

A Sparsity-Controlled Vector Autoregressive Model

Emilio Carrizosa^a, Alba V. Olivares-Nadal^{1, a} and Pepa Ramírez-Cobo^b
^a*Instituto de Matemáticas de la Universidad de Sevilla (Spain),*
^b*Department of Statistics and Operational Research, Universidad de Cádiz (Spain),*

Abstract

Vector autoregressive (VAR) models constitute a powerful and well studied tool to analyze multivariate time series. Since sparseness, crucial to identify and visualize joint dependencies and relevant causalities, is not expected to happen in the standard VAR model, several sparse variants have been introduced in the literature. However, in some cases it might be of interest to control some dimensions of the sparsity, as e.g. the number of causal features allowed in the prediction. To authors extent none of the existent methods endows the user with full control over the different aspects of the sparsity of the solution. In this paper we propose a sparsity-controlled VAR model which allows to control different dimensions of the sparsity, enabling a proper visualization of potential causalities and dependencies. The model coefficients are found as the solution to a mathematical optimization problem, solvable by standard numerical optimization routines. The tests performed on both simulated and real-life multivariate time series show that our approach may outperform both the standard and Group Lasso in terms of prediction errors specially when highly sparse graphs are sought, while avoiding the VAR's overfitting for more dense graphs. Causality; Mixed Integer Non Linear Programming; multivariate time series; sparse models; Vector autoregressive process.

¹Corresponding author: Alba V. Olivares-Nadal, Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Av. Reina Mercedes, s/n, 41012 Sevilla, Spain. *Phone number:* +34 637 872 829 *Email address:* aolivares@us.es (Alba V. Olivares-Nadal)

1 Introduction

A plethora of real world data, such as e.g. air pollution measures or brain functional connections, involve multivariate time series, i.e., different and inter-related features, evolving in time, are simultaneously measured and are to be forecasted. Since the components of such multivariate time series are not independent, inaccurate predictions are expected if the series are analyzed separately by repeatedly using one-dimensional time series forecasting tools. In order to properly address such dependencies, vector autoregressive models (VAR) are frequently applied. However, although capturing features dependencies, there is no reason to expect the so-obtained VAR to be sparse. In other words, the output may be too complex when, on top of obtaining sharp forecasts, visualization of relevant causalities is sought, (Eichler, 2012; Shojaie and Michailidis, 2010; Lozano et al., 2009; Arnold et al., 2007; Valdés-Sosa et al., 2005).

Visualization is a trending topic nowadays, and it is being widely studied to unravel potential causalities in biological systems. In particular, graphical models have been developed to deal with genetic networks (Abegaz and Wit, 2013; Shojaie and Michailidis, 2010; Lozano et al., 2009; Hu and Hu, 2009; Dobra et al., 2004), which includes *E. coli* and *Arabidopsis thaliana* regulatory networks, or human cancer cell data. Moreover, Gorrostieta et al. (2013); Valdés-Sosa et al. (2005) develop sparse autoregressive models to enhance visualization of brain functional connectivity. Also, understanding relevant causalities has played an important role in evaluating the effect of air pollution and exposure over human health (Dominici et al., 2000).

Due to its wide range of applications, several attempts have been proposed in the literature to obtain VAR models in which sparsity, as a potential for easy visualization of complex causal relations, is pursued. This is the case, for instance, of Stochastic Search Variable Selection (George, 2000; George and McCulloch, 1997), Bayesian approaches (Doan et al., 1984) and the Lasso (Tibshirani, 1996). Minimizing the forecasting errors plus an ℓ_1 -penalty regularization term has not led to a unique Lasso method, but it has evolved into a full class of Lasso approaches, such as the so-called Adaptive Lasso, (Zou, 2006), the Group Lasso, (Song and Bickel, 2011; Haufe et al., 2010; Zhao et al., 2009; Yuan and Lin, 2006), the Maximum Likelihood Estimated Lasso (Davis et al., 2012; Hsu et al., 2008), or Lasso Granger methods, (Arnold et al., 2007). Such VAR models attempt to gain overall sparsity, without an explicit analysis of its different aspects.

Nevertheless, it might be of interest in certain real life situations to control not only the overall number of depicted dependencies but also other levels of sparsity. For example, the number of causal features might be wanted to be limited when acceding to the historical records of the features incurs into a cost. Consider a patient who is under surveillance and several tests need to be undertaken to control

her health periodically. Those tests may not only be costly but also invasive for the patient and therefore it may be desirable to reduce their application without affecting much the quality of the diagnose. As an example, Griffin et al. (2005) suggest heart rate to predict neonatal sepsis, instead of obtaining blood from the infants for laboratory tests. This is also useful for the companies shares prices, whose time series are to be paid but, once done, all the historical records are available to use. In all these cases the number of causal features and the number of dependencies from each causal feature are valued differently: as we might seek to limit the first one, we might not be that strict with the second.

This subject was previously noted by Lozano et al. (2009), who stated that “*as a method of Granger graphical modeling, the relevant variable selection question is not whether an individual lagged variable is to be included in regression, but whether the lagged variables for a given time series as a group (i.e. the feature), are to be included*”. To address this issue they proposed to use Group Lasso, in which all the lagged variables of a feature were assigned to the same group. Although this approach reinforces to choose all the past values of a feature once one of them has already been selected for the prediction, it still does not grant control over the exact level of sparsity of the outcome, like none of the other Lasso methods. Moreover, the Lasso approaches perform a shrinkage over the coefficients which, as will be seen in Section 4, might not be advisable when highly sparse graphs are sought.

In contrast we propose a novel sparse approach, formulated as a mathematical optimization problem, which endows the user with the power to control different aspects of the sparsity of the solution. Sparsity is meant here in some of its many different dimensions, as the total number of nonzero coefficients, the total number of features used for the forecast, or the number of past observations of each feature used by the model to make predictions. The performance of this sparsity-controlled VAR method (SC-VAR henceforth) will be compared with three benchmark approaches, namely, the VAR and both the standard and Group Lasso, on simulated and real-life multivariate time series. The results show that the proposed approach outperforms the benchmark Lasso methods in terms of prediction errors when highly sparse graphs are sought, while avoiding the VAR’s overfitting for more dense graphs.

The paper is structured as follows. Next section introduces mathematically the three benchmark methods, the VAR, the classic Lasso and the Group Lasso, and motivates our approach. In Section 3 the SC-VAR model is introduced and expressed as a mixed integer non-linear optimization program, solvable by standard optimization software. Also, a discussion about the choice of the parameters of the model is included. Competing approaches are compared against the proposed method in Section 4 in both simulated and real datasets. Finally, conclusions and

future extensions are collected in Section 5.

2 Preliminaries

Let $\{\mathbf{X}_t\}_{t \geq 0}$ be an N -dimensional vector autoregressive process of order p , $\text{VAR}(p)$, i.e., each series i , $i = 1, \dots, N$, can be expressed as

$$X_{i,t} = c^i + \sum_{j=1}^N \sum_{k=1}^p \alpha_{jk}^i X_{j,t-k} + e_t^i \quad t \geq 0$$

where $\{e_t^i\}_{t \geq 0}$ denotes the series of contemporaneous shocks that affect feature i , and c^i and α_{jk}^i are real numbers. The usual estimation procedures for the coefficients c^i and α_{jk}^i are Maximum Likelihood (which implies making distributional assumptions on the errors e_t^i) or, without imposing any statistical assumption, the Ordinary Least Squares method:

$$\min_{\mathbf{c}, \mathbf{A}} \sum_{i=1}^N \sum_{t=p}^T \left(X_{i,t+1} - c^i - \sum_{j=1}^N \sum_{k=1}^p \alpha_{jk}^i X_{j,t+1-k} \right)^2 \quad (1)$$

where $\mathbf{c} = (c^i)_i \in \mathbb{R}^{1 \times N}$ and $\mathbf{A} = (A^1 | \dots | A^N) \in \mathbb{R}^{N \times Np}$ stand for all unknown coefficients of the process to be estimated. Here $A^i = (\alpha_{jk}^i)_{j,k}$ represents the $N \times p$ matrix of coefficients used to model series i .

There is no reason to expect sparsity in the estimates obtained by maximum likelihood estimation or by solving the nonlinear program (1), and therefore, it may be difficult to visualize causalities while leading to overfitting (Kalli and Griffin, 2014; Li, 2012; Kojima et al., 2009). Among the different procedures proposed in the literature with the aim of obtaining more sparse solutions, a prominent role is given to the Lasso-VAR (Lasso thereafter), in which an ℓ_1 regularization term is added to the objective function (1), and thus estimates are obtained by solving the following nonlinear nonsmooth optimization program:

$$\min_{\mathbf{c}, \mathbf{A}} \sum_{i=1}^N \sum_{t=p}^T \left(X_{i,t+1} - c^i - \sum_{j=1}^N \sum_{k=1}^p \alpha_{jk}^i X_{j,t+1-k} \right)^2 + \sum_{i=1}^N \lambda^i \left(\sum_{j=1}^N \sum_{k=1}^p |\alpha_{jk}^i| \right). \quad (2)$$

As previously mentioned, when features are presented as time series, the use of Group Lasso is encouraged in the literature if more than one-lagged values are considered; see, for instance, Lozano et al. (2009). Such a method groups lagged variables of the same feature, giving more importance to the number of causal features rather than to the overall number of dependencies depicted. This is done by adding an ℓ_2 -penalty separately for each group (Yuan and Lin, 2006):

$$\min_{\mathbf{c}, \mathbf{A}} \sum_{i=1}^N \sum_{t=p}^T \left(X_{i,t+1} - c^i - \sum_{j=1}^N \sum_{k=1}^p \alpha_{jk}^i X_{j,t+1-k} \right)^2 + \sum_{i=1}^N \lambda^i \left(\sum_{r=1}^R \|\boldsymbol{\alpha}_{j(\mathbf{r})}^i\|_2 \right) \quad (3)$$

where $\boldsymbol{\alpha}_{j(\mathbf{r})}^i = (\alpha_{j1}^i, \alpha_{j2}^i, \dots, \alpha_{jp}^i)$.

Infinitely many solutions may be obtained for the Lasso approaches when varying the parameter $\lambda^i \in \mathbb{R}^+$, where more sparse solutions are attained when increasing the value of this penalty (in fact, note that all $\alpha_{jk}^i = 0$ when $\lambda^i \rightarrow +\infty$). However, although one may think that these popular Lasso approaches already cope with sparsity-controlled graphs, there is no way to discern the exact level of sparsity of the outcome when values for the penalties λ^i are given a priori, calling for a (multidimensional) parameter tuning which, at the end, show solutions with different levels of sparsity, but there is no guarantee that sparsity is fulfilled. Moreover, these methods perform a shrinkage over the coefficients, which may deteriorate their prediction power when highly sparse graphs are sought. Even although these methods often provide reasonable sparse solutions, according to some authors (Bertsimas and Copenhaver, 2014; Caramanis et al., 2012) adding an ℓ_1 or ℓ_2 penalty is a robust approach rather than a sparse method. In contrast, Bertsimas and Copenhaver (2014) present Mixed Integer Programming as a useful tool to attain sparsity. This recommendation is supported by the recent improvement on integer optimization solvers, which can attain considerably *good* solutions at a reasonable computational cost. In this paper a sparsity controlled VAR is formulated as a Mixed Integer Non-Linear Problem (MINLP), in which we manage to control different aspects of the sparsity of the outcome.

When control over sparsity is sought, a natural approach is as follows: start with a standard VAR and then, in a naive way, select the largest estimated coefficients (in absolute value) and set to zero the remaining (smaller) coefficients. This naive approach is easy to implement, quick to execute and sounds reasonable. However, its performance may be poor, as we show next. Consider the top left panel of Figure 1, where a simulated VAR of order $p = 3$ is graphically represented. The multivariate time series is represented as a directed graph; nodes correspond to the different features, i.e., the different one-dimensional time series composing the multivariate time series; edges in the graph visualize causality: an arrow from node j to node i means that the model uses feature j in the forecast of feature i ; the thickness of the edges is proportional to the magnitude (in absolute value) of the coefficient relating the features, and therefore it measures the causality's strength when the series are normalized. The color of the edges is related to the lag: the arrow is plotted in black if the present (t) value of a feature is related with a data one period behind ($t - 1$), red for two periods ($t - 2$) and green for three ($t - 3$). Here node 1 receives arrows from nodes 2 and 3, meaning that, in

order to forecast feature 1, past values of features 2 and 3 are used. Note also that node 1 receives here three arrows from node 3, so the present value t of feature 1 is caused by the previous three values $(t - 1)$, $(t - 2)$ and $(t - 3)$ of series 3.

Assume that only one observation of one single feature is allowed to be used to explain a feature. Then the naive approach, illustrated in the top right panel of Figure 1, underestimates strong persistence in favor of just one significant coefficient. This undesired phenomenon is a consequence of the nature of the naive approach: arrows are considered to be kept or removed one by one, and thus the overall picture may be lost. For this reason, instead of using the above-described naive approach, we suggest to express the problem of arrows selection as a mathematical optimization model, solvable by current standard numerical optimization routines. In particular, when applied to the time series of the example, the output of our procedure is visualized through the directed graph in Figure 1 (bottom right panel). It can be observed that the MSE obtained by the SC-VAR solution is considerably smaller than that of the naive approach. In order to illustrate the effect of the shrinkage of Lasso methods over the forecasting power we depict the solution under the standard Lasso in the bottom left panel. This effect is clear since the arrow is thinner, which could be detrimental when highly sparse solutions are sought. This phenomenon, also observed in the results of the experiments carried out in Section 4, will be discussed in more detail. Also it is interesting to note that the chosen causal feature differs for each method: as variable 1 is considered to cause itself for the Group Lasso, feature 3 is considered for our approach instead. This is not surprising since it is known that the VAR might not be identifiable (Lütkepohl, 2005).

3 The SC-VAR.

In this section we will first discuss the sparsity parameters in the SC-VAR model (Section 3.1), which will be later written as a mathematical optimization problem (Section 3.2). In Section 3.3 we will briefly discuss the choice of parameters of our model.

3.1 Different aspects of the sparsity

To the best of our knowledge the existent sparse VAR models attempt to gain overall sparsity, without an explicit analysis of the different aspects of sparsity. The Lasso approaches do not grant the user with the ability to manage the sparsity of the solution in some desirable dimensions either. Indeed, the number of features allowed to be used by the model to explain feature i (V_S^i) or the overall number of nonzero coefficients (V_A), are different measures usually masked under the generic

term of sparsity. Other aspects of the sparsity, that will be also under the user's control in our proposed sparse VAR, are the number of non-zeroes used to explain feature i (V_T^i) or the number of observations per each causal feature of variable i ($V_{S_a}^i$). Moreover, when series are normalized then the strength of a potential causality might be related with the magnitude of its associated coefficient. Hence, in order to avoid spurious dependencies, clear cut-offs ϵ_j^i will be introduced, so that only coefficients with an absolute value greater than or equal to ϵ_j^i are allowed when relating feature i with feature j . All these parameters, included in our SC-VAR, allow the user to obtain an output with the desired level and structure of sparsity.

3.2 Mathematical Programming formulation

The objective of the SC-VAR approach is twofold: control the sparsity of the solution while not damaging much the forecasting capacity. Therefore, in our SC-VAR model, the VAR estimates are obtained by solving the optimization problem (1) imposing on the coefficients of \mathbf{A} the bounds represented by the sparsity parameters $V_A, V_T^i, V_S^i, V_{S_a}^i$ and ϵ_j^i above. They can be expressed as linear constraints by adding logical (binary) variables. Indeed, define the variables δ_{jk}^i to indicate whether a coefficient α_{jk}^i is zero or not, and variables γ_j^i to indicate if feature j is meant to cause variable i (i.e., if $\alpha_{jk}^i \neq 0$ for some k).

Now the SC-VAR model is formulated as the optimization problem (4)-(11), whose outputs are the estimates of the sparse coefficients c^i and α_{jk}^i , as well as the

solution for the indicator variables δ_{jk}^i and γ_j^i .

$$\min_{\mathbf{c}, \mathbf{A}, \Delta, \Gamma} \sum_{i=1}^N \sum_{t=p}^T \left(X_{i,t+1} - c^i - \sum_{j=1}^N \sum_{k=1}^p \alpha_{jk}^i X_{j,t+1-k} \right)^2 \quad (4)$$

s.t

$$\delta_{jk}^i \leq \gamma_j^i \quad \forall k \in K, j, i \in I \quad (5)$$

$$\sum_{j=1}^N \gamma_j^i \leq V_S^i \quad \forall i \in I \quad (6)$$

$$\sum_{k=1}^p \gamma_j^i \delta_{jk}^i \leq V_{Sa}^i \quad \forall j, i \in I \quad (7)$$

$$\sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^p \delta_{jk}^i \leq V_A \quad (8)$$

$$\sum_{j=1}^N \sum_{k=1}^p \delta_{jk}^i \leq V_T^i \quad \forall i \in I \quad (9)$$

$$M \delta_{jk}^i \geq |\alpha_{jk}^i| \geq \epsilon_j^i \delta_{jk}^i \quad \forall k \in K, j, i \in I \quad (10)$$

$$\delta_{jk}^i, \gamma_j^i \in \{0, 1\} \quad \forall k \in K, j, i \in I \quad (11)$$

where $K = \{1, \dots, p\}$, $I = \{1, \dots, N\}$, $\Delta = (\delta_{jk}^i)_{i,j,k}$, $\Gamma = (\gamma_j^i)_{i,j}$ and M is a *large* constant.

Let us briefly discuss the correctness of the formulation above. The objective function (4) minimizes the sum of squared errors. Constraint (5) forces the variable γ_j^i to take the value 1 when some δ_{jk}^i takes the value 1, i.e., as soon as some α_{jk}^i is non-zero. The remaining constraints model different aspects of the sparsity of the process. Indeed, constraints (6) and (7) bound the number of features that are said to cause variable i and the number of non-zero coefficients per each of the chosen causal features of variable i , respectively, for $i = 1, \dots, N$. Constraints (8) and (9) bound respectively the total number of non-zero entries in matrix \mathbf{A} and the total number of non-zero coefficients for each variable i . The shrinking parameter ϵ_j^i is included in the model via constraint (10), which assigns zero to any coefficient that is not allowed to appear on the model, but otherwise it requires $|\alpha_{jk}^i|$ to belong to the interval $[\epsilon_j^i, M]$. Here M is assumed to be a *large* fixed number, and thus this constraint does not exclude reasonable values of the parameters α_{jk}^i .

Problem (4)-(11) is a Mixed Integer Non Linear Program (Burer and Letchford, 2012; Lee and Leyffer, 2012), called MINLP henceforth, with convex quadratic objective function. See e.g. Bertsimas and Copenhaver (2014); Bertsimas and

Mazumder (2014) for other statistical problems recently addressed via optimization in integer numbers. All constraints are linear, except for (10). However, this can be rewritten by introducing new auxiliary variables ν_{jk}^{i+} , ν_{jk}^{i-} , allowing to reformulate Problem (4)-(11) as:

$$\begin{aligned}
& \min_{\mathbf{c}, \mathbf{A}, \Delta, \Gamma} \sum_{i=1}^N \sum_{t=p}^T \left(X_{i,t+1} - c^i - \sum_{j=1}^N \sum_{k=1}^p \alpha_{jk}^i X_{j,t+1-k} \right)^2 \\
& \text{s.t} \\
& \quad (5)-(9), (11) \\
& \quad \alpha_{jk}^i \geq \epsilon_j^i \nu_{jk}^{i+} - (1 - \nu_{jk}^{i+})M \quad \forall k \in K, j, i \in I \quad (\text{P}) \\
& \quad \alpha_{jk}^i \leq -\epsilon_j^i \nu_{jk}^{i-} + (1 - \nu_{jk}^{i-})M \quad \forall k \in K, j, i \in I \\
& \quad \delta_{jk}^i \leq \nu_{jk}^{i+} + \nu_{jk}^{i-} \leq 1 \quad \forall k \in K, j, i \in I \\
& \quad \nu_{jk}^{i+}, \nu_{jk}^{i-} \in \{0, 1\} \quad \forall k \in K, j, i \in I
\end{aligned}$$

Note that the new constraints require that either $-M \leq \alpha_{jk}^i \leq -\epsilon_j^i$ or $\epsilon_j^i \leq \alpha_{jk}^i \leq M$ if coefficient α_{jk}^i is chosen to appear in the model (i.e., if $\delta_{jk}^i = 1$). Now, Problem (P) is a MINLP with quadratic convex objective function and linear constraints. Hence, it can be solved using standard solvers, such as **CPLEX** or **Gurobi**, and it can be easily written using a simple algebraic language such as **AMPL** (Fourer et al., 2002). The lines of the **AMPL** code are included in Appendix A of the Supplementary Material.

Note that (8) is the only constraint of Problem (P) linking the count of the non-zeroes of all features. Hence, if constraint (8) were redundant, the separability of the objective function (4) would allow to solve Problem (P) by solving separately the problem for each feature i :

$$\min_{c^i, A^i} \sum_{t=p}^T \left(X_{i,t+1} - c^i - \sum_{j=1}^N \sum_{k=1}^p \alpha_{jk}^i X_{j,t+1-k} \right)^2 .$$

Since such problems are of much smaller dimension, we see that removing constraint (8) would allow one to cope with databases with a larger number of time series.

3.3 Choice of parameters

Compared with other methods seeking sparsity, our approach has many more parameters, which may be, at first glance, discouraging. Some comments are in

order to treat this issue. First note that, as opposed to the Lasso approaches, the parameters which are to be determined in Problem (P) have a precise meaning in terms of the sparsity, and therefore can be fixed by the user a priori without carrying out any tuning if a determined level of sparsity is sought. In other words, they are not parameters to tune but decisions to make with respect to the sparsity desired. Second, although apparently the proposed methodology consists of four parameters per series (V_T^i , V_S^i , $V_{S_a}^i$ and ϵ_j^i) and one global parameter (V_A), not all of them need to be fixed to address the problem. Indeed, it suffices to fix the parameters that are significant for the user and choose the rest of them so that the remaining constraints are redundant. For instance, if the user only cares about the number of causal features, it suffices to fix only V_S^i . As an illustration consider Figure 1, discussed in Section 2, where the SC-VAR approach was solved fixing nothing but $V_T^1 = 1$.

Finally, Problem (P) can be solved by fixing solely V_A , which represents the overall sparsity of the whole graph. This means that the SC-VAR can be solved for all features simultaneously, while the Lasso approaches are always solved separately for each feature. This provides our model with more flexibility than Lasso methods to obtain a graph with a specified level of overall sparsity, placing pools of non-zeroes where necessary, as will be seen in Figure 2. Summarizing, the more thorough the user wants to be with the structure of the solution, the more parameters she has to fix. This is the price to pay to control sparsity, if sought, in many of its dimensions.

4 Numerical illustrations

In this section the SC-VAR is compared with three benchmark approaches, the VAR, the Lasso and the Group Lasso, on simulated as well as real datasets. Table 1 summarizes the methods under comparison. Further descriptions on such methods and the choice of the parameters used are provided below.

4.1 Comparison methodology

For the proposed numerical examples, the Lasso coefficients were obtained via the Least Angle Regression algorithm, LAR. This algorithm solves the Lasso for each variable i separately. Since at each step the LAR incorporates a new predictor for variable i , it provides a set of solutions with different levels of sparsity (i.e. different number of non-zeroes per node, V_T^i). The Lasso set of solutions was obtained by using the `lars()` function of R-cran package `lars` (Hastie and Efron, 2013).

Solutions for the Group Lasso were obtained by using the functions of the R-cran package `gglasso` (Yang and Zou, 2015). In particular, the function `cv.gglasso`

was used to perform a 5-fold cross-validation over the fits obtained for each value of the penalties λ^i and calculate their mean cross-validated errors. Recall that the Group Lasso understands sparsity in terms of the number of causal features, rather than the overall number of non-zero coefficients per node. As the sparsity of the output cannot be discerned in advance, we had to tune its parameters: we solved the Group Lasso for a grid of λ^i large enough, so solutions for all possible values of V_S^i were obtained.

The SC-VAR solution was obtained by solving Problem (P) using **Gurobi** version 6.0.0. As previously commented, the Lasso approaches are solved independently for each node. Hence, in order to fairly compare against these methods, the SC-VAR parameter V_A was fixed so as the constraint (8) was redundant. In this way, Problem (P) could be solved separately for each feature. To test the performance of our approach under different requirements of sparsity, the SC-VAR was solved for all possible combinations of parameters V_T^i , V_S^i and $V_{S_a}^i$. In order to test the influence of the shrinking parameter of constraint (10), $\epsilon_j^i = \epsilon$ for all i, j and the sparse problem was solved for $\epsilon = 0$ and $\epsilon = 0.2$. However, the obtained MSEs were mainly equivalent and thus we only report here the results for $\epsilon = 0.2$ to conserve space. We found interesting that considering $\epsilon = 0.2$ not only leads to a neater graph, but it may also yield different solutions to those obtained for the case $\epsilon = 0$. This might often be caused by the non-identifiability of the process. Therefore it can be concluded that when no small non-zero coefficients are wanted, then it is not optimal to simply delete or increase such coefficients, as a better solution can be obtained by dismissing old links and adding new ones between other features. In contrast to the VAR, constraint (10) of the SC-VAR bounds the sparse coefficients by a constant M . In our experiments M was fixed to 2, although other values were tested, with similar performance. Observe that, as customary in the literature, the choice of such M is problematic, since a very small value may exclude reasonable values of the coefficients α_{jk}^i , whilst a very large value of M is to cause severe numerical troubles (Camm et al. (1990)).

Two criteria are used to compare the methods, namely, the MSE and the sparsity. Each time series is divided in train and test sets with cardinality T_{train} and T_{test} , respectively. The solutions for each method are obtained using only the data of the train set, and the MSE for such solutions is calculated in the test set. All the results presented now are normalized by dividing by the VAR solution; that is to say, when the SC-VAR or the Lasso approaches attain a MSE greater than 1 their prediction capacity is estimated to be worse than that of the VAR, while for smaller values the forecasting power is expected to improve. For VAR(1) processes, the SC-VAR was compared against the Lasso, since no grouping effects amongst lagged variables were necessary for these cases. For VAR(p) processes with $p > 1$ the SC-VAR was also compared against the Group Lasso when it was

sought to test the influence of the number of causal features over the MSE.

4.2 Simulation study

In order to test the performance of all the approaches we generate synthetic data following VAR(1) and VAR(2) processes of 10 nodes with i.i.d. errors drawn from a standard Normal distribution. To generate the coefficients of such multivariate time series we roughly follow the experiments conducted in Arnold et al. (2007); Lozano et al. (2009), in which an *affinity* parameter is fixed. Such a parameter is the probability that an edge is included in the graph. For example, if we want a graph to have a 20% density we fix the *affinity* parameter to 0.2. Then, we randomly generate all the coefficients of the process matrix \mathbf{A} and decide whether each off-diagonal element is included by simulating a Binomial distribution with success probability 0.2.

So as to test the performance of the sparse approaches under graphs with different levels of sparsity, VAR processes with densities 0%, 10%, 20%,...,100% have been generated. A 0% density means that each feature follows an independent autoregressive process (diagonal matrix) and a 100% density implies that there exists correlation amongst all nodes. For each level of density 100 instances of VAR processes were generated. Each time series has 1000 observations, whose first 500 were assigned to the train set and the remaining to the test set. Although the VAR, Lasso, Group Lasso and SC-VAR were solved for all levels of density, the results were mainly equivalent and thus only the results for simulations with a 10%, 50% and 90% density are included here. Table 3 reports the median of the normalized MSEs for such densities for Lasso and SC-VAR for VAR(1) processes, while Table 4 reports them for Group Lasso and SC-VAR for VAR(2) processes.

4.2.1 $p = 1$.

From Table 3 it can be observed that the SC-VAR outperforms the Lasso in terms of MSE when highly sparse graphs are sought (that is, for small values of V_T). This outperformance over the Lasso increases as the density of the process does. Indeed, the difference between the Lasso and the SC-VAR is at least a 5% deterioration over the VAR MSE when the maximum number of non-zeroes required is less than 2, 4 or 6 for processes with densities 10%, 50% and 90%, respectively. While the sparsity-controlled approach can sometimes attain an MSE 9% better than the Lasso ($V_T = 2$, 90% density), the Lasso outperforms it for highly dense graphs ($V_T \geq 8$, 90% density). In order to illustrate more in depth the results, randomly selected instances are depicted in Figure 2, together with their VAR solutions. For the sake of abbreviation, only the most extreme solutions of the SC-VAR and the Lasso ($V_T = 1$ and $V_T = 10$) have been depicted.

From Figure 2 some observations arise. First, the VAR leads to overfitting. This is clearer as more sparse the real graph is. Second, the Lasso is equivalent to the VAR for $V_T = 10$. However, the SC-VAR attains a much more sparse solution while providing a similar MSE. It seems that requiring the absolute value of the non-zero coefficients to be larger than a threshold ($\epsilon = 0.2$ in constraint (10)) helps avoiding overfitting in these particular cases, leading to graphs that are more similar to the original ones. Observe that this idea of defining a clear cut between non-zero and zero coefficients is rather different to the behaviour of the Lasso, which shrinks coefficients towards zero. Third, although the two sparse approaches under comparison attain equally sparse graphs for $V_T = 1$, the SC-VAR yields better MSEs. Note that the chosen causal features are different for the two approaches. Moreover, as noted in Section 2, larger penalties for the Lasso imply more sparsity but stronger shrinkage, entailing a loss in its prediction power. Finally, as an illustration results for the SC-VAR fixing nothing but parameter V_A have been also depicted. We required a total of 10 non-zero coefficients for the whole graph, obtaining the same number of arrows as when requiring a maximum of 1 non-zero per feature. Observe that the obtained graphs for processes with 50 and 90% density are similar to the outputs of SC-VAR solved fixing $V_T = 1$ for each node. However, for the process with 10% density the MSE is improved by encouraging extra-diagonal elements. Note that some features receive more than one arrow, although the level of sparsity of the whole graph is the same. In conclusion, the forecasting power can be improved while maintaining the same level of sparsity by allowing a pooling effect on the non-zeroes (i.e., by solving the SC-VAR with binding (8) constraint).

4.2.2 $p = 2$.

In Table 4 it can be observed that the SC-VAR usually helps avoiding VAR overfitting. Note that the SC-VAR outperformance over the VAR is clearer for $p = 2$ than for $p = 1$, which is reasonable since the number of parameters of the VAR increases with p . As solving the SC-VAR with fixing nothing but $V_S = 10$ is equivalent to the standard VAR, the improvement over its MSE is thanks to the choice $\epsilon = 0.2$.

In Table 4 it can be observed that the SC-VAR always outperforms the Group Lasso in terms of MSE, the outperformance being clearer as the sparsity of the output decreases. Moreover, it seems that the SC-VAR performs equivalently with respect to the VAR no matter the density of the original graph, but the Group Lasso performs worse for large V_S in truly dense processes. For instance, while the SC-VAR improves the VAR's forecasting error in a 13% for $V_S = 10$ and processes with 90% density, the Group Lasso attains a 48% of worsening.

4.2.3 Computational times.

Although it would be interesting to study how data parameters (N , p and T) and chosen parameters (such as V_T , V_S , V_{Sa} ...etc.) affect computational times when solving the MINLP (P), carrying out such costly experiments are out of the scope of this paper. However, some intuitions arise from our numerical experience. As the VAR consists of $N^2p + N$ coefficients, the computational times are expected to increase specially with N and p , and we have experienced so in our experiments. On the other hand, we found that the SC-VAR is not that sensitive to the length of the time series T .

In order to illustrate the effect of the sparsity requirements of the SC-VAR over the computational cost, consider Figure 3, where the median elapsed time in seconds has been depicted against the sparsity of the output, measured in terms of V_S . For the sake of clarity, only the times obtained for VAR(1) and VAR(2) processes with 10, 50 and 90% densities are depicted. Some comments are in order here. First, Figure 3 supports the intuition about computational times increasing with p . Second, although there is no much difference in the computational times for VAR(1) processes, it seems that for VAR(2) processes the computational times increase with the density of the true graph. Third, for $p = 2$ it is clear that the SC-VAR computational burden is consistently lower when either highly sparse or dense graphs are sought. To conclude, for $p = 1$ the SC-VAR takes to solve around 0.6 seconds on a PC Intel® Core™ Quad CPU 4GB RAM, while for $p = 2$ the behavior is less consistent, but it usually obtains highly sparse graphs ($V_S \leq 2$) in less than 1 second.

4.3 Real data sets

In this section the performance of the three competing approaches is compared in two real databases, whose main features are summarized in Table 2. The series have not only been normalized but they have also undergone the Augmented Dickey-Fuller test for stationarity, implemented in the function `adf.test()` of the R-cran's package `tseries`. The order p in the autoregressive model was unknown and hence chosen from $p = \{1, 2, \dots, 7\}$ by the Schwartz criterion, implemented on the function `VARselect()`, available in the R-cran's package `vars`.

4.3.1 Google flu database.

This database is derived by the 45 Google user search terms that are considered to be indicative of influenza activity in the U.S. The sample is measured weekly from the beginning of the year 2006 until the week of June 6, 2014. According to the Centers for Disease Control and Surveillance (CDC) the probability that a

patient query is related to influenza-like-illness is closely related to the data in the Google flu database. From the 51 considered regions (50 states and the District of Columbia), North and South Dakota as well as Wyoming have been removed due to missing data.

Figure 4 shows the results obtained for Lasso and SC-VAR in terms of MSE normalized with respect to the VAR, which is represented against the upper-bound on the number of causal features per node (or, equivalently for $p = 1$ cases, the upper-bound on the number of non-zero entries per node V_T).

From Figure 4 some conclusions arise. First, the SC-VAR can considerably reduce the density of the matrix \mathbf{A} , containing the coefficients of the VAR, with an improvement of the MSE. Although one may think that fewer non-zero entries in \mathbf{A} would lead to better MSE, increasing the freedom of the model may lead to overfitting. For instance, see that the SC-VAR solutions for $V_T \leq 15$ or Lasso solutions for $V_T \geq 3$ report a MSE smaller than 1; i.e., the prediction power of those sparse graphs is better than that of the solution provided by the VAR. Finally, it can be observed that although Lasso outperforms the SC-VAR for $V_T \geq 6$, it can attain extremely worse MSE when highly sparse graphs are sought.

Together with the MSE plots, Figure 5 depicts heatmaps of the VAR, Lasso and SC-VAR solutions for the Google flu data set. For the sake of abbreviation, only the most sparse solutions of the SC-VAR and Lasso are included. The color represents the sign of the coefficients α_{jk}^i (blue for negative, red for positive) and the intensity is related to the magnitude of such coefficients. The names of the states have been replaced by their abbreviations.

Although Lasso provides a much more sparse solution than the VAR and enhance the visualization of potential causalities of the flu for the different regions of the US, the price of gaining such a level of sparsity is a 192% deterioration over its MSE. Nevertheless, the SC-VAR solution for $V_T = 1$ considerably improves the VAR's forecasting capacity (it attains a MSE with a 38% improvement over the VAR) while also allowing for an easier interpretation thanks to its sparsity. Also note that, since the absolute value of the coefficients must be larger than the threshold 0.2, the obtained heat map is sharper than that of the Lasso. It is also interesting to note that SC-VAR approach tends to strengthen diagonal elements; i.e., with the SC-VAR in most cases the chosen causal feature for a node is itself when only one non-zero is allowed. This behavior is not as evident with the Lasso.

4.3.2 Air pollution database.

The data consists of hourly records of the solar radiation intensity (R) and the levels of four air pollutants, namely CO, NO, NO₂ and O₃, measured in Azusa, California during the year 2006. Figure 6 depicts the normalized MSEs for the Lasso and the SC-VAR against two different measures of the sparsity: the upper-

bound on the total number for non-zeroes per feature (V_T , left panel), and the upper-bound on the number of causal features (V_S , right panel). Note that for this second case the SC-VAR is compared against the more suitable Group Lasso instead.

In the left panel of Figure 6 we observe that there exists a clear difference in the impact of the parameters V_S and V_{Sa} over the MSE: increasing the parameter V_{Sa} seems more efficient to reduce the prediction error than increasing V_S . Therefore it is advisable to treat both parameters separately, since they clearly represent different aspects of the sparsity. This is done in a natural way with our approach, but not with the Lasso. Furthermore, note that our approach seems to stabilize for $V_T \geq 6$, whatever the values of V_S and V_{Sa} are. The constant MSE seems to denote that no further coefficients are being added to the model. Also, some of these constant lines explain that the constraint involving V_T is redundant, as the $V_S \cdot V_{Sa} \leq V_T$.

The behaviour observed from both plots of Figure 6 is analogous to that of the previous results: when highly sparse graphs are sought, the SC-VAR seems to be more appropriate as it attains a better MSE. We point out that when only one non-zero is allowed ($V_T = 1$), the SC-VAR yields a 16% worsening over the VAR, while the Lasso reports a 49% deterioration. Moreover, when the number of causal features wants to be limited, the differences between the prediction errors of the SC-VAR and Group Lasso approaches are roughly a 200% and 50% for one and two causal features, respectively. Note also that the MSE deterioration of the SC-VAR over the VAR is less than a 2% for $V_S \geq 2$.

Figure 7 depicts, through directed graphs, the solutions of the VAR, and a couple of solutions for the Group Lasso and the SC-VAR for the air pollution data set. In such graphs color blue is associated with a 4-lag dependency. From Figure 7 some comments arise. First, note that although the chosen relevant causal features are the same for the Group Lasso and SC-VAR when $V_S = 1$, the Group Lasso implies a 191% worsening over the VAR's prediction error, while the SC-VAR deteriorates it in a 8%. Second, the unraveled potential causal features obtained for each method are different when $V_S = 2$. The choice of the SC-VAR provides a much more sparse graph than the VAR while obtaining 26% less deterioration than its sparse counterpart.

5 Concluding remarks and extensions

In this paper a sparse vector autoregressive model, the SC-VAR, that allows the user to control the sparsity of the output from various perspectives has been introduced. The model's sparsity is expressed in terms of different parameters, such as the number of total non-zero entries per series, the number of features involved

or the number of periods chosen per feature. The method is expressed as an optimization problem, solvable by standard numerical optimization software.

The ability of the proposed approach to unravel potential causalities and, in many cases, to improve the fit, has been tested in simulated multivariate time series as well as in two real data bases, referred in the existent literature. It is concluded from the experiments that (i) the proposed approach is able to yield very sparse solutions either improving or without significantly increasing the VAR's forecasting error, (ii) the SC-VAR usually yields better MSEs than both the classic and Group Lasso when highly sparse graphs are sought, leading to a much better visualization of the process dependencies, as e.g. depicted in Figure 2, and (iii) the parameters considered to measure the sparsity play different roles, thus it might not be advisable to aggregate them into one single measure, as done by other sparse methods.

Although the computational times were often negligible for the simulations, where ten nodes were considered, they are expected to increase dramatically within the dimensionality of the data sets. Hence, the authors are planning to develop in the future heuristics enabling to cope with very large data sets.

6 Supplementary Material

Supplementary material is available online at (to appear)

Acknowledgments

This research is supported by projects MTM2012-36163 (Ministerio de Economía y Competitividad, Spain), P11-FQM-7603 and FQM-329 (Junta de Andalucía), all with EU ERD Funds.

We thank Richard A. Davis, Pengfei Zang and Tian Zheng for providing the *concentration of air pollutants* database.

Conflict of Interest: None declared.

References

- Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, 14(3):586–599.
- Arnold, A., Liu, Y., and Abe, N. (2007). Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge discovery and data mining*, 66–75. ACM.

- Bertsimas, D. and Copenhaver, M. S. (2014). Characterization of the equivalence of robustification and regularization in linear, median, and matrix regression. arXiv preprint arXiv:1411.6160.
- Bertsimas, D. and Mazumder, R. (2014). Least quantile regression via modern optimization. *The Annals of Statistics*, 42(6):2429–2525.
- Burer, S. and Letchford, A. N. (2012). Non-convex mixed-integer nonlinear programming: A survey. *Surveys in Operations Research and Management Science*, 17(2):97–106.
- Camm, J. D., Raturi, A. S., and Tsubakitani, S. (1990). Cutting big M down to size. *Interfaces*, 20(5):61–66.
- Caramanis, C., Mannor, S., and Xu, H. (2012). Robust optimization in machine learning. In Sra, S., Nowozin, S., and Wright, S. J., eds., *Optimization for machine learning*, 369–402. MIT Press, Michigan.
- Davis, R. A., Zang, P., and Zheng, T. (2012). Sparse vector autoregressive modeling. arXiv preprint arXiv:1207.0520.
- Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.
- Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212.
- Dominici, F., Zeger, S. L., and Samet, J. M. (2000). A measurement error model for time-series studies of air pollution and mortality. *Biostatistics*, 1(2):157–175.
- Eichler, M. (2012). *Causality*, chapter Causal inference in time series analysis, 327–354. Wiley Series in Probability and Statistics.
- Fourer, R., Gay, D., and Kernighan, B. W. (2002). *The AMPL book*. Duxbury Press, Pacific Grove.
- George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308.
- George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica sinica*, 7(2):339–373.
- Gorrostieta, C., Fiecas, M., Ombao, H., Burke, E., and Cramer, S. (2013). Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity.

- Griffin, M. P., Lake, D. E., and Moorman, J. R. (2005). Heart rate characteristics and laboratory tests in neonatal sepsis. *Pediatrics*, 115(4):937–941.
- Hastie, T. and Efron, B. (2013). Least Angle Regression, Lasso and Forward Stagewise. <http://cran.r-project.org/web/packages/lars/lars.pdf>.
- Haufe, S., Nolte, G., Mueller, K.-R., and Krämer, N. (2010). Sparse causal discovery in multivariate time series. *JMLR W&CP*, 6:97–106.
- Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.
- Hu, J. and Hu, F. (2009). Estimating equation-based causality analysis with application to microarray time series data. *Biostatistics*, 10(3):468–480.
- Kalli, M. and Griffin, J. E. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793.
- Kojima, K., Yamaguchi, R., Imoto, S., Yamauchi, M., Nagasaki, M., Yoshida, R., Shimamura, T., Ueno, K., Higuchi, T., Gotoh, N., et al. (2009). A state space representation of VAR models with sparse learning for dynamic gene networks. *Genome Informatics*, 22:56–68.
- Lee, J. and Leyffer, S. (2012). *Mixed integer nonlinear programming*. Springer.
- Li, J. (2012). Monetary policy analysis based on Lasso-Assisted Vector Autoregression (Lavar). Available at SSRN 2017877.
- Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118.
- Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.
- Shojaie, A. and Michailidis, G. (2010). Discovering graphical granger causality using the truncating Lasso penalty. *Bioinformatics*, 26(18):i517–i523.
- Song, S. and Bickel, P. J. (2011). Large vector auto regressions. arXiv preprint arXiv:1106.3915.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

- Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., and Canales-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):969–981.
- Yang, Y. and Zou, H. (2015). Group Lasso Penalized Learning Using A Unified BMD Algorithm. <https://cran.r-project.org/web/packages/gglasso/gglasso.pdf>.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.
- Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 3468–3497.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Abbreviation	Method	Expression
VAR	Classic vector autoregressive approach	(1)
Lasso-VAR	Vector autoregression + Lasso	(2)
Group Lasso	Vector autoregression + Group Lasso + 5-fold cross-validation	(3)
SC-VAR	Sparsity-controlled vector autoregressive method	(P)

Table 1: Summary of methods under comparison in the computational illustrations

Abbreviation	Name	N	T_{train}	T_{test}	p	Reference
Google flu	Google Flu Trends	48	221	220	1	Davis et al. (2012)
Air pollution	Concentration levels of air pollutants	5	4185	4185	4	Davis et al. (2012)

Table 2: Summary of real databases used for numerical illustrations

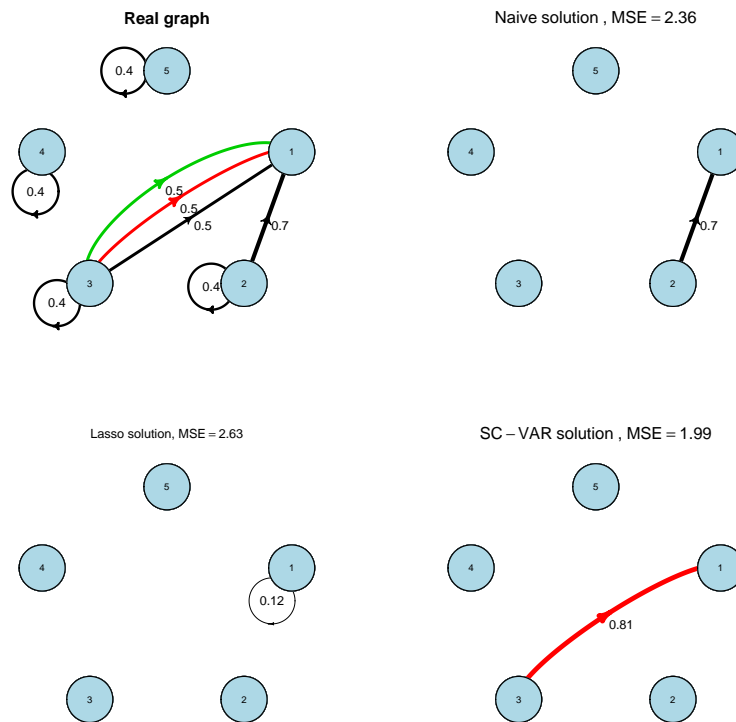


Figure 1: Graphical representation of a simulated VAR and its sparse counterparts. Naive algorithm (center) versus the proposed SC-VAR method (right).

	Density 10%		Density 50%		Density 90%	
	Lasso	SC-VAR	Lasso	SC-VAR	Lasso	SC-VAR
$V_T = 1$	1.09	1.04	1.23	1.15	1.23	1.17
$V_T = 2$	1.03	1.00	1.15	1.07	1.20	1.11
$V_T = 3$	1.01	0.99	1.09	1.03	1.15	1.08
$V_T = 4$	0.99	0.99	1.05	1.02	1.12	1.05
$V_T = 5$	0.99	0.99	1.03	1.01	1.09	1.04
$V_T = 6$	1.00	0.99	1.01	1.01	1.06	1.04
$V_T = 7$	1.00	0.99	1.00	1.01	1.03	1.03
$V_T = 8$	1.00	0.99	1.00	1.01	1.02	1.03
$V_T = 9$	1.00	0.99	1.00	1.01	1.00	1.03
$V_T = 10$	1.00	0.99	1.00	1.01	1.00	1.03

Table 3: Normalized median MSE under the Lasso and the SC-VAR for VAR(1) processes generated with a 10, 50 and 90% density, for different requirements of sparsity ($V_T=1,\dots,10$)

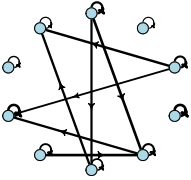
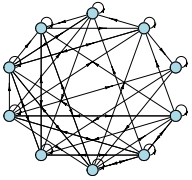
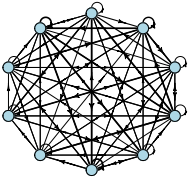
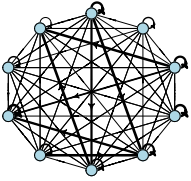
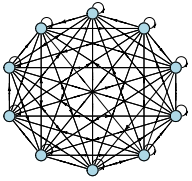
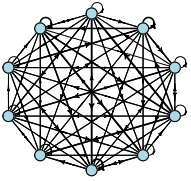
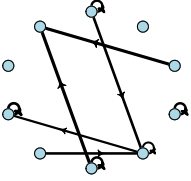
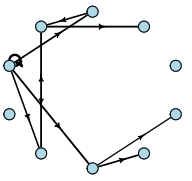
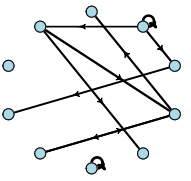
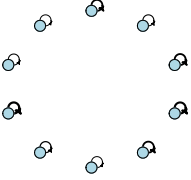
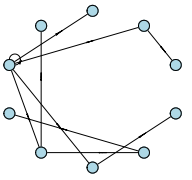
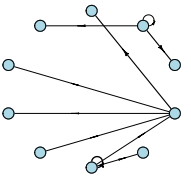
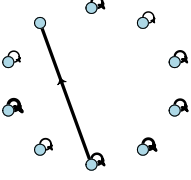
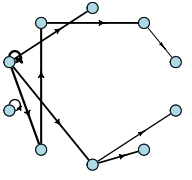
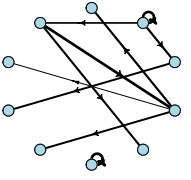
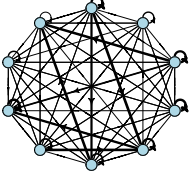
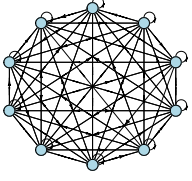
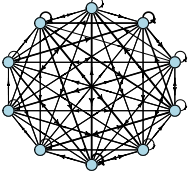
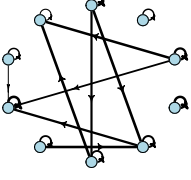
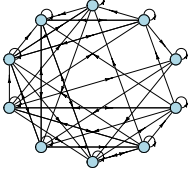
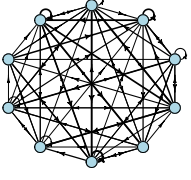
Density		10%	50%	90%
Real				
VAR		MSE = 1 	MSE = 1 	MSE = 1 
$V_A = 10$	SC-VAR	MSE = 1.19 	MSE = 1.32 	MSE = 1.78 
	Lasso	MSE = 1.54 	MSE = 1.9 	MSE = 2.4 
$V_T = 1$	SC-VAR	MSE = 1.34 	MSE = 1.35 	MSE = 1.8 
	Lasso	MSE = 1 	MSE = 1 	MSE = 1 
$V_T = 10$	SC-VAR	MSE = 0.99 	MSE = 1 	MSE = 1.03 

Figure 2: Real graphs of randomly selected instances of processes with different densities, represented along with their VAR solutions. SC-VAR and Lasso solutions are represented for a couple of levels of sparsity, $V_T = 1$ and $V_T = 10$, and the SC-VAR is also depicted fixing solely $V_A=10$.

	Density 10%		Density 50%		Density 90%	
	Group Lasso	SC-VAR	Group Lasso	SC-VAR	Group Lasso	SC-VAR
$V_S = 1$	1.13	0.95	1.11	1.01	1.12	1.02
$V_S = 2$	1.15	0.90	1.11	0.96	1.13	0.99
$V_S = 3$	1.16	0.90	1.14	0.90	1.14	0.94
$V_S = 4$	1.17	0.90	1.17	0.89	1.15	0.91
$V_S = 5$	1.18	0.90	1.20	0.87	1.18	0.89
$V_S = 6$	1.19	0.90	1.24	0.87	1.21	0.88
$V_S = 7$	1.20	0.90	1.28	0.87	1.25	0.87
$V_S = 8$	1.20	0.90	1.30	0.87	1.31	0.87
$V_S = 9$	1.21	0.90	1.33	0.87	1.40	0.87
$V_S = 10$	1.24	0.90	1.36	0.87	1.48	0.87

Table 4: Normalized median MSE under the Group Lasso and the SC-VAR for VAR(2) processes generated with a 10, 50 and 90% density, for different requirements of sparsity in terms of number of causal features ($V_S=1,\dots,10$)

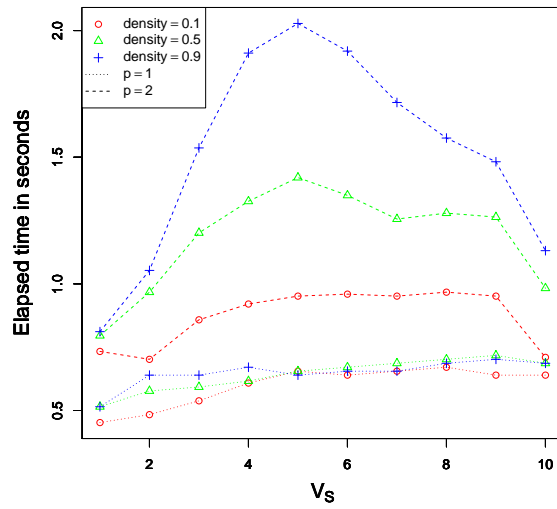


Figure 3: Median elapsed times in seconds taken by the SC-VAR to be solved for different requirements of the sparsity in VAR(1) and VAR(2) processes with various levels of density.

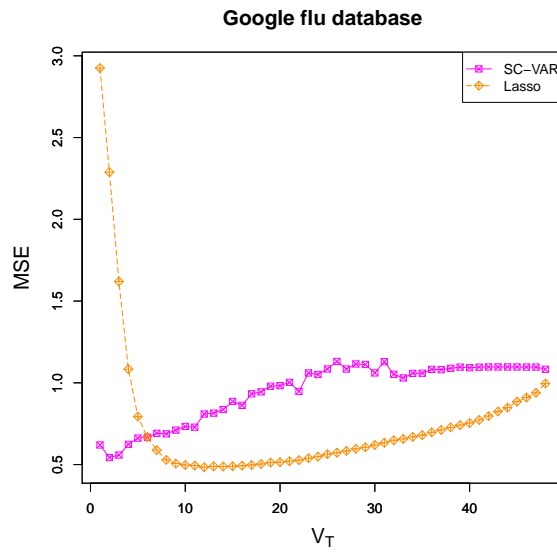


Figure 4: Normalized MSEs for the SC-VAR and the Lasso under different levels of sparsity.

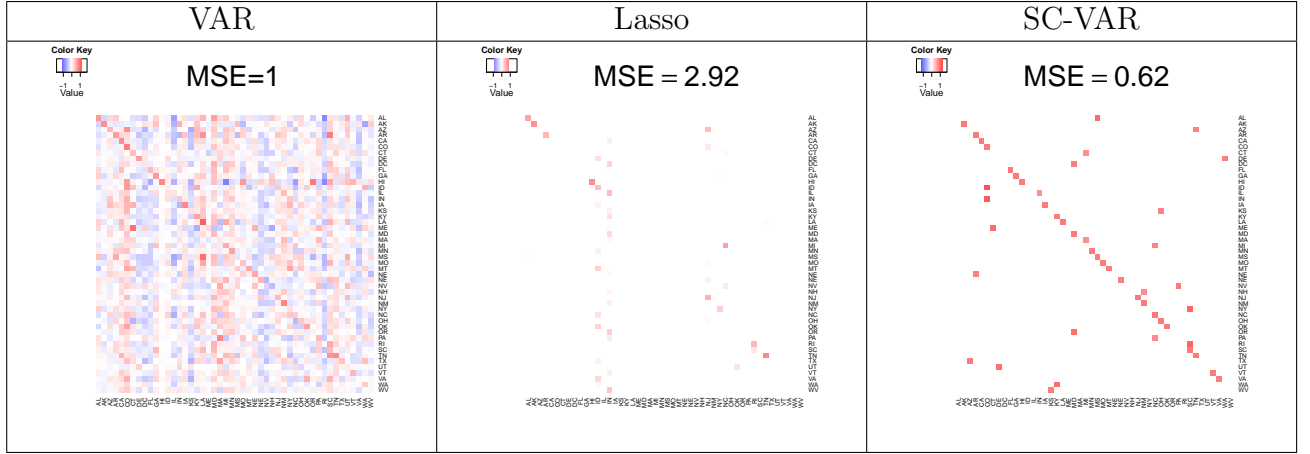


Figure 5: Heat maps representing the solutions of the VAR, and the SC-VAR and Lasso (for $V_T = 1$) for the Google flu database.

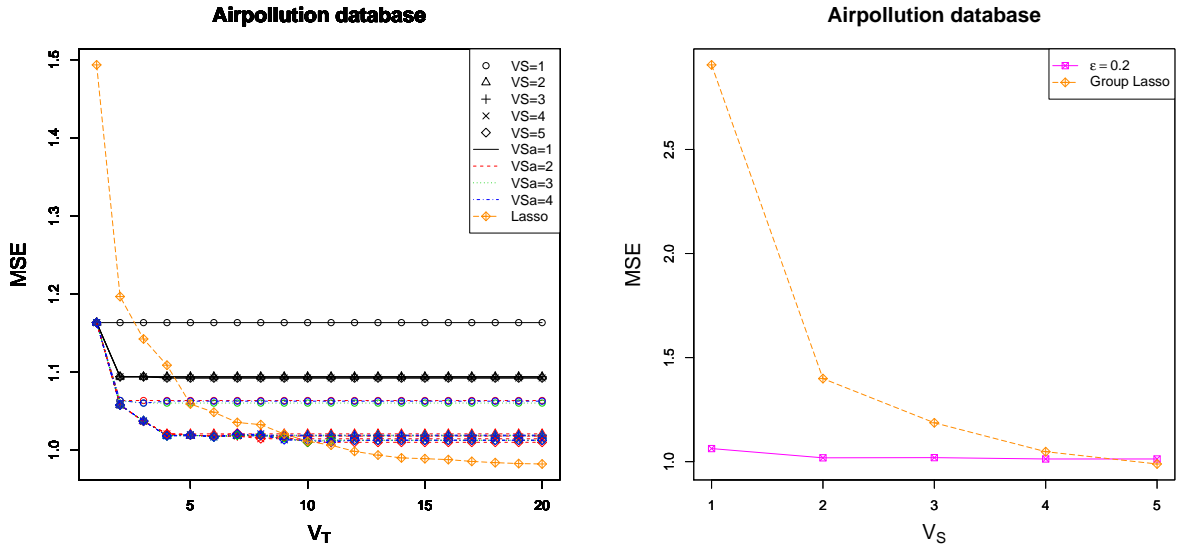


Figure 6: Normalized MSEs for the SC-VAR and Lasso approaches under different aspects of sparsity for the air pollution database. In the left panel, the x axis represents the upper bound on the total number of arrows received by a node (V_T), while in the right panel it represents the upper bound on the total number of causal features (V_S).

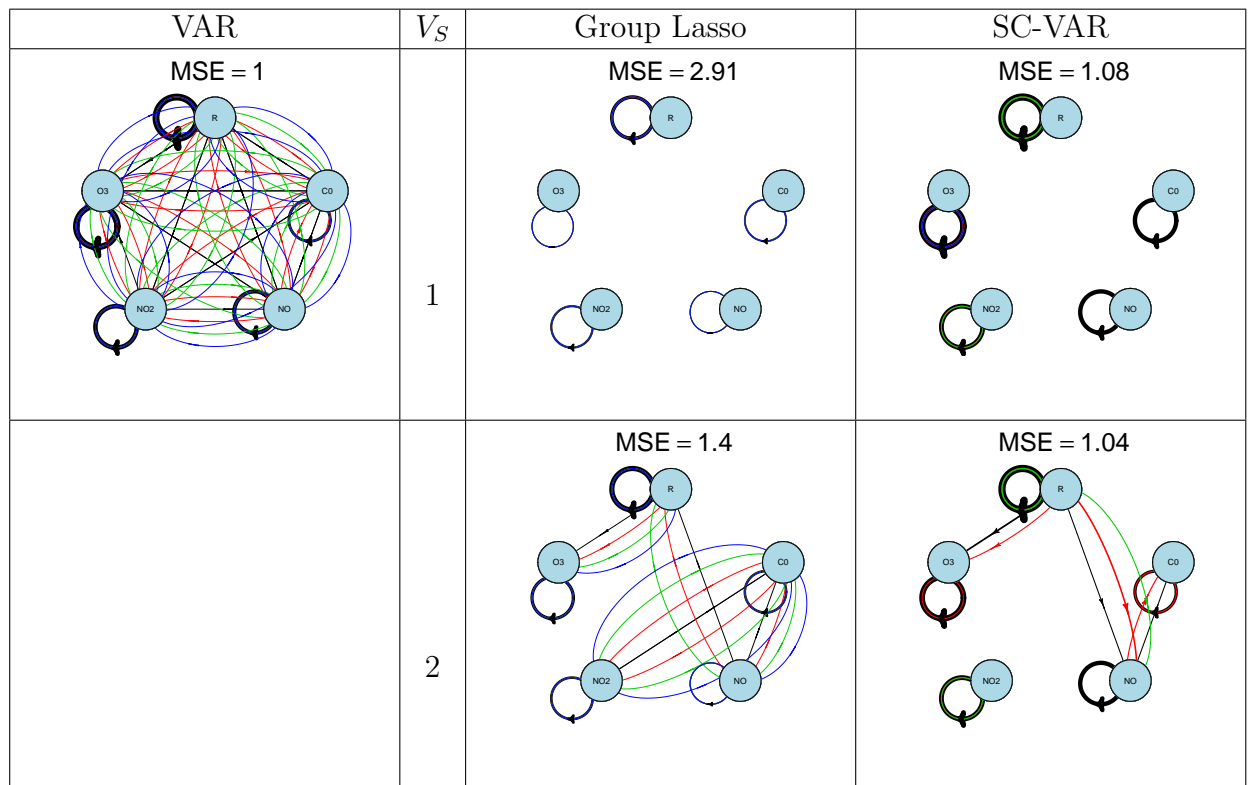


Figure 7: VAR solution (top-left panel) together with the SC-VAR (right) and Group Lasso (central) outputs when allowing one ($V_S = 1$) or two ($V_S = 2$) causal features for the airpollution database.