we propose to use the LASSO and adaLASSO for linear and STR models. The STR model will allow us to identify some possible non-linear effects on the data.

First, we apply original LASSO and adaLASSO approach to fit a linear regression with all genes as candidate variables. In a second step, we use the variables selected in the LASSO and adaLASSO as linear main effects in the STR model. The set of candidate variables for the STR model is composed by the main linear and non-linear effects, and interactions of first order, similarly to (14). For the STR model we apply the direct LASSO and adaLASSO, and the Group Stepwise LASSO and adaLASSO (GS-LASSO and GS-adaLASSO), according to the methodology presented in Section 3.2.

We evaluate 1000 permutation on the data observations, creating 1000 different in-sample and out-of-sample sets. The results presented next are the average statistics of the 1000 fitted models.

3.5.3. Results

We compare 6 models: LASSO and adaLASSO, for the linear regression (LASSO – linear and adaLASSO – linear); LASSO and adaLASSO in the STR model (LASSO – STR and adaLASSO – STR); and Group Stepwise LASSO and adaLASSO in the STR model (GS-LASSO – STR and GS-adaLASSO – STR).

PUC-Rio - Certificação Digital Nº 0821483/CA

The STR models were specified for 2 sets of linear main effects: first, using the selected variables in LASSO, and second, using the selected variables in adaLASSO. Table 41 presents the in-sample and out-of-sample average statistics for the 1000 permutations on the observations.

TABLE 41. PSORIASIS 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 R2, the out-of-sample mean squared error (MSE), the Bayesian Information Criterion (BIC), and the number of parameters, for each model selection technique, and benchmark models.

Model	R2_in	R2_out	MSE_out	BIC	num par
		using selected main effects by LASSO			
LASSO - linear	0.981	0.273	80.059	3.681	51.991
adaLASSO - linear	0.999	0.238	82.329	0.788	45.704
LASSO - STR	0.987	0.127	94.929	3.532	54.455
adaLASSO - STR	0.997	0.101	96.960	0.556	40.158
GS-LASSO - STR	0.997	0.264	80.107	0.402	49.773
GS-adaLASSO - STR	0.997	0.265	80.109	0.318	48.914
		using selected main effects by adaLASSO			
LASSO - linear	0.982	0.270	82.089	3.671	52.414
adaLASSO - linear	0.999	0.242	83.934	0.771	45.880
LASSO - STR	0.986	0.110	98.525	3.521	53.400
adaLASSO - STR	0.999	0.086	100.648	0.575	39.904
GS-LASSO - STR	0.999	0.241	83.955	0.347	45.658
GS-adaLASSO - STR	0.999	0.240	84.025	0.271	45.103

Table 41 shows that models selected by GS-LASSO – STR and GS-adaLASSO – STR present better in-sample results than the others. However, for predictive power (out-of-sample forecast), these models present similar results than LASSO – linear and adaLASSO – linear. We tested the predictive accuracy of all models against the others in relation to the absolute and squared error using the modify Diebold and Mariano test, proposed by Harvey et al. (1997).

The test statistics are presented in Table 42. The critical value at a 0.05 significant level, for a sample of size 17, is 1.75. Therefore, the test shows that none of the model presents out-of-sample absolute or squared errors significantly lower than the others. We can say that all models are equivalent concerning predictive power.

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 42. TEST OF PREDICTIVE ACCURACY

The table reports the modify Diebold and Mariano test statistic for all models, for absolute error and squared error. Models in columns are compared with models in rows (reference). The statistics are presented for the cases where the main effects of the STR models were determined by the LASSO –linear and adaLASSO – linear.

	adaLASSO - linear	LASSO - STR	adaLASSO - STR	GS-LASSO - STR	GS-adaLASSO - STR
			Using selected main effects by LASSO		
			absolute error		
LASSO - linear	-0.090	-0.641	-0.649	0.216	0.209
adaLASSO - linear	-	-0.405	-0.515	0.199	0.196
LASSO - STR	-	-	-0.083	0.648	0.643
adaLASSO - STR	-	-	-	0.656	0.650
GS-LASSO - STR	-	-	-	-	-0.014
			squared error		
LASSO - linear	-0.266	-0.672	-0.765	0.017	0.011
adaLASSO - linear	-	-0.622	-0.493	0.213	0.217
LASSO - STR	-	-	-0.294	0.587	0.580
adaLASSO - STR	-	-	-	0.675	0.668
GS-LASSO - STR	-	-	-	-	0.005
			Using selected main effects by adaLASSO		
			absolute error		
LASSO - linear	0.010	-0.687	-0.658	-0.005	-0.025
adaLASSO - linear	-	-0.144	-0.586	0.049	0.022
LASSO - STR	-	-	-0.094	0.579	0.567
adaLASSO - STR	-	-	-	0.579	0.566
GS-LASSO - STR	-	-	-	0.000	-0.047
			squared error		
LASSO - linear	-0.209	-0.730	-0.787	-0.208	-0.230
adaLASSO - linear	-	0.010	-0.578	0.028	0.009
LASSO - STR	-	-	-0.325	0.475	0.462
adaLASSO - STR	-	-	-	0.567	0.555
GS-LASSO - STR	-	-	-	-	-0.025

Analyzing the LASSO and GS-LASSO for STR models, we evaluate the number of selected genes. Table 43 shows the average total of selected variables in each group of variables, where X1 represents the linear main effects, X2 represents the non-linear main effects, and X3 is composed by the interactions of first order.

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 43. RELEVANT VARIABLES

The table shows the average total of variables included per group of variables for LASSO and adaLASSO, and the average total of candidate variables per group.

	num.	STR		GS - STR	
Group var	candidates	LASSO	adaLASSO	LASSO	adaLASSO
		using selected main effects by LASSO			
X1	51.991	0.035	0.024	47.960	47.070
X2	259.955	1.657	1.392	1.697	1.745
X3	13828.585	52.763	38.742	0.116	0.099
Total	**14140.531**	**54.455**	**40.158**	**49.773**	**48.914**
		using selected main effects by adaLASSO			
X1	45.880	0.033	0.023	43.394	42.710
X2	229.400	1.756	1.553	2.182	2.330
X3	10593.490	51.611	38.328	0.082	0.063
Total	**10868.770**	**53.400**	**39.904**	**45.658**	**45.103**

Analyzing Table 43, we notice that when applying direct methodology for LASSO and adaLASSO in STR models, most of selected genes are interactions of first order (X3), but when using the Group Stepwise LASSO and adaLASSO, the selected genes are mostly from the linear main set of variables (X1). This result gives an indication that most of the relevant information for the psoriasis is in the linear main effects, and when they are fixed in the model, as in the case of Group Stepwise LASSO, the non-linear effects are not significant, and therefore, not selected.

Table 44 presents the genes included in most than 70% of the permuted linear regression models. This analysis is of extreme importance for the interpretability of the model. If these genes are present in 70% of the models, we have an indication that they have a strong correlation to psoriasis severity, even with such a small in-sample set. The last column of Table 44 shows the selected genes and a brief description. The proportion of 70% was chosen randomly, the analysis could be also based in 80% or 90% of the models, for example.

PUC-Rio - Certificação Digital Nº 0821483/CA

TABLE 44. RELEVANT VARIABLES – 70%

The table shows the relevant variables (regressors) in LASSO and adaLASSO for linear regression that appear in more than 70% of the models, the number of parameters of each model, and the genes description.

Variables that appear in more than 70% of linear regression models			
Genomic expression	LASSO	adaLASSO	Related to
'203395_s_at'	X	X	immunological diseases
'203421_at'	X	X	cancer
'205554_s_at'	X	X	connectivity in tissues disorder
'214708_at'	X	X	inflammatory diseases
'1561336_at'	X		connectivity in tissues disorder
num var	**5**	**4**	

For out-of-sample forecast the models did not present satisfactory results in any of the applied methodologies. Our number of observation is relatively small, and we have a high dimensional statistic model to fit ($n >> p$). Considering this, all methods presented satisfactory in-sample results and we manage to select 5 relevant genes out of a set of 54675 genes.

3.6. Conclusion

The model selection methodology presented in this chapter, allows simultaneous estimation of the STR/STAR model and variable selection. The advantage of the logistic function is that it can approximate the true non-linear model, as shown in simulation 4, in Section 3.3.4.

Besides a good fit of the selected models, an important advantage of the presented methodologies is the interpretability. The variable selection of linear and non-linear main effects, and interaction terms, allows the user to interpret and understand non-linear effects in the true model.

We showed by simulations that the methodologies consistently choose the relevant variables as the number of observations increases. The Group Stepwise LASSO has the advantage of spending significantly less computational time than the direct approach. An application to genetic data was evaluated comparing the methodology to linear LASSO. For out-of-sample forecasting purpose, the results were not conclusive due to the small number of observations on the data.

PUC-Rio - Certificação Digital Nº 0821483/CA