Consistent Causal Inference for High Dimensional Time Series*

Francesco Cordoni† and Alessio Sancetta‡

January 30, 2023

Abstract

A methodology for high dimensional causal inference in a time series context is introduced. It is assumed that there is a monotonic transformation of the data such that the dynamics of the transformed variables are described by a Gaussian vector autoregressive process. This is tantamount to assume that the dynamics are captured by a Gaussian copula. No knowledge or estimation of the marginal distribution of the data is required. The procedure consistently identifies the parameters that describe the dynamics of the process and the conditional causal relations among the possibly high dimensional variables under sparsity conditions. The methodology allows us to identify such causal relations in the form of a directed acyclic graph. As an application we estimate the directed acyclic graph for the order book on one-minute aggregated data on four stock constituents of the S&P500.

Key Words: High frequency trading, high dimensional model, nonlinear model, order book.

JEL Codes: C14, G10.

*We are grateful to Yanqin Fan for having shared the latest version of Fan et al. (2022). The first author acknowledges financial support from the project "How good is your model? Empirical evaluation and validation of quantitative models in economics" funded by MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2017.

†Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK. Email: francesco.cordoni@rhul.ac.uk

‡Corresponding Author. Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK. Email: asancetta@gmail.com

1 Introduction

Identifying and estimating causal relations is a problem that has received much interest in economics. In the last two decades the statistical and machine learning literature has made a number of advances on the front of identification and estimation within the framework of causal graphs (Comon, 1994, Hyvärinen and Oja, 2000, Pearl, 2000, Spirtes et al., 2000, Hyvärinen et al., 2001, Shimizu et al., 2006, Meinshausen and Bühlmann, 2006, Kalisch and Bühlmann, 2007, Cai et al., 2011, Bühlmann et al., 2014, Peters et al., 2014), where the data generating process can be characterized as a system of structural equations. This complex causal relations system might be represented through the causal graph, which conveys essential topological information to estimate causal effects.

However, the true data generating process is often a latent object to researchers, which can only rely on finite sample observations to infer the causal structure and mechanism of the true system. A causal model entails a probabilistic model from which a researcher can learn from observations and outcomes about changes and interventions of the system variables (Pearl, 2000, Peters et al., 2014). Thus, causality can be formally defined using the do-notation of Pearl (2000) in terms of intervention distributions. This definition of causality is quite different from the well known concept of Granger causality. However, causal relations in economics and finance require to account for time series dependence.

In this paper we develop a methodology to extract the causal relations of time series data, conditioning on the past in a flexible way. We assume that there is a monotone transformation of the data that maps the original variables into a Gaussian vector autoregressive (VAR) model (see also Fan et al., 2022). There are a number of advantages to this approach. First, we are able to retain the interpretability of VAR models building on the rich econometrics literature on structural VAR models. Second, we do not need any assumptions on the marginal distribution of the data. This means that the procedure is robust to fat tails, as we do not make any assumption on the existence of any moments. For instance, given that the existence of a second moment for financial data has been a much debated topic in the past (Mandelbrot, 1963, Clarke, 1973, for some of the earliest references) dispensing all together of this unverifiable condition should be welcomed. Third, we can model variables that take values in some subset of the real line, for example variables that only take positive

values or are truncated. This is not possible using a standard VAR model.

The estimation of the contemporaneous causal structure of a time series is equivalent to solving the identification problem of a structural VAR model. The latter can be achieved by finding a unique Choleski type decomposition of the covariance matrix of the VAR innovations (Rigobon, 2003, Moneta et al., 2013, Gouriéroux et al., 2017, Lanne et al., 2017). However, the time series dynamics of economic and financial data may not be captured well by a linear VAR model when the data is not Gaussian. For example, some variables may only be positive. The problem of estimation is exacerbated if the data have fat tails. This may distort the estimates. Such problems reflect negatively on the estimation of causal relations for time series data. Furthermore, due to the curse of dimensionality issue, SVAR analysis is only feasible in a low-dimensional context. Restricting the VAR model only to a few variables may lead to unreasonable adverse effects such as 'price-puzzles' in impulse responses (Sims, 1992, Christiano et al., 1999, Hanson, 2004). To avoid the curse of dimensionality, factor augmented VAR models (Bernanke et al., 2005) and dynamical factor models (Forni et al., 2000, Forni et al., 2009) are often employed. However, the interpretation of the causal relations with factor models is not always straightforward. Our methodology does not require the machinery of factor models.

This paper builds on a number of previous contributions and develops a methodology to address the aforementioned problems. Our approach is tantamount to the assumption that the cross-sectional and transition distribution of the variables can be represented using a Gaussian copula. The procedure builds on the work of Liu et al. (2012) and does not require us to estimate any transformation of the variables or the marginal distribution of the data, as commonly done when estimating a copula. In fact, our procedure bypasses the estimation of the innovations of the model altogether. Our methodology is built for high dimensional time series, as commonly found in some economics and financial applications. What we require is some form of sparsity in the partial dependence of the data. This is different from assuming that the covariance matrix of innovations or the matrix of autoregressive coefficients are sparse. Such two restrictions can be restrictive. We shall make this clear in the text when we discuss our assumptions. Finally, even when not all causal relations are identified, we are able to identify the largest number of causal relations. This statement is formalized by the concept of complete partially acyclic graph using the PC algorithm (Spirtes et al., 2000, Kalisch and Bühlmann, 2007). These concepts

are reviewed in the main body of the paper (Section 3).

We conclude this introduction with a few remarks whose aim is to put the goals of this paper into a wider perspective. The process of scientific discovery is usually based on 1. the observation of reality, 2. the formulation of a theory, and 3. tests of that the theory. The plethora of data available allows the researcher to observe different aspects of reality that might have been precluded in the past. High dimensional estimation methods are particularly suited to explore the present data-centric reality. However, the next step forward requires formulation of a theory or hypothesis. Such theory needs to be able to explain rather than predict in order to enhance our understanding. This very process requires the identification of a relatively small number of explanatory causes for the phenomenon that we are trying to understand. The problem's solution, in a complex and rather random environment, should then be a simple approximation. This approximation can then be tested in a variety of situations in order to verify its applicability. The program of this paper is to follow this process of scientific discovery. We start from possibly high dimensional dynamic datasets. We aim to provide a reduced set of possible contemporaneous causes conditioning on the past.

1.1 Relation to Other Work

One of the main empirical econometric tools for the study of policy intervention effects is the VAR approach (Sims, 1980, Kilian and Lütkepohl, 2017). In the first step, the so called reduced form model is estimated. Then, the structural counterpart needs to be recovered. This gives rise to an identification problem, which is equivalent to finding the contemporaneous causal relations among the variables.

Traditionally, the identification of Structural Vector Autoregressive (SVAR) models was achieved by imposing model restrictions. Such restrictions can be derived from an underlying economic model, such as short and long-run restrictions on the shocks impact (Bernanke, 1986, Blanchard and Quah, 1989, Faust and Leeper, 1997), or imposing sign restrictions on impulse response functions (Uhlig, 2005, Chari et al. 2008).

The success of the VAR approach is its reliance on data characteristics, thus allowing the validation of economic models under reasonably weak assumptions. However, standard restrictions necessary for the identification invalidate the data-driven nature

of SVAR. In recent years, it has been shown that different statistical features of the data can be exploited to achieve identification of the SVAR model. For instance, identification can be obtained by relying on either heteroskedasticity (Sentana and Fiorentini, 2001, Rigobon, 2003, Lütkepohl and Netšunajev, 2017) or non-Gaussianity of the residuals (Moneta et al., 2013, Gouriéroux et al., 2017, Lanne et al., 2017) or instrumental variables (Mertens and Ravn, 2013, Stock and Watson, 2018). Another approach relies on the graphical causal model literature (Swanson and Granger, 1997, Demiralp and Hoover, 2003, Moneta, 2008). There, identification can be achieved by exploiting the statistical distribution of estimated residuals. We shall show that this last approach is related to our method.

Our work is also related to the statistical and machine learning literature for the identification of causal graph structures in a high dimensional setting (Meinshausen and Bühlmann, 2006, Kalisch and Bühlmann, 2007, Liu et al., 2009, Zhou et al., 2011, Bühlmann et al., 2014). However, these approaches do not account for contemporaneous causal inference conditioning on the past, as required for time series problems.

To account for the time series dependence, we employ a modelling assumption that can be viewed as a Gaussian copula VAR model, a definition that will be made clear in the text. We recently discovered that Fan et al. (2022) have used the same time series assumption for the analysis of high dimensional Granger causality. The present paper is concerned with conditional causal relations and identification of the Gaussian copula VAR. Moreover, some basic assumptions are also different. For example, Fan et al. (2022) assume that the autoregressive matrix of the Gaussian copula VAR is sparse. We instead assume that the inverse of the scaling matrix of the Gaussian copula that leads to a VAR representation is sparse. This is a very different assumption. Hence, the contributions are related, but complementary.

1.2 Outline of the Paper

The plan for the paper is as follows. In Section 2, we introduce the model and briefly discuss its statistical properties. In Section 3 we discuss identification of the model and the causal relations. In Section 4 we describe algorithms to find estimators for the population quantities, including the complete partially acyclic graph. In Section 5 we state conditions and results for the consistency of the quantities derived from the

algorithms. Section 6 applies the methodology to shed light on the causal relations in the order book and trades, in high frequency electronic trading. Section 7 concludes. Proofs and additional details can be found in the Electronic Supplement to this paper. There we also present the main conclusions from a simulation study as evidence of the finite sample properties of our methodology (Section A.3 in the Electronic Supplement).

2 The Model

Let $X := (X_t)_{t\in\mathbb{Z}}$ be a sequence of stationary random variables taking values in $\mathbb{R}^K$ or some subset of it. For each $k = 1, 2, ..., K$, we suppose that there is a monotone function f_k such that $Z_{t,k} = f_k(X_{t,k})$ is a standard Gaussian random variable such that $Z_t = (Z_{t,1}, Z_{t,2}, ..., Z_{t,K})'$

$$Z_t = AZ_{t-1} + \varepsilon_t \tag{1}$$

where A has singular values in $(0, 1)$ and $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a sequence of independent identically distributed random variables with values in $\mathbb{R}^K$ and covariance matrix Σ_ε. Throughout, the prime symbol $'$ denotes transposition. All vectors in the paper are arranged as column vectors. We do not require knowledge of the functions f_k. We also note that there is always a monotone transformation that maps any univariate random variable into a standard Gaussian (Rüschendorf and de Valk, 1993). Hence, the assumption is that such transformed variables satisfy the VAR dynamics in (1). We do not consider higher order VAR models, as these can always be recast into a VAR of order one. Under stationarity assumptions, all the information of the model can be obtained from the covariance matrix of the $2K$-dimensional vector $\left(Z_t', Z_{t-1}'\right)'$, which we denote by Σ. We can then partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Gamma & A\Gamma \\ \Gamma A' & \Gamma \end{pmatrix} \tag{2}$$

with obvious notation, once we note that A is as in (1) and $\Gamma := \mathbb{E}Z_tZ_t'$. Clearly, $\Sigma_\varepsilon := \Gamma - A\Gamma A'$ (recall $\Sigma_\varepsilon := \mathbb{E}\varepsilon_t\varepsilon_t'$).

The above setup can be recast into a formal probabilistic framework using the copula function to model Markov processes (Darsow et al., 1992). The copula transition density would be the ratio of two Gaussian copulae: one with scaling matrix

Σ and one with scaling matrix Γ. Given that we shall not use this in the rest of the paper, we omit the details. However, given this fact, for short, we refer to our model as a Gaussian copula VAR. We note that when X_t has an invariant distribution with marginals that are continuous, the functions f_k are necessarily equal to the unconditional distribution of $X_{t,k}$, by Sklar's Theorem (Joe, 1997).

We consider a high dimensional framework, where K can go to infinity with the sample size. Formally, this would require us to consider a family of models (1) indexed by the sample size n to allow for increasing dimension K (Han and Wu, 2019, for more details). We do not make explicit this in the notation. Next, we summarise the main properties of the model under the possibility that $K \to \infty$.

Proposition 1 *Define $Z_{t,k} = f_k(X_{t,k})$ for some increasing monotonic transformation $f_k : \mathbb{R} \to \mathbb{R}$, $k = 1, 2, ..., K$, such that $(Z_t)_{t\in\mathbb{Z}}$ follows a Gaussian VAR as described in (1). Furthermore, suppose that the singular values of A are in a compact interval inside $(0, 1)$ and the eigenvalues of Σ_ε are in a compact interval inside $(0, \infty)$, uniformly in K. Then, $(X_t)_{t\in\mathbb{Z}}$ is a stationary Markov chain with strong mixing coefficients that decay exponentially fast, uniformly in K even for $K \to \infty$.*

Recall that the singular values of a matrix A are the square root of the eigenvalues of $A'A$. Hence, the condition means that A is full rank with eigenvalues inside the unit circle. We note that for fixed K the model is not only strong mixing, but also absolutely regular (beta mixing), with exponentially decaying coefficients (Doukhan, 1995, Theorem 5, p.97). However, when K is allowed to increase, this is not the case anymore (Han and Wu, 2019, Theorem 3.2). Nevertheless, allowing for increasing dimension K, it is still strong mixing with exponentially decaying coefficients.

3 Identification

3.1 Preliminary Concepts

A graph $G = (\mathcal{V}, \mathcal{E})$ consists of a set of vertices $\mathcal{V} = \{1, 2, ..., p\}$, where p is the number of vertices, and edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The edges are a set of ordered pairs of distinct vertices. The edges are directed if the order matters, $(k, l) \in \mathcal{E}$ but $(l, k) \notin \mathcal{E}$, otherwise it is undirected. Arrows are commonly used to define the direction when there is one. In our context, $\mathcal{V}$ is the set of indices of $W_t = \left(X_t', X_{t-1}'\right)'$, i.e. $p = 2K$,

while $\mathcal{E}$ contains the direction in the causal relations if any. For example, we know that we cannot have $X_{t,i} \to X_{t-1,i}$ while the other way around is possible if $X_{t-1,i}$ Granger causes $X_{t,i}$. In the language of graphs we say that $X_{t-1,i}$ is a parent of $X_{t,i}$. In this paper we focus on the causal relations of X_t conditioning on X_{t-1}. This is different from Granger causality. Given that the statistical relations of the elements in X_t conditioning on X_{t-1} are defined by ε_t, we focus on finding the set of parents of each $\varepsilon_{t,i}$. For example, $\varepsilon_{t,1}$ is a parent of $\varepsilon_{t,2}$ if $\varepsilon_{t,1}$ causes $\varepsilon_{t,2}$ and not the other way around. We write $\varepsilon_{t,1} \to \varepsilon_{t,2}$. When the variables $\varepsilon_{t,k}$ are jointly Gaussian, we immediately see that conditional independence is not enough to identify the direction of the relation (Moneta et al., 2013, Peters et al., 2014).

In the case when all causal relations are identified with no cycles, the causal graph is a directed acyclic graph (DAG): all edges are directed and there are no cycles. There are no cycles if no descendant can be a parent of their ancestor. When the direction cannot be identified we shall content to obtain the undirected edges. The graph with no directions is called the skeleton. When we use observational data, we work with their distribution, possibly under model assumptions as in (1). We say that the distribution of the data is faithful to the graph if the set of all (possibly conditional) independence relations of the distribution of the data and the graph coincide. The (possibly conditional) independence relations of the graph are defined as the set of vertices for which there is no edge between them. Such relations only require to identify the skeleton. Unfortunately, a given distribution of data can generate an infinite number of DAG's. In the case of a VAR this is equivalent to say that the structural VAR cannot be identified. This means that we cannot draw arrows for all edges. A triangular system always allow us to draw edges, but this can be the exception rather than the norm. Hence, we may need to content ourselves with the complete partially directed acyclic graph (CPDAG), which is a graph where some edges are undirected because they cannot be identified.

The PC algorithm (Spirtes et al., 2000) is an algorithm that identifies the skeleton of the graph from conditional independence relations. It then uses some rules to find the edges when identified. The skeleton needs to be estimated when we use observational data (Kalisch and Bühlmann, 2007). For high dimensional time series data, we require special tools as devised in the present paper. Hence, a main goal is to identify the skeleton of ε_t. The first step in this direction is to be able to estimate Σ_ε. The inverse of this matrix plays a special role as it allows us to identify all

the partial regression coefficients. In particular, the set of nonzero entries in row i of the inverse of Σ_ε identifies the neighbours of $\varepsilon_{t,i}$. The set of all neighbours defines the so called moral graph. This is larger than the skeleton as it includes edges between two vertices even when these are unconditionally independent, but conditionally dependent. Such situation arises when there is a so called immorality, e.g. $\varepsilon_{t,1}$ and $\varepsilon_{t,3}$ are unconditionally independent and cause $\varepsilon_{t,2}$. Clearly, conditioning on $\varepsilon_{t,2}$, the variables $\varepsilon_{t,1}$ and $\varepsilon_{t,3}$ are not independent anymore.

3.2 Identification of the Gaussian Copula VAR

We conclude with two results that show the identification strategy in our methodology. We define the precision matrix $\Theta = \Sigma^{-1}$. As we did for Σ in (2), we partition it with same dimensions as in (2):

$$\Theta = \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{pmatrix}. \tag{3}$$

The parameters in (1) are identified from the precision matrix (3). The following, is a consequence of the classical result on graphical Gaussian models (Lauritzen, 1996, eq. C3 and C4).

Lemma 1 *Suppose that the conditions of Proposition 1 hold. Then,* $A = -\Theta_{11}^{-1}\Theta_{12}$ *and* $\Sigma_\varepsilon = \Theta_{11}^{-1}$.

When the DAG is identified, we can identify the SVAR. To this end, we introduce some notation. Let Π be a $K \times K$ matrix that can be transformed into the identity by simple permutation of its rows. We call Π a permutation matrix as it permutes the rows of the conformable matrix that it premultiplies. We have the following result for identification of the SVAR.

Lemma 2 *Suppose that the conditions of Proposition 1 hold and that the causal graph for* ε_t *in (1) is a DAG. Then, we can find a permutation matrix* Π *such that*

$$\Pi Z_t = D\Pi Z_t + (I - D)\,\Pi A Z_{t-1} + \xi_t \tag{4}$$

where D *is lower triangular with diagonal elements equal to zero, and* ξ_t *is a vector of independent Gaussian random variables such that* $\mathbb{E}\xi_t\xi_t'$ *is a diagonal full rank matrix. In particular, the innovation in (4) satisfies* $\Pi\varepsilon_t = H\xi_t$ *where* $H := (I - D)^{-1}$ *is a*

full rank lower triangular matrix with diagonal elements equal to one. Furthermore, the process admits the infinite moving average representation

$$Z_t = \sum_{i=0}^{\infty} \Upsilon_i \xi_{t-i}, \ \textit{where } \Upsilon_i = A^i \Pi^{-1} H. \tag{5}$$

The matrix Υ_i represents the impulse response of Z_t to the shock[1] ξ_{t-i} , $i \geq 0$. The permutation matrix Π can be recovered from the topological order of the contemporaneous causal DAG, where each row of Π identifies an ancestor in its nonzero entry, ordered by "birth". For example, $\varepsilon_{t,3} \to \varepsilon_{t,1} \to \varepsilon_{t,2}$ says that $\varepsilon_{t,3}$ is the first ancestor, $\varepsilon_{t,1}$ is the second ancestor and $\varepsilon_{t,2}$ is the third one. Clearly, $\varepsilon_{t,2}$ has no descendant. Then, the permutation matrix is $\Pi = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, and is unique. When an ancestor has more than one descendant, Π is not unique. The simplest example is $\varepsilon_{t,1} \to \varepsilon_{t,2}$ and $\varepsilon_{t,1} \to \varepsilon_{t,3}$, so that the first variable has two descendants. It is not difficult to see that we have two possible permutation matrices Π because $\varepsilon_{t,1}$ is the first ancestor while $\varepsilon_{t,2}$ and $\varepsilon_{t,3}$ are not ancestors of each other. Hence we can choose to have either $\varepsilon_{t,2}$ or $\varepsilon_{t,3}$ in the second row of Π. One choice is $\Pi = I$, the identity matrix, the second is $\Pi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. In what follows, we shall always refer to the Π matrix as the one that is obtained from the least number of row permutations of the identity matrix. In this case Π is unique. Hence, estimation of the DAG is equivalent to estimation of the permutation matrix Π.

Finally, we remark that a significant instantaneous effect on the Impulse Response Function (IRF) does not provide any information about the true contemporaneous causal structure. This is because correlation does not imply causation. It is easy to construct an example with a causal chain $\varepsilon_1 \to \varepsilon_2 \to \varepsilon_3$ to show that a shock on $\xi_{t,1}$ might instantaneously propagate to $\varepsilon_{t,3}$ even if there is no direct path in the contemporary causal graph. Therefore, the restrictions derived by the causal structure cannot be employed to estimate directly the matrix H from the covariance matrix

[1] We have to permute the vector ξ_t by Π, so that ξ_t and the shocked variables $Z_t$have the same ordering. For instance, if we want to observe the impact of the shock related to the first variable on Z_t we have to consider the vector $\Pi \cdot (1, 0, \cdots, 0)'$, since in the topological order described by Π the first variable of Z_t might be in another position.

of ε_t. Imposing restrictions on H would correspond to a different causal structure with respect to the true one. First, the permutation matrix Π has to be estimated. Then, we can estimate D and recover an estimator for the matrix H of instantaneous effect, as defined in Lemma 2. Next, we introduce algorithms that will be shown to be produce consistent estimators, under regularity conditions.

4 Estimation Algorithms

For any positive integer p, $[p] := \{1, 2, ..., p\}$. For any matrix Q of dimensions $p \times q$ and sets $\mathcal{A} \subseteq [p]$ and $\mathcal{B} \subseteq [q]$, $A_{\mathcal{A},\mathcal{B}}$ is the submatrix with rows in $\mathcal{A}$ and columns in $\mathcal{B}$. In $A_{\mathcal{A},\mathcal{B}}$, when $\mathcal{A} = [p]$ we write $A_{\cdot,\mathcal{B}}$ and similarly if $\mathcal{B} = [q]$. When $\mathcal{A} = [p] \setminus \{i\}$ for some $i \in [p]$, we write $A_{-i,\mathcal{B}}$ and similarly for $\mathcal{B}$. When A is a vector, it is always assumed that it is a column vector and we shall use the same notation, but with one single subscript. This notation will be used throughout the paper with no further mention.

The estimation methodology is based on a number of steps which extend the methodology in Liu et al. (2012). First, we find an estimator of the matrix Σ in (2), which is the Gaussian copula scaling matrix of the vector $W_t = \left(X_t', X_{t-1}'\right)'$. This is achieved using Algorithm 1. Once, the estimator for Σ is available, we identify the set of zero entries in the precision matrix, i.e., the inverse of Σ. This can be achieved using Lasso, as described in Algorithm 2. This algorithm follows the approach of Meinshausen and Bühlmann (2006) to find the zeros in the inverse of (2). However, the algorithm also thresholds the resulting Lasso estimators in order to achieve sign consistency. In this form, the algorithm is equivalent to Gelato (Zhou et al., 2011).

In Algorithm 2, (6) is solved by the x that satisfies the first order conditions in a Lasso minimization problem. The constraint $x_i = 0$ is needed to avoid running the regression of the i^{th} variable on all the other covariates and itself. We need the estimator to be in this form for later use. A competing algorithm to find the zeros of the precision matrix is the CLIME estimation algorithm with thresholding (Cai et al., 2011). The procedure is described in Algorithm 3. The minimization problem in Algorithm 3 can be solved for one column of Ω at the time, with Ω as defined there, due to the use of the uniform norm. We shall show the validity of both algorithms within the time series context of this paper.

Algorithm 4 allows us to estimate the parameters in (1). In particular, it uses the

information on the zeros of the estimator for the precision matrix Θ to construct a sparse estimator (Le and Zhong, 2021). Using Lemma 1, such sparse estimator of the precision matrix is used to estimate the autoregressive matrix A and the covariance matrix of the innovations ε_t in (1).

Finally, using Algorithm 5, we identify the PCDAG. Algorithm 5 makes reference to the PC-algorithm. We do not report the details in Algorithm 5, as the number of steps is relatively large and can be found in Spirtes et al. (2000) among many other places. The aim of the PC-algorithm is to start with a dense graph with undirected edges for all variables. It then aims at removing edges to obtain the skeleton of the graph. Finally, it uses a set of rules to direct all possible edges based on deterministic rules. It is not guaranteed that all edges can be directed, of course.

In order to delete edges, the PC algorithm uses the correlation coefficients between two variables, conditional on subsets of other variables. Note that the innovations in the latent model (1) are Gaussian so that zero correlation implies independence. As soon as we find a set of conditioning variables such that the two variables are conditionally uncorrelated, we remove an edge between these two variables. Given that the conditional correlations are unknown, Kalisch and Bühlmann (2007) suggest to replace these with sample versions as in Algorithm 5. They define a parameter α, as in Algorithm 5, and show that for $\alpha \to 0$ at a certain speed we can obtain a consistent estimator of the PCDAG, as if we knew the true conditional correlations. For this reason, Algorithm 5 only gives details on the sample estimator leaving out the deterministic steps, to avoid distracting details.

Identification of the SVAR requires that all edges are directed. Assuming that Algorithm 5 can direct all the edges, for each $i \in [K]$, we obtain estimators $\hat{\mathcal{V}}(i)$ for the set of parents of $\varepsilon_{t,i}$, using the notation in Algorithm 6. According to Lemma 2, to find the matrix D, we need to find the regression coefficients of the innovation $\varepsilon_{t,i}$ on $\varepsilon_{t,\hat{\mathcal{V}}(i)}$, $i \in [K]$. Algorithm 6 finds such regression coefficients and collects them into a $K \times K$ matrix $\hat{\Delta}$, $i = 1, 2, ..., K$. In particular, the i^{th} row of $\hat{\Delta}$ has entries $\hat{\mathcal{V}}(i)$ equal to the coefficients found regressing $\varepsilon_{t,i}$ on $\varepsilon_{t,\hat{\mathcal{V}}(i)}$ and zeros elsewhere. By the fact that the graph is a DAG, there is a permutation matrix $\hat{\Pi}$ such that $\hat{\Pi}\hat{\Delta}\hat{\Pi}^{-1}$ is an estimator for D and is a lower triangular matrix with zeros along the diagonal. The regression coefficients are obtained relying on $\hat{\Sigma}_\varepsilon := \hat{\Theta}_{11}^{-1}$. This is because $\hat{\Theta}_{11}$ is a sparse estimator with good asymptotic properties. Such properties are inherited by $\hat{\Sigma}_\varepsilon$ even though Σ_ε is not sparse. The estimator $\hat{\Sigma}_\varepsilon$ is not necessarily sparse.

Algorithm 1 Copula Scaling Matrix Estimation.

Define $W_t := \left(X_t', X_{t-1}'\right)'$, $t \in [n]$.
For $1 \leq i < j \leq 2K$:
Let $\hat{\rho}_{i,j}$ be the sample Spearman's rho coefficient between $(W_{s,i})_{s\in[n]}$ and $(W_{s,j})_{s\in[n]}$ (i.e. the sample correlation of their ranks).
Define the $2K \times 2K$ matrix estimator $\hat{\Sigma}$ for (2) with i, j entry $\hat{\Sigma}_{i,j} = 2\sin\left(\frac{\pi}{6}\hat{\rho}_{i,j}\right)$ and set $\hat{\Sigma}_{j,i} = \hat{\Sigma}_{i,j}$.
Ensure that the entries in $\hat{\Sigma}$ corresponding to Σ_{11} and Σ_{22} in (2) are the same by taking averages of the two estimators if needed.

Algorithm 2 High Dimensional Causal Estimation with Lasso. Use Lasso (Meinshausen and Bühlmann, 2006) to find the moral graph of W_t.

Set $\tau > \lambda > 0$.
Run Algorithm 1 to obtain $\hat{\Sigma}$.
For $i \in [K]$:
Denote by $\hat{\beta}^{(i)} \in \mathbb{R}^{2K}$ the solution to

$$\hat{\Sigma}_{\cdot,i} - \hat{\Sigma}x = \lambda \text{sign}\left(x\right), \text{ s.t. } x_i = 0,\ x \in \mathbb{R}^K \tag{6}$$

Redefine $\hat{\beta}_j^{(i)}$ as $\hat{\beta}_j^{(i)} 1_{\left\{\left|\hat{\beta}_j^{(i)}\right| \geq \tau\right\}}$.
Let j be a neighbour of i if $\hat{\beta}_j^{(i)} \neq 0$.
For each $i \in [K]$:
Set $\hat{\Omega}^{(i)}$ equal to $\hat{\beta}^{(i)}$, but let $\hat{\Omega}_i^{(i)} = 1$, where $\hat{\Omega}_i^{(i)}$ is the i^{th} entry.

Moreover, regression coefficients are found directly from $\hat{\Sigma}_\varepsilon$ with no need to estimate the innovations.

The tuning parameters for Algorithms 2 and 3 are chosen using crossvalidation (Section A.2 in the Electronic Supplement, for details).

In the Electronic Supplement, we also use simulations to investigate the finite sample properties of the estimators in our algorithms (see Section A.3 in the Electronic Supplement). There, we also evaluate the performance of the PC algorithm when we use the zeros in $\hat{\Theta}_{11}$ to remove edges from the skeleton with the purpose of skipping some time consuming steps in the PC Algorithm. This considerably reduces the compute time in the high dimensional case. However, when imposing such a priori restrictions, we need to be careful not to increase the possibility of not including an edge that should instead be included (a false negative). The PC algorithm can only delete edges and not add them back. This means that we should undersmooth by

Algorithm 3 High Dimensional Causal Estimation with CLIME. Use CLIME (Cai et al., 2011) to find the moral graph of W_t.

Set $\tau > \lambda > 0$.
Run Algorithm 1 to obtain $\hat{\Sigma}$.
Let $\hat{\Omega} \in \mathbb{R}^{2K \times 2K}$ be the solution to $\min |\Omega|_{1,1}$ s.t. $\left|\hat{\Sigma}\Omega - I\right|_\infty \leq \lambda$.
Redefine $\hat{\Omega}_{i,j}$ as $\hat{\Omega}_{i,j} 1_{\left\{|\hat{\Omega}_{i,j}| \geq \tau\right\}}$ and denote by $\hat{\Omega}^{(i)}$ the i^{th} column of the redefined $\hat{\Omega}$.

Algorithm 4 Estimation of the Parameters in (1).

Run either Algorithm 2 or 3 to find $\hat{\Sigma}$ and $\hat{\Omega}^{(i)}$, $i = 1, 2, ..., 2K$.
Let $\tilde{\Omega}^{(i)}$ be the subvector obtained by deleting the zero elements in $\hat{\Omega}^{(i)}$ and denote by $\hat{s}_i$ its size.
Denote by $\hat{B}_i$ the $2K \times \hat{s}_i$ matrix such that $\hat{\Omega}^{(i)} = \hat{B}_i \tilde{\Omega}^{(i)}$
Define $\hat{\Theta}^{(i)} = \hat{B}_i \left(\hat{B}_i' \hat{\Sigma} \hat{B}_i\right)^{-1} \hat{B}_i' e_i$ where e_i is the $2K \times 1$ vector with i^{th} entry equal to one and zero otherwise.
Let $\hat{\Theta} = \frac{1}{2}\left[\left(\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, ..., \hat{\Theta}^{(2K)}\right) + \left(\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, ..., \hat{\Theta}^{(2K)}\right)'\right]$.
Denote by $\hat{\Theta}_{11}$ the entries (k, l) in $\hat{\Theta}$, $k, l = 1, 2, ..., K$.
Denote by $\hat{\Theta}_{12}$ the entries (k, l) of $\hat{\Theta}$ with $k = 1, 2, ..., K$, and $l = K+1, K+2, ..., 2K$.
Define $\hat{A} = -\hat{\Theta}_{11}^{-1} \hat{\Theta}_{12}$ as an estimator for A in (2).
Define $\hat{\Sigma}_\varepsilon = \hat{\Theta}_{11}^{-1}$ as an estimator for $\Sigma_\varepsilon := \mathbb{E}\varepsilon_t \varepsilon_t'$.

Algorithm 5 Estimation of the PCDAG.

Run Algorithm 4 to find $\hat{\Sigma}_\varepsilon$.
Use $\hat{\Sigma}_\varepsilon$ to find the estimator of the correlation coefficient of $\varepsilon_{t,i}$ and $\varepsilon_{t,j}$ conditioning on $\{\varepsilon_{t,l} : l \in \mathbf{k}\}$ where $\mathbf{k} \subset [K]$ is a set that excludes i, j. Denote such correlation coefficient by $\hat{\Xi}_{i,j|\mathbf{k}}$.
Use the PC-algorithm (Spirtes et al., 2000) and delete a node between (i, j) if $\sqrt{n - |\mathbf{k}| - 3} \times g\left(\hat{\Xi}_{i,j|k}\right) \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ where $g(x) = 2^{-1} \ln\left(\frac{1+x}{1-x}\right)$ $(x \in (-1, 1))$ and $\alpha \in (0, 1)$.

Algorithm 6 Estimation of the impulse response.

Run Algorithm 5 and suppose that the PC algorithms identifies the DAG in the sense that it produces and estimator $\hat{\mathcal{E}} \subseteq \mathcal{V} \times \mathcal{V}$ for the true edges $\mathcal{E}$, such that all elements in $\hat{\mathcal{E}}$ are directed.
For $i \in [K]$:
Find all $j \in \mathcal{V}$ such that $(j, i) \in \hat{\mathcal{E}}$ so that conditioning on the Z_{t-1}, the j covariate is a parent of the i one (i.e. $\varepsilon_{t,j} \to \varepsilon_{t,i}$). Denote such set by $\hat{\mathcal{V}}(i)$.
Find $\hat{d}_i = \hat{\Sigma}^{-1}_{\varepsilon,\hat{\mathcal{V}}(i),\hat{\mathcal{V}}(i)} \hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(i),i}$.
Let $\hat{\Delta}$ be the matrix such that $\hat{\Delta}_{i,\hat{\mathcal{V}}(i)} = \hat{d}_i'$ and zero otherwise.
Find the matrix $\hat{\Pi}$ obtained from the least number of row permutations of the identity matrix and such that $\hat{D} := \hat{\Pi}\hat{\Delta}\hat{\Pi}^{-1}$ is lower diagonal with diagonal elements equal to zero.

choosing tuning parameters that are smaller than the ones obtained by crossvalidation. This produces a $\hat{\Theta}_{11}$ with less zeros. However, it has the negative consequence of increasing the estimation error of $\hat{\Sigma}_\varepsilon := \hat{\Theta}_{11}^{-1}$. In conclusion, as far as sample properties of the estimator are concerned, we found that imposing such restrictions does not improve the performance of the estimator of the PC algorithm, but for some special causal structures.

The simulation analysis show that our approach produces more reliable results than methods that do not account for either sparsity or time series dependence, i.e. setting $\lambda = 0$ in Algorithms 2 and 3 or assuming $A = 0$ in (1). Even when the persistence of the time series is reduced, our methodology produces the best results for estimation of the causal structure and the VAR parameters (for details, see Tables 2-7 in Section A.3 in the Electronic Supplement). Although our approach is designed for a high dimensional setting, it provides competitive results even in the low dimensional case.

5 Asymptotic Analysis of the Algorithms

The consistency of the algorithms relies on a set of conditions. Before introducing our conditions, we introduce some additional notation.

5.1 Additional Notation

For any vector, the ℓ_p norm is denoted by $|\cdot|_p$, $p \in [0, \infty]$. For any $I \times J$ dimensional matrix A, $|A|_{p,q} = \left(\sum_{j=1}^{J} \left(\sum_{i=1}^{I} |A_{i,j}|^p \right)^{q/p} \right)^{1/q}$ is the elementwise norm. When $q = \infty$ we define $|A|_{p,\infty} = \max_{j \leq J} \left(\sum_{i=1}^{I} |A_{i,j}|^p \right)^{1/p}$. When both $p = q = \infty$ we simply write $|A|_\infty = \max_{i \leq I, j \leq J} |A_{i,j}|$, and this should not cause confusion with the ℓ_∞ norm. For $p = 0$, $|A|_{0,\infty} = \max_{j \leq J} \sum_{i=1}^{I} 1_{\{|A_{i,j}|>0\}}$. When $p = q = 0$, this is just the total number of non-zero elements in A. Finally, $|\cdot|_{\text{op}}$ is used to define the following operator norm: $|A|_{\text{op}} = \max_{x: x'x \leq 1} |Ax|_2$. Then, $|A|_{\text{op}}$ is the largest singular value of A. For ease of reference, we call this norm the operator's norm.

Let $\mathcal{U}(\omega, s) = \left\{ \Omega \in \mathbb{R}^{2K \times 2K} : \Omega \succ 0, |\Omega|_{1,\infty} \leq \omega, |\Omega|_{0,\infty} \leq s \right\}$. The symbol $\Omega \succ 0$ is used to mean that Ω is a symmetric strictly positive definite matrix. Then, $\mathcal{U}(\omega, s)$ is the set of symmetric strictly positive definite matrices whose absolute sum of column entries is at most ω, and with maximum number of non-zero entries in each row equals s.

We shorten left and right and side with l.h.s. and r.h.s., respectively. Finally, $\lesssim$ is used when the l.h.s. is bounded above by a constant times the r.h.s.; $\gtrsim$ is bounded below by a constant times the r.h.s.; $\asymp$ is used when the l.h.s. is bounded below and above by constants times the r.h.s.. Finally, to avoid notational trivialities, we assume that $K \geq 2$.

5.2 Regularity Conditions

Assumption 1 *(Model) There are monotone functions f_k such that $Z_{t,k} = f_k(X_{t,k})$ is a standard Gaussian random variable such that (1) holds.*

Assumption 2 *(Dimension) The state space is a subset of $\mathbb{R}^K$, where $K = O(n^{\eta_K})$ for some $\eta_K < \infty$.*

Assumption 3 *(Precision matrix sparsity) The precision matrix $\Theta = \Sigma^{-1}$ is an element of $\mathcal{U}(\omega, s)$ for $s = O(n^{\eta_s})$ for some $\eta_s < 1/2$.*

Assumption 4 *(Identifiability) $\theta_{\min} \gtrsim n^{-\eta_\theta}$, $\eta_\theta < 1/2$, where $\theta_{\min}$ is the smallest absolute value of the nonzero elements in Θ.*

Assumption 5 *(Eigenvalues) The singular values of A are in a compact interval inside $(0,1)$ and the eigenvalues of Σ_ε are in a compact interval inside $(0,\infty)$, uniformly in K.*

Strictly speaking, if $K \to \infty$ as $n \to \infty$, we should index both the process X and its law by n and think in terms of a sequence of processes. We refrain to do so for notational simplicity. No part in the proofs makes implicitly use of assumptions that contradicts this.

5.3 Remarks on the Regularity Conditions

Condition 1. The modelling assumption includes a Gaussian linear vector autoregressive model as special case. However, it is clearly more general than that. Once, we assume that the data satisfy a VAR model after a monotone transformation, we do not need to impose any moment condition on the original data. Hence the procedure is robust to fat tails. As discussed in Section 2, we can view this assumption as a Gaussian copula assumption for the cross-sectional and time series dependence. Condition 1 can be viewed as a generalization of the framework of Liu et al. (2012) in the time series direction and has been recently exploited by Fan et al. (2022) to test for Granger causality in high dimensional models.

Condition 3. The precision matrix is supposed to have maximum absolute sum of each column bounded by a constant ω. Our bounds make explicit the dependence on ω so that we can have $\omega \to \infty$ if needed. This constant is only used in Algorithms 2 and 3. The total number of non zero elements in each row is supposed to be bounded by a constant s. This is allowed to grow to infinity with the sample size at a certain rate. This assumption is different from Fan et al. (2022) who assumes that the autoregressive matrix A in (1) is sparse. This is not the case here. By Lemma 1, sparsity of Θ does not imply sparsity of either A or Σ_ε.

Condition 4. This condition is only used to ensure that we can identify the zero entries in Θ. It is necessary in order to ensure the validity of post selection asymptotic, though the rate can be arbitrarily slow when $\theta_{\min} \to 0$ (Leeb and Pötscher, 2005, p.29ff).

Condition 5. The eigenvalues condition means that the variables are linearly independent in the population. This could be weakened, but at the cost of technical complexity. This condition also implies the following.

Lemma 3 *Under Condition 5 the following statements hold uniformly in K:*

1. *The eigenvalues of $\Gamma = Var(Z_t)$ are bounded away from zero and infinity;*
2. *There are constants $\sigma_{\min}, \sigma_{\max} \in (0, \infty)$ such that the eigenvalues of Σ in (2) are in the interval $[\sigma_{\min}, \sigma_{\max}]$;*
3. *There is a $\nu > 0$ such that $|\Theta_{i,i}| \geq \nu^2$;*
4. *The partial correlations of $\varepsilon_{t,i}$ and $\varepsilon_{t,j}$ conditioning on any other subset of remaining innovations is bounded above by a constant $\bar{\sigma} < 1$.*

5.4 Uniform Convergence of the Scaling Matrix Estimator

The uniform consistency of the covariance estimator from Algorithm 1 is well known (Liu et al., 2012). It is still consistent for dependent data.

Theorem 1 *Under the Regularity Conditions,* $\left|\hat{\Sigma} - \Sigma\right|_\infty = O_P\left(\sqrt{\frac{\ln K}{n}}\right)$.

Fan et al. (2022) show a similar result using Kendall's tau instead of Spearman's rho with a different method of proof.

5.5 Estimation of the Undirected Graph

5.5.1 Consistency for Algorithm 2

The reader is referred to the Regularity Conditions and Algorithm 2 for the notation. Let $\beta^{(i)}$ be the population regression coefficient including a zero in the i^{th} entry, i.e. the solution to $\Sigma_{\cdot,i}x - \Sigma = 0$ s.t. $x_i = 0$.

Theorem 2 *Suppose that the Regularity Conditions hold. There is a finite constant c large enough such that in Algorithm 2, choosing $\lambda = \lambda_n = c\omega\sqrt{\frac{\ln K}{n}}$, with ω is as in Condition 3 we have that* $\max_{i\in[K]} \left|\hat{\beta}^{(i)} - \beta^{(i)}\right|_1 = O_P\left(\omega s\sqrt{\frac{\ln K}{n}}\right)$.

One could choose $c \to \infty$ slowly enough, in which case the bound would be $O_P\left(c \times \omega s\sqrt{\frac{\ln K}{n}}\right)$ instead of $O_P\left(\omega s\sqrt{\frac{\ln K}{n}}\right)$. The proof of this result shows that we could have stated the results as finite sample one with high probability. However, such statement would still depend on an unknown constant. Hence, for simplicity, we have chosen not to do so.

Using appropriate thresholding, with threshold constant greater than the noise level, but smaller than $\theta_{\min}$, the absolute value of the smallest nonzero entry in Θ, leads to set identification. In what follows $\operatorname{sign}(x)$ is the sign of the real variable x with $\operatorname{sign}(0) = 0$.

Theorem 3 *Suppose that the Regularity Conditions hold. In Algorithm 2, set* $\tau = \tau_n = o(\theta_{\min})$ *such that* $\lambda = \lambda_n = o(\tau_n)$ *with* λ *as in Theorem 2. If* $\omega s\sqrt{n^{-1}\ln K} \to 0$, *then,*

$$\Pr\left(\operatorname{sign}\left(\hat{\beta}_j^{(i)}\right) \neq \operatorname{sign}\left(\beta_j^{(i)}\right) \text{ for at least one } i \in [K], j \in [2K]\right) \to 0.$$

5.5.2 Consistency Results for Algorithm 3

The reader is referred to the Regularity Conditions and Algorithm 3 for the notation.

Theorem 4 *Suppose that the Regularity Conditions hold. There is a finite constant* c *large enough such that in Algorithm 3,* $\lambda = \lambda_n = c\omega\sqrt{\frac{\ln K}{n}}$, *where* ω *is as in Condition 3, implies that* $\left|\hat{\Omega} - \Theta\right|_\infty = O_P\left(\omega^2\sqrt{\frac{\ln K}{n}}\right)$.

The same remark we made about c in Theorem 2 applies here. Also here, we could have stated the result as a finite sample one with high probability.

Using the appropriate level of thresholding, Theorem 4 implies the following.

Theorem 5 *Suppose that the Regularity Conditions hold. In Algorithm 3, set* $\tau = \tau_n = o(\theta_{\min})$ *and* $\lambda = \lambda_n = o(\tau_n/\omega)$ *with* λ *as in Theorem 4. If* $\omega^2\sqrt{n^{-1}\ln K} \to 0$, *then,*

$$\Pr\left(\operatorname{sign}\left(\hat{\Omega}_{i,j}\right) \neq \operatorname{sign}(\Theta_{i,j}) \text{ for some } i, j \in [2K]\right) \to 0.$$

5.6 Estimation of the Process Parameters and Causal Graph

In what follows, we suppose that the conditions of either Theorem 3 or Theorem 5 hold, depending on which algorithm is used. For short we generically refer to these

as the Regularity Conditions (λ, τ) as they also involve restrictions on the choice of penalty λ and threshold τ.

5.6.1 Consistency of Precision Matrix Estimation

The estimator for the precision matrix is elementwise uniformly consistent under sparseness conditions.

Theorem 6 *Suppose that the Regularity Conditions* (λ, τ) *hold. Then, the estimator* $\hat{\Theta}$ *from Algorithm 4 satisfies* $\left|\hat{\Theta} - \Theta\right|_\infty = O_P\left(\sqrt{\frac{\ln K}{n}}\right)$.

While the quantity $s = |\Theta|_{0,\infty}$ does not enter the bound, a constraint on its growth rate, as prescribed by Condition 3, is required for Theorem 6 to hold.

5.6.2 Consistency of the Estimators for the Autoregressive Matrix and Innovation Covariance Matrix

Recall that by Lemma 1, using the notation in (1) and (3), $A = -\Theta_{11}^{-1}\Theta_{12}$ and $\Sigma_\varepsilon = \Theta_{11}^{-1}$. Hence, we need consistency of Θ_{12} and the inverse of Θ_{11}, which is the case under sparseness. Recall that $s = |\Theta|_{0,\infty}$ as in Condition 3. We have the following bounds in terms of the operator's norm.

Theorem 7 *Suppose that the Regularity Conditions* (λ, τ) *hold. Then,* $\left|\hat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right|_{\text{op}} = O_P\left(s\sqrt{\frac{\ln K}{n}}\right)$ *and* $\left|\hat{A} - A\right|_{\text{op}} = O_P\left(s\sqrt{\frac{\ln K}{n}}\right)$.

5.6.3 PC-Algorithm

Let $\hat{G}$ be the estimated PCDAG from Algorithm 4 and G the true PCDAG. The next result requires faithfulness of the distribution of the data to the graph, as defined in Section 3.1. In what follows, $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

Theorem 8 *Suppose that the Regularity Conditions* (λ, τ) *hold and that the joint distribution of the innovations* ε_t *in (1) is faithful to the DAG for all* K. *Run the PC algorithm as referenced in Algorithm 5 with* $\alpha = \alpha_n$ *such that* $\alpha_n = 2\left(1 - \Phi\left(n^{1/2}c_n/2\right)\right)$ *for* $c_n \asymp n^{-\eta_c}$ *where* $2\eta_c + 3\eta_s < 1$ *with* η_s *as in Condition 3. Then,* $\Pr\left(\hat{G} \neq G\right) \lesssim n^{-p}$ *for any constant* $p < \infty$.

Theorem 8 says that the estimator for the PCDAG converges to the true one at an arbitrarily fast polynomial rate. This is worse that the exponential rate obtained by Kalisch and Bühlmann (2007) for causal discovery using independent identically distributed data.

5.6.4 Consistency of Impulse Response Function

We show that $\hat{D}$ from Algorithm 6 is consistent for D, with D as in Lemma 2. When the PC-Algorithms in Algorithm 5 produces edges that are all directed, we interpret D to be the one corresponding to the permutation matrix Π that is obtained by the least number of row permutations of the identity. Then, D is unique.

In the following, we state the consistency of $\hat{D}$ for D, and the consistency of an estimator $\hat{H}$ for H, in (5), with convergence rates. We shall denote by κ the maximum number of direct descendants among all parents. It is not difficult to show that this is the same as the maximum number of nonzero elements among the columns of D. Such number is bounded above by s, which corresponds to the maximum number of adjacent variables across all the nodes.

Theorem 9 *Suppose that the Regularity Conditions (λ, τ) hold, that the joint distribution of the innovations ε_t in (1) is faithful to the DAG for all K, and that all the estimated edges resulting from Algorithm 5 are directed. Then, using Algorithm 6, $\left|\hat{D} - D\right|_{\text{op}} = O_P\left(s\sqrt{\frac{\kappa \ln K}{n}}\right)$, where D is as in (4) with Π obtained by the least number of row permutations of the identity. Moreover, we also have that $\hat{H} = \left(I - \hat{D}\right)^{-1}$ satisfies $\left|\hat{H} - H\right|_{\text{op}} = O_P\left(s\sqrt{\frac{\kappa \ln K}{n}}\right)$.*

6 Empirical Application

We apply our methodology to study the causal relations between aggregated order book and trades variables in high frequency electronic trading. We aggregate the information to one minute in order to filter out noise and be able to extract one-minute causal relations. This is different from the analysis of order book tick data which has been studied extensively in the literature (Cont et al., 2014, Kercheval and Zhang, 2015, Sancetta, 2018, Mucciante and Sancetta, 2022a, 2022b). It is well known that market participants look at the order book to extract market information

(MacKenzie, 2017). We want to extract average causal relations. For example, such relations are useful to decide how to place trading orders and understand how on average these affect the order book and prices.

We shall estimate a model with 5 stocks to investigate the direction of information dissemination within each stock, via the order book and trades, as well intra stocks. This requires the estimation of a large dimensional model. Our results will also show how the methodology of this paper allows us to disentangle contemporaneous causal effects from time series effects.

6.1 The Data

We consider four stocks constituents of the S&P500 traded on the NYSE: Amazon (AMZN), Cisco (CSCO), Disney (DIS) and Coca Cola (KO). We also consider the ETF on the S&P500 (SPY). The stock tickers are given inside the parenthesis. The sample period is from 01/March/2019 to 30/April/2019, from 9:30am until 4:30pm on every trading day. The data were collected from the LOBSTER data provider[2] (Huang and Polak, 2011). This is a Level 3 dataset, meaning that it contains all limit orders and cancellations for the first 10 levels of the order book as well as trades, all in a sequential order.

6.2 The Covariates

We construct a set of covariates related to the ones that are commonly found in the studies of high frequency order book and trades. However, we use aggregated data to one minute equally spaced frequency. We do so to reduce noise and to be able to estimate an average propensity of each covariate to cause the other. In particular the covariates are the book imbalance up to ten levels, a geometric average return, and the trade imbalance, often termed order flow imbalance. The covariates are listed in Table 1, where their definition can be found. In Table 1, $\text{Mid} = (\text{AskPrice}_1 + \text{BidPrice}_1)/2$ and LagMid is the Mid from the previous minute bucket, where AskPrice_i is the ask price at level i and similarly for BidPrice_i. The operator $\text{avg}(\cdot)$ takes the data from the same one minute bucket and computes the average value. In case of much market activity, the exchange will use the same timestamp for a number of messages at different levels. In the case of the orderbook, we use the last book snapshot of

[2] https://lobsterdata.com/.

Table 1: List of Covariates. The covariates are listed together with their definition.

Name	Short Name	Definition
Book imbalance at level $i \in [10]$	BookImb_i	$\frac{\text{avg}(\text{BidSize}_i - \text{AskSize}_i)}{\text{avg}(\text{BidSize}_i + \text{AskSize}_i)}$
Return	Ret	$100 \times [\text{avg}(\ln(\text{Mid})) - \text{avg}(\ln(\text{LagMid}))]$
Trade Imbalance	TradeImb	$\frac{\text{avg}(\text{SignedTrdSize})}{\text{avg}(\text{TrdSize})}$

the many with the same time stamp. We do not apply this logic to trades. These covariates are directional ones. For this reason, we have omitted other interesting ones, such as the spread. Moreover, the instruments we use are all very liquid and the spread does not change much in this case.

For ease of reference, in what follows, we shall use the convention of merging the ticker and covariate short name.

6.3 Estimation

We estimate the causal graph using our proposed methodology. We used both Lasso (Algorithm 2) and CLIME (Algorithm 3) for the estimation of the sparse precision matrix. For these algorithms, the penalization parameter λ and the threshold parameter τ were selected using cross-validation (see Section A.2 in the Electronic Supplement for details). We then applied Algorithms 4, 5, and 6 to estimate the Gaussian copula VAR parameters, recover the contemporaneous causal structure and identify the matrix of contemporaneous relations D for estimation of the IRFs. The code to implement the PC-algorithm using the sample correlations and parameter α is available as part of the R-package pcalg `https://cran.r-project.org/web/packages/pcalg/pcalg.pdf`. The PC algorithm was initialized with the restrictions provided by Lasso and CLIME to speed up computations and obtain a more restricted graph. We found that all the edges of the causal graph were directed.

It is well known that subset selection procedures are inherently unstable (Meinshausen and Bühlmann, 2010). For this reason, we resample the data 100 times and carry out the above estimation procedure (Algorithms 1, 2 or 3, 4 and 5) for each

sample. To ensure that we do not alter the time series structure of the data, the resampling was performed to select the days. The total number of days in our sample is 42. Then, to obtain the causal structure G from the PC, we keep the edges selected at least 75% of the times within the 100 resamples. The above procedure can produce cycles so that the graph is no longer a DAG. In this case, we would discard the less frequently observed edge for each cycle, in order to obtain a DAG. However, we remark that cycles were not observed.

Given that our estimated causal graph had no undirected edges, we recovered the matrix of contemporaneous effects D from Algorithm 6 and then the (mixing) matrix $H = (I - D)^{-1}$ necessary for recovering the IRFs (see Section 3.2).

To account for uncertainty in the estimation of IRFs we performed 500 bootstrap sampling conditioning on the moral graph and skeleton obtained from the original sample. This means that we only ran Algorithms 1, 4 and 6 on each sample, using the matrices $\hat{B}_i$ and the sets $\hat{\mathcal{V}}(i)$ estimated on the original sample. Using the 500 samples, we computed the median and the related 95% confidence interval.

6.4 Summary of Results

The results for Lasso and CLIME were very similar. In the interest of space, we report and discuss only the results when Lasso (Algorithm 2) is used as intermediate step, with no further mention. Our results show that the causal structure of the order book of each instrument exhibits a dense network structure. The first level of order book imbalance is a source node for each instrument. This means that it is not contemporaneously caused by any other variable. In general we observe how the causal structure goes from top levels to deeper ones. Usually, the return is affected directly by the deeper levels of the order book imbalance. For all but one instrument (KO), the return is a parent of the trade imbalance variable that happens to be a sink node variable. A sink node variable is a variable that is no parent of other variables. We also observe crosscausal effects across instruments, where in particular we observe how the SPY return is affected by the other returns. The details can be found in Figure 1 that shows the DAG of contemporaneous causal relations obtained from our estimation procedure.

The results from the IRF convey a complementary picture to the DAG, as the two are distinct. However, identification of the SVAR requires identification of a

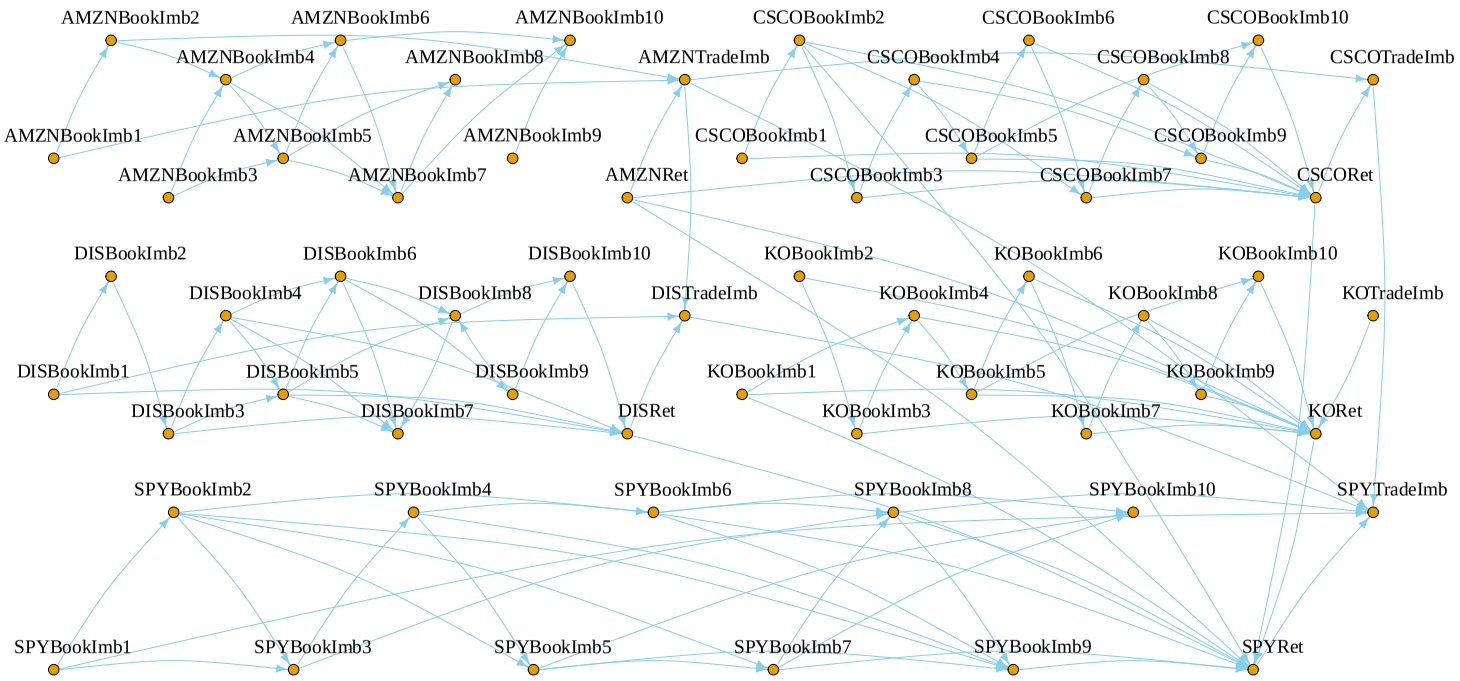


Figure 1: Contemporaneous Causal Graph (DAG). The DAG is obtained using Lasso and then the PC, through the stability selection procedure discussed in the text.

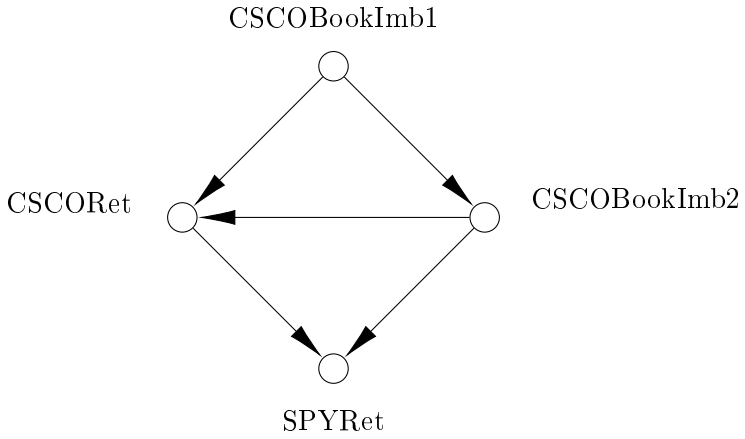


Figure 2: Subgraph of Estimated DAG. The subgraph only considers the contemporaneous causal relations between CSCOBookImb_1, CSCOBookImb_2, CSCORet and SPYRet.

permutation matrix Π (Lemma 2). Such matrix is identified from the DAG. By looking only at the IRFs, one could conclude that a variable contemporaneously affects another. This is because the IRF does not show how the information propagates from one variable to the other at the contemporaneous level. For this reason, we need the causal graph. This point is made vivid by some of our results.

For the sake of definiteness we consider the subgraph composed by CSCOBookImb_1, CSCOBookImb_2, CSCORet and SPYRet as shown in Figure 2. The related IRF's are plotted in Figure 3. From the latter, we observe that a shock on either CSCOBookImb_1 and CSCOBookImb_2 produces an instantaneous effect on both CSCORet and SPYRet. Therefore, by looking only at the IRFs, we can conclude that CSCOBookImb_1 and CSCOBookImb_2 are directly affecting CSCORet and SPYRet. However, this is not the case (see Figure 2). There, we can see that CSCOBookImb_1 and SPYRet are independent when we condition on CSCOBookImb_2 and CSCORet. This means that CSCOBookImb_1 is confounding CSCOBookImb_2 and SPYRet. The information derived from the causal graph makes explicit the difference between the instantaneous effects exhibited in Panel (a) of Figure 3. Any shock to CSCOBookImb_1 will first affect the CSCOBookImb_2 and CSCORet. Hence, SPYRet is only affected through the latter covariates.

From Panel (b) of Figure 3 we observe that, despite the contemporaneous causal relations, SPYRet affects CSCORet with a lagged impact. This effect is also observed for the other instruments: changes in SPYRet affect other variables through its lags.

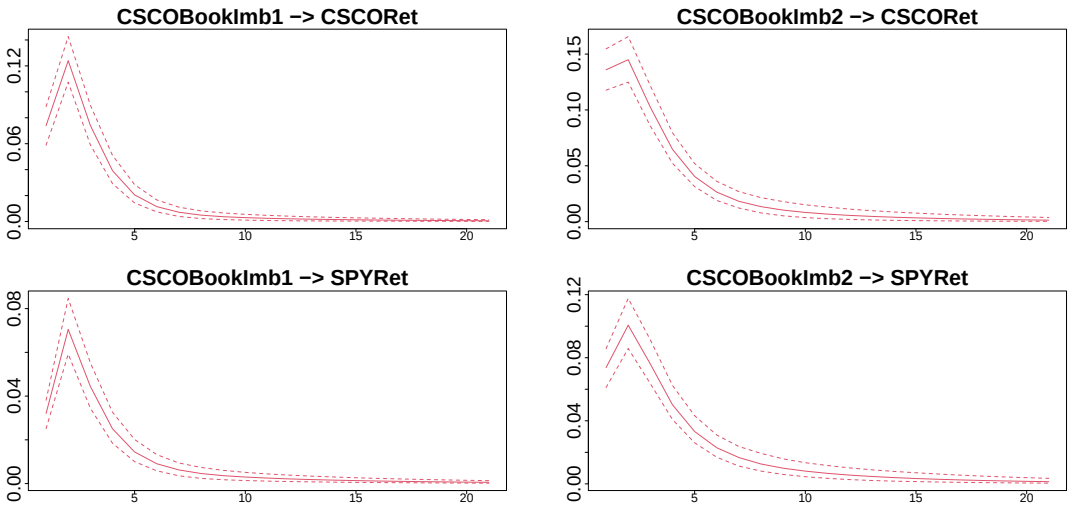
In summary, a thorough analysis of relations between these variables does require to look both at the contemporaneous causal effects via a DAG and the IRF's. The former helps us identify causal effects within simultaneously occurring events. The latter sheds light on the time series propagation of such effects.

7 Conclusion

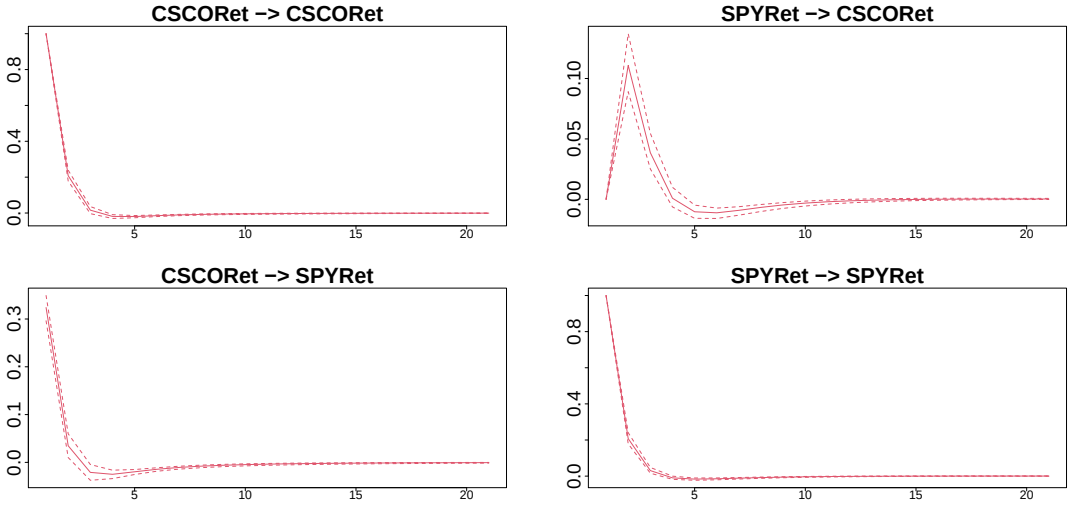
This paper has introduced a novel approach for the estimation of causal relations in time series. It essentially uses a Gaussian copula VAR model. Such causal relations differ from Granger causality. Our methodology, allows us to identify causal relations in high dimensional models. Using a sparsity condition we are able to consistently estimate the model parameters. Our sparsity condition does not imply sparsity of the autoregressive matrix and of the covariance matrix of the innovations implied by the Gaussian copula VAR model. Our sparsity conditions can be viewed as weak assumptions on conditional independence. We are then able to identify the related directed acyclic graph of causal relations, using observational data, as if we knew the true distribution of the data.

Asymptotic results and finite sample investigation confirm the viability of our methodology and its practical usefulness for high dimensional problems. A finite sample analysis, carried out using simulation (Section A.3 in the Electronic Supplement), confirms the asymptotic results of the paper. Moreover, the simulations show that not accounting for time series dependence leads to wrong causal inference. Failing to exploit sparsity leads to suboptimal results, even in low dimensions.

We applied our methodology to the analysis of the conditional contemporaneous causal relations of order book data in high frequency financial data. To the best of our knowledge this has not been done before and has important implications for understanding the aetiology of electronic trading. The shape of the order book appears to be a main causal factor for price changes. The shape of the order book of SPY does not necessarily cause contemporaneous price changes in some of its constituents. Nevertheless, the analysis of the estimated impulse response functions shows that the order book of SPY can have a lagging effect on the price changes of other instruments.



(a)



(b)

Figure 3: IRF's for a Subset of the Covariates. Panel (a) shows the median IRF's (solid line) with its 95% confidence interval (dotted lines) for CSCORet, SPYRet as a result of a unitary shock on CSCOBookImb_1, CSCOBookImb_2. Panel (b) show the same information for CSCORet and SPYRet on each other. The time 1 on the x-axis corresponds to the instantaneous effect of the shock, i.e., it is related to time $t = 0$.

Hence, the approach put forward in this paper allows us to disentangle contemporaneous causality from time series effects. Causal inference and IRF's analysis show in a complementary way the nature of how the information propagates among the variables of a dynamical system accounting for both contemporaneous and time series effects.

References

[1] Acid, S. and L.M. de Campos (2003) Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs. Journal of Artificial Intelligence Research 18, 445–490.

[2] Bernanke, B. (1986) Alternative Explanations of the Money-Income Correlation. In Carnegie-Rochester Conference Series on Public Policy 25, 49-99. North- Holland.

[3] Bernanke, B.S., J. Boivin and P. Eliasz (2005) Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. The Quarterly Journal of Economics 120, 387-422.

[4] Bhatia, R. (1996) Matrix Analysis. New York: Springer.

[5] Blanchard, O. and D. Quah (1989) The Dynamic Effects of Aggregate Demand and Supply Disturbances. American Economic Review 79, 655-673.

[6] Burman, P. and D. Nolan (1992) Data Dependent Estimation of Prediction Functions. Journal of Time Series Analysis 13, 189-207.

[7] Bühlmann, P., J. Peters and J. Ernest (2014) CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression. The Annals of Statistics 42, 2526-2556.

[8] Cai, T., W. Liu and X. Luo (2011) A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. Journal of the American Statistical Association 106, 594-607.

[9] Chari, V., P.J. Kehoe and E.R. McGrattan (2008) Are Structural VARs with Long-Run Restrictions Useful in Developing Business Cycle Theory?. Journal of Monetary Economics 55, 1337-1352.

[10] Christiano, L.J., M. Eichenbaum and C. L. Evans (1999) Monetary Policy Shocks: What Have We Learned and to What End?. Handbook of Macroeconomics 1, 65–148.

[11] Clarke, P.K. (1973) A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. Econometrica 41, 135-155.

[12] Comon, P. (1994) Independent Component Analysis a New Concept?. Signal Processing 36, 287–314.

[13] Cont, R., A. Kukanov and S. Stoikov (2014) The Price Impact of Order Book Events. Journal of Financial Econometrics 12, 47-88.

[14] Darsow, W.F., B. Nguyen and E.T. Olsen (1992) Copulas and Markov processes. Illinois Journal of Mathematics 36, 600-642.

[15] Demiralp, S. and K.D. Hoover (2003) Searching for the Causal Structure of a Vector Autoregression. Oxford Bulletin of Economics and Statistics 65, 745-767.

[16] Doukhan, P. (1995) Mixing. New York: Springer.

[17] Fan Y., F. Han and H. Park (2022) Estimation and Inference in a High-dimensional Semiparametric Gaussian Copula Vector Autoregressive Model. Preprint.

[18] Faust, J. and E.M. Leeper (1997) When Do Long-Run Identifying Restrictions Give Reliable Results?. Journal of Business & Economic Statistics 15, 345-353.

[19] Forni, M., D. Giannone, M. Lippi and L. Reichlin (2009) Opening the Black Box: Structural Factor Models with Large Cross-Sections. Econometric Theory 25, 1319-1347.

[20] Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000) The Generalized Dynamic-Factor Model: Identification and Estimation. Review of Economics and Statistics 82, 540-554.

[21] Gouriéroux, C., A. Monfort and J.-P. Renne (2017) Statistical Inference for Independent Component Analysis: Application to Structural VAR Models. Journal of Econometrics 196, 111-126.

[22] Han, F. (2018) An Exponential Inequality for U-statistics under Mixing Conditions. Journal of Theoretical Probability 31, 556–578.

[23] Han, F. and W.B. Wu (2019) Probability Inequalities for High Dimensional Time Series Under a Triangular Array Framework. https://arxiv.org/abs/1907.06577v1.

[24] Hanson, M. S. (2004) The "Price Puzzle" Reconsidered. Journal of Monetary Economics 51, 1385–1413.

[25] Hyvärinen, A., J. Karhunen and E. Oja (2001) Independent Component Analysis. Wiley, New York.

[26] Hyvärinen, A. and E. Oja (2000) Independent Component Analysis: Algorithms and Applications. Neural Networks 13, 411–430.

[27] Huang, R. and T. Polak (2011) LOBSTER: The Limit Order Book Reconstructor. School of Business and Economics, Humboldt Universität zu Berlin, Techenical Report.

[28] Joe, H. (1997) Multivariate Models and Dependence Models. London: Chapman & Hall.

[29] Kalisch, M. and P. Bühlmann (2007) Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. Journal of Machine Learning Research 8, 613-636.

[30] Kercheval, A.N., Y. Zhang (2015) Modelling High-Frequency Limit Order Book Dynamics with Support Vector Machines. Quantitative Finance 15, 1-15.

[31] Kilian, L. and H. Lütkepohl (2017) Structural Vector Autoregressive Analysis. Cambridge University Press.

[32] Lanne, M., M. Meitz and P. Saikkonen (2017) Identification and Estimation of NonGaussian Structural Vector Autoregressions. Journal of Econometrics 196, 288-304.

[33] Lauritzen, S. L. (1996) Graphical Models. Oxford: Oxford University Press.

[34] Le, T.-M. and P.-S. Zhong (2021) High-Dimensional Precision Matrix Estimation with a Known Graphical Structure. Stat 11, e424.

[35] Leeb, H. and B. M. Pötscher (2005) Model Selection and Inference: Facts and Fiction. Econometric Theory 21, 21-59.

[36] Liu, H., F. Han, M. Yuan, J. Lafferty and L. Wasserman (2012) High Dimensional Semiparametric Gaussian Copula Graphical Models. The Annals of Statistics 40, 2293-2326.

[37] Liu, H., J. Lafferty and L. Wasserman (2009) The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. Journal of Machine Learning Research 10, 2295-2328.

[38] Loh, P.-L. and M. J. Wainwright (2012) High-Dimensional Regression With Noisy and Missing Data: Provable Guarantees With Nonconvexity. The Annals of Statistics 40, 1637-1664.

[39] Lütkepohl, H. and A. Netšunajev (2017) Structural Vector Autoregressions with Heteroskedasticity: A Review of Different Volatility Models. Econometrics and Statistics 1, 2-18.

[40] MacKenzie, D. (2017) A Material Political Economy: Automated Trading Desk and Price Prediction in High - Frequency Trading. Social Studies of Science 47, 172-194 .

[41] Mandelbrot, B. (1963) The Variation of Certain Speculative Prices. Journal of Business 36, 394-419.

[42] Meinshausen, N. and P. Bühlmann (2006) High-Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34, 1436-1462.

[43] Meinshausen, N. and P. Bühlmann (2010) Stability Selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417-473

[44] Mertens, K. and M. O. Ravn (2013) The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States. American Economic Review 103, 1212-47.

[45] Moneta, A. (2008) Graphical Causal Models and VARs: An Empirical Assessment of the Real Business Cycles Hypothesis. Empirical Economics 35, 275-300.

[46] Moneta, A., D. Entner, P. O. Hoyer and A. Coad (2013) Causal Inference by Independent Component Analysis: Theory and Applications. Oxford Bulletin of Economics and Statistics 75, 705-730.

[47] Mucciante, L. and A. Sancetta (2022a) Estimation of a High Dimensional Counting Process Without Penalty for High Frequency Events. Econometric Theory: https://doi.org/10.1017/S0266466622000238.

[48] Mucciante, L. and A. Sancetta (2022b) Estimation of an Order Book Dependent Hawkes Process for Large Datasets. Preprint.

[49] Pearl, J. (2000) Causality: Models, Reasoning, and Inference. Cambridge, UK: Cambridge University Press.

[50] Peters, J., J. M. Mooij, D. Janzing and B. Schölkopf (2014) Causal Discovery with Continuous Additive Noise Models. Journal of Machine Learning Research 15, 2009–2053.

[51] Piterbarg, V.I. (1995) Asymptotic Methods in the Theory of Gaussian Processes and Fields. Providence, RI: American Mathematical Society.

[52] Rigobon, R. (2003) Identification through Heteroskedasticity. The Review of Economics and Statistics 85, 777-792.

[53] Rüschendorf, L. and V. de Valk (1993) On Regression Representation of Stochastic Processes. Stochastic Processes and their Applications 46, 183-198.

[54] Sancetta, A. (2018) Estimation for the Prediction of Point Processes with Many Covariates. Econometric Theory 34, 598-627.89-107.

[55] Sentana, E. and G. Fiorentini (2001) Identification, Estimation and Testing of Conditionally heteroskedastic Factor Models. Journal of Econometrics 102, 143-164.

[56] Shimizu, S., P. O. Hoyer, A. Hyvärinen and A. Kerminen (2006) A Linear Non-Gaussian Acyclic Model for Causal Discovery. Journal of Machine Learning Research 7, 2003–2030.

[57] Sims, C. A. (1980) Macroeconomics and Reality. Econometrica 48, 1-48.

[58] Sims, C. A. (1992) Interpreting the Macroeconomic Time Series Facts: The effects of Monetary Policy. European Economic Review 36, 975–1000.

[59] Spirtes, P., C. Glymour and R. Scheines (2000) Causation, Prediction, and Search. Boston: The MIT Press.

[60] Stock, J. H. and M. W. Watson (2018) Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. The Economic Journal 128, 917-948.

[61] Swanson, N. R. and C.W. Granger (1997) Impulse Response Functions Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions. Journal of the American Statistical Association 92, 357-367.

[62] Tsamardinos, I., L. E. Brown and C. F. Aliferis (2006) The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning 65, 31–78.

[63] Uhlig, H. (2005) What are the Effects of Monetary Policy on Output? Result from an Agnostic Identification procedure. Journal of Monetary Economics 52, 381-419.

[64] van De Geer, S. A. and P. Bühlmann (2009) On the conditions used to prove oracle results for the lasso. Electronic Journal of Statistics 3, 1360-1392.

[65] Zhou, S., P. Rütimann, M. Xu and P. Bühlmann (2011) High-dimensional Covariance Estimation Based On Gaussian Graphical Models. Journal of Machine Learning Research 12, 2975-3026.

Supplementary Material to "Consistent Causal Inference for High Dimensional Time Series" by F. Cordoni and A. Sancetta

A.1 Proofs

Throughout, we use $c_0, c_1, c_2, ...$ to denote constants.

We also recall a property of symmetric strictly positive definite partitioned matrices. Let $\Sigma = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{pmatrix}$ where $A_{i,j}$ $i,j \in \{1,2\}$ is a partition of Σ. Then, $\Sigma^{-1} = \Theta = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}' & B_{22} \end{pmatrix}$ where

$$B_{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}, B_{12} = -B_{11}A_{12}A_{22}^{-1}, B_{22} = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \quad \text{(A.1)}$$

(e.g. Lauritzen, 1996, eq. B.2).

The conclusions from Lemma 3 will be used in a number of places. Hence, we prove this first.

A.1.1 Proof of Lemma 3

We prove one point at the time.

Proof of Point 1. From the condition on A, we have that $Var\left(X_t\right) = \sum_{i=0}^{\infty} A^i \Sigma_\varepsilon \left(A'\right)^i$. We note that

$$\sum_{i=0}^{\infty} \text{eig}_{\min}\left(A^i \Sigma_\varepsilon \left(A'\right)^i\right) \leq \text{eig}_j\left(\sum_{i=0}^{\infty} A^i \Sigma_\varepsilon \left(A'\right)^i\right) \leq \sum_{i=0}^{\infty} \text{eig}_{\max}\left(A^i \Sigma_\varepsilon \left(A'\right)^i\right)$$

$j = 1, 2, ..., K$, where $\text{eig}_j(\cdot)$, $\text{eig}_{\min}(\cdot)$ and $\text{eig}_{\max}(\cdot)$ are the j^{th} eigenvalue, the minimum and the maximum eigenvalue of the argument (Bhatia, 1996, eq. III.13, using induction). Moreover, we have that

$$\text{eig}_{\min}\left(\Sigma_\varepsilon\right) \text{eig}_{\min}\left(A^i \left(A'\right)^i\right) \leq \text{eig}_{\min}\left(A^i \Sigma_\varepsilon \left(A'\right)^i\right)$$

and

$$\text{eig}_{\max}\left(A^i\Sigma_\varepsilon\left(A'\right)^i\right) \leq \text{eig}_{\max}\left(A^i\left(A'\right)^i\right)\text{eig}_{\max}\left(\Sigma_\varepsilon\right).$$

To see this note that

$$\max_{x:x'x=1} x'A\Sigma_\varepsilon A'x \leq \max_{y:y'y=x'A'Ax} y'\Sigma_\varepsilon y = \text{eig}_{\max}\left(A'A\right)\text{eig}_{\max}\left(\Sigma_\varepsilon\right)$$

and similarly for the lower bound and for $i > 1$. Given that the eigenvalues $\text{eig}_j\left(A'A\right)$ are in $(0,1)$ and the eigenvalues $\text{eig}_j\left(\Sigma_\varepsilon\right)$ are in $(0,\infty)$ by assumption, we conclude that the eigenvalues of $Var\left(X_t\right)$ are bounded away from zero and infinity, uniformly in K.

Proof of Point 2. From the definition in (2), we have the following equality,

$$\Sigma = \left[\begin{pmatrix} I & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} + \begin{pmatrix} \mathbf{0} & A \\ A' & \mathbf{0} \end{pmatrix}\right]\begin{pmatrix} \Gamma & \mathbf{0} \\ \mathbf{0} & \Gamma \end{pmatrix},$$

where, here, $\mathbf{0}$ represents a $K \times K$ matrix of zeros. From the assumption on A and the fact that $\Gamma = Var\left(X_t\right)$, we can use the definition of eigenvalues and, mutatis mutandis, the previous inequalities, from the proof of Point 1, to deduce the result.

Proof of Point 3. From (A.1) and the definition of Σ as variance of $\left(Z_t', Z_{t-1}'\right)'$, we deduce that the (i,i) element in Θ_{11} is the inverse of the variance of $Z_{t,i}$ conditioning on $Z_{t-1,i}$, all the other variables and their first lag. Given that the eigenvalues of Σ are bounded away from zero, uniformly in K, the random variables are not perfectly correlated. Hence there must be a constant $\nu > 0$ as in the statement of the lemma.

Proof of Point 4. The eigenvalues of Σ_ε are in some compact interval inside $(0,\infty)$, uniformly in K, by assumption. Hence, the innovation vector has entries that are not perfectly dependent. This means that no conditional correlation between any two variables can be equal to one, uniformly in K.

A.1.2 Proof of Proposition 1

It is clear that the process X is a stationary Markov chain. The mixing coefficients are invariant of monotone transformations of the random variables. Hence, we can

consider the mixing coefficients of Z in (1). For the Gaussian VAR model in (1), Theorem 3.1 in Han and Wu (2019) says that the strong mixing coefficient $\alpha(k)$ for variables k periods apart satisfies $\alpha(k) \leq c|A|_{\text{op}}^k$ where c is the square root of the ratio between the largest and smallest eigenvalue of $Var(Z_t)$. This ratio is bounded by Lemma 3. On the other hand, $|A|_{\text{op}}$ is the largest singular value of A, which is smaller than one, uniformly in K, by assumption. Hence, the strong mixing coefficients decay exponentially fast.

A.1.3 Proof of Lemmas 1 and 2

The conditions in Proposition 1 ensure that the model is stationary. We use this with no explicit mention in the following.

A.1.3.1 Proof of Lemma 1

This follows from (2) and Lauritzen (1996, eq. C3-C4) or from (A.1).

A.1.3.2 Proof of Lemma 2

By the assumption of the lemma, all edges of the graph of ε_t are directed. There are also no cycles. Hence, there must be a permutation matrix Π of the elements in ε_t such that the i element in $\Pi\varepsilon_t$ is not a parent of the $i-1$ element. This implies the structure $\Pi\varepsilon_t = H\xi_t$ where H is a lower triangular matrix with diagonal entries equal to one. Note that H can have diagonal elements equal to one because we are not assuming that $\mathbb{E}\xi_t\xi_t'$ is the identity. The fact that the graph is acyclic means that H is full rank. Otherwise, we would have a descendant that is an ancestor of itself. Now note that the inverse of a lower triangular matrix is also lower triangular. Moreover, if the matrix has diagonal elements equal to one, also the inverse has diagonal elements equal to one. Hence, we can write $H^{-1} = I - D$ where D is as in the statement of the lemma and obtain (4). To find the infinite moving average representation, rewrite (4) as $H^{-1}\Pi(I - AL)Z_t = \xi_t$ where, here, L is the lag operator. By assumption, $(I - AL)$ can be inverted and has an infinite convergent series representation. Hence, we deduce (5) by standard algebra and the aforementioned remarks on H.

A.1.4 Exponential Inequality for Spearman's Rho

Given that Spearman's rho is invariant of monotonically increasing transformations, within our framework, we may consider variables that have been transformed into Gaussian. The following, which is a special case of Theorem 1.5 in Piterbarg (1995), will be useful to bound functions of Gaussian random vectors.

Lemma 4 *Suppose that X and Y are $p \times 1$ mean zero Gaussian random vectors with covariance matrices Σ_X and Σ_Y, respectively. Suppose that the eigenvalues of such matrices are in some compact interval inside $(0, \infty)$. Let h be a bounded function on $\mathbb{R}^p$. Then, there is a finite constant c such that*

$$\left|\mathbb{E}h\left(X\right) - \mathbb{E}h\left(Y\right)\right| \leq c \sum_{i,j} \left|\Sigma_{X,i,j} - \Sigma_{Y,i,j}\right|.$$

With the help of Lemma 4, we bound the bias that arises from using dependent data in the calculation of a U-statistic closely related to Spearman's rho.

Lemma 5 *Let $Z := (Z_t)_{t\in\mathbb{Z}}$ be a sequence of 2×1 dimensional stationary Gaussian random variables with mean zero and variance one. Suppose that its 2×2 autocovariance function (ACF) is full rank for any lag value, and has elements that are absolutely summable w.r.t. the lag value. Let $\tilde{Z} := \left(\tilde{Z}_t\right)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. random variables such that $\tilde{Z}_1$ has same distribution as Z_1. For any sequence of 2×1 dimensional stationary random variables $(X_t)_{t\in\mathbb{Z}}$ define*

$$\rho_3\left(X_1, X_2, ..., X_n\right) := \frac{3}{n\left(n-1\right)\left(n-2\right)} \sum_{t_1\neq t_2\neq t_3} \operatorname{sign}\left(X_{t_1,1} - X_{t_2,1}\right) \operatorname{sign}\left(X_{t_1,2} - X_{t_3,2}\right).$$

Then, there is a finite constant c_Z such that

$$\left|\mathbb{E}\rho_3\left(Z_1, Z_2, ..., Z_n\right) - \mathbb{E}\rho_3\left(\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_n\right)\right| \leq c_Z/n.$$

Proof. We shall first bound the expectation of the summand under two different expectations. With some abuse of notation, let $\Gamma\left(k\right)$ be the 2×2 ACF of Z at lag k. (In the text we have been using Γ to denote $Var\left(Z_t\right)$, which here we shall denote by $\Gamma\left(0\right)$.) By assumption, we have that $Var\left(\tilde{Z}_t\right) = Var\left(Z_t\right) = \Gamma\left(0\right)$ and $\Gamma_{1,1}\left(0\right) = \Gamma_{2,2}\left(0\right) = 1$ because the variables have variance one. We shall use this fact

momentarily. Let $U := (Z_{t_1,1}, Z_{t_2,1}, Z_{t_1,2}, Z_{t_3,2})'$ and $\tilde{U} := \left(\tilde{Z}_{t_1,1}, \tilde{Z}_{t_2,1}, \tilde{Z}_{t_1,2}, \tilde{Z}_{t_3,2}\right)'$ and let $\Sigma_U := Var(U)$ and $\Sigma_{\tilde{U}} := Var\left(\tilde{U}\right)$. The covariance matrices are functions of the ACF Γ. Define $V = RU$ where $R = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$. The variable V is Gaussian with mean zero and variance $R\Sigma_U R'$. Define $k := t_1 - t_2$, $l := t_1 - t_3$ and $v := t_2 - t_3$. By direct calculation, we find that

$$R\Sigma_U R' = \begin{pmatrix} 2 - 2\Gamma_{1,1}(k), & \Gamma_{1,2}(0) + \Gamma_{1,2}(v) - \Gamma_{1,2}(-k) - \Gamma_{1,2}(l) \\ \Gamma_{1,2}(0) + \Gamma_{1,2}(v) - \Gamma_{1,2}(-k) - \Gamma_{1,2}(l) & 2 - 2\Gamma_{2,2}(l) \end{pmatrix}.$$

On the other hand $R\Sigma_{\tilde{U}} R'$ is as in the above display with $\Gamma_{1,1}(i) = \Gamma_{2,2}(i) = \Gamma_{1,2}(j) = 0$ for $i, j \neq 0$.

Now, note that

$$\operatorname{sign}(V_1)\operatorname{sign}(V_2) = \operatorname{sign}(Z_{t_1,1} - Z_{t_2,1})\operatorname{sign}(Z_{t_1,2} - Z_{t_3,2})$$

using the symmetry properties of mean zero Gaussian random variables. Moreover, $\operatorname{sign}(x) = 1_{\{x>0\}} - 1_{\{x<0\}}$, where $1_{\{\cdot\}}$ is the indicator function. Hence,

$$\mathbb{E}\operatorname{sign}(V_1)\operatorname{sign}(V_2) = 2\Pr(V_1 < 0, V_2 < 0) - 2\Pr(V_1 < 0, V_2 > 0).$$

Moreover, $\Pr(V_1 < 0, V_2 > 0) = 1/2 - \Pr(V_1 < 0, V_2 < 0)$ by standard set manipulation and using the fact that $\Pr(V_1 < 0) = 1/2$ because V_1 is Gaussian with mean zero. Hence, we deduce that

$$\left|\mathbb{E}\operatorname{sign}(V_1)\operatorname{sign}(V_2) - \mathbb{E}\operatorname{sign}\left(\tilde{V}_1\right)\operatorname{sign}\left(\tilde{V}_2\right)\right| = 4\left|\Pr(V_1 < 0, V_2 < 0) - \Pr\left(\tilde{V}_1 < 0, \tilde{V}_2 < 0\right)\right|.$$

To bound the r.h.s. we shall use Lemma 4 with function $h(x) = 1_{\{x_1<0\}}1_{\{x_2<0\}}$. We also note that the assumption of that lemma on the eigenvalues is satisfied because $R\Sigma_U R'$ and $R\Sigma_{\tilde{U}} R'$ are full rank and have bounded maximum eigenvalue. By assumption, we can also deduce that there is a function γ such that $\Gamma_{i,j}(k) \leq \gamma(k)$ for $i, j = 1, 2$ and all integers k and $\sum_k \gamma(k) \leq \bar{\gamma} < \infty$. Hence, by Lemma 4, there is a

constant c (the same as in Lemma 4) such that

$$\begin{aligned}&\left|\mathbb{E}\rho_3\left(Z_1, Z_2, ..., Z_n\right) - \mathbb{E}\rho_3\left(\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_n\right)\right| \\ &\leq \frac{12c}{n\left(n-1\right)\left(n-2\right)} \sum_{t_1 \neq t_2 \neq t_3} \left[\gamma\left(t_1 - t_2\right) + \gamma\left(t_1 - t_3\right) + \gamma\left(t_2 - t_3\right)\right].\end{aligned}$$

By summability of $\gamma\left(k\right)$ w.r.t. $k \in \mathbb{Z}$ we deduce the result, where the constant c_Z used in the statement of the lemma can be chosen equal to $36c\bar{\gamma}$. ■

The following is a rephrasing of Theorem 3.1 in Han (2018) where we have added the bias that results from the use of dependent data (see Han, 2018, eq. 3.1).

Lemma 6 *Let $X := \left(X_t\right)_{t\in\mathbb{Z}}$ be a sequence of stationary random variables, possibly vector valued, with exponentially decaying strong mixing coefficients. Let $\tilde{X} := \left(\tilde{X}_t\right)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. random variables such that X_1 and $\tilde{X}_1$ have same distribution. Suppose that $\rho\left(X_1, X_2, ..., X_n\right)$ is a U-statistic of finite order with kernel bounded by one. Define*

$$\text{bias} := \left|\mathbb{E}\rho\left(X_1, X_2, ..., X_n\right) - \mathbb{E}\rho\left(\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n\right)\right| \tag{A.2}$$

Then, there is a strictly positive constant c such that for any $x > 0$,

$$\Pr\left(\left|\rho\left(X_1, X_2, ..., X_n\right) - \mathbb{E}\rho\left(\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n\right)\right| \geq x + \text{bias}\right) \leq \exp\left\{-\frac{cnx^2}{1 + x\left(\ln n\right)\left(\ln\ln\left(4n\right)\right)}\right\}.$$

The proof of the above inequality in Han (2018) uses the strong mixing condition to bound the estimation error (Theorem 3.1 in Han, 2018). Theorem 2.1 in Han (2018) uses the beta mixing condition to bound the bias (A.2). We shall exploit the structure of Spearman's rho and use Lemma 5, instead.

The definition of the population version of Spearman's rho (e.g., Joe, 1997, p.32) between two random variables with joint distribution F_{XY} and marginals F_X and F_Y is $\rho = 12\int\int F_X\left(x\right)F_Y\left(x\right)dF_{XY}\left(x,y\right) - 3$. It is not difficult to deduce that $\mathbb{E}\rho_3\left(\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_n\right)$ in Lemma 5 is the Spearman's rho population coefficient of $\tilde{Z}$. We shall denote by $\rho_{i,j}$ the Spearman's rho population coefficient (the rank correlation) between the random variables $W_{t,i}$ and $W_{t,j}$ in our dataset used in Algorithm 1 (recall that $W_t := \left(X_t', X_{t-1}'\right)'$. Then, we have the following.

Lemma 7 *Under the Regularity Conditions, for $\hat{\rho}_{i,j}$ as in Algorithm 1,*

$$\Pr\left(\max_{i,j\leq K}|\hat{\rho}_{i,j}-\rho_{i,j}|\geq x+c_1 n^{-1}\right)\leq K^2\exp\left\{-\frac{c_2 n x^2}{1+x\left(\ln n\right)\left(\ln\ln\left(4n\right)\right)}\right\}.$$

Here, c_1 and c_2 are absolute constants.

Proof. At first, we note that $\hat{\rho}_{i,j}=\frac{n-2}{n+1}\rho_{3,i.j}+\frac{3}{n+1}\rho_{\tau,i,j}$ where

$$\rho_{3,i,j}=\rho_3\left(\left(X_{1,i},X_{1,j}\right),\left(X_{2,i},X_{2,j}\right),...,\left(X_{n,i},X_{n,j}\right)\right)$$

is the U-statistic ρ_3 from Lemma 5, while $\rho_{\tau,i,j}$ is the sample Kendall's tau between the i and j variables (Han, 2018, eq. 2.4). By the invariance of Spearman's rho under monotonically increasing transformations, we can replace the observable $X:=(X_t)_{t\in\mathbb{Z}}$ with the unobservable $Z:=(Z_t)_{t\in\mathbb{Z}}$, which is Gaussian with dynamics as in (1). The ACF of the VAR model in (1) has entries that are all absolutely summable by the Regularity Conditions on A. Hence, by Lemma 5, (A.2) is bounded above by some constant times n^{-1}. Noting that $\frac{n-2}{n+1}\rho_{3,i.j}=\rho_{3,i.j}-\frac{3}{n+1}\rho_{3,i.j}$ and that $\left|\frac{3}{n+1}\rho_{3,i,j}\right|+\left|\frac{3}{n+1}\rho_{\tau,i,j}\right|\leq 6/n$, we can find a finite constant c_1 such that Lemma 6 applies with bias replaced by a c_1/n. Applying the union bound, we deduce the statement of the lemma. ■

A.1.5 Lemmas on Control of the Sample Covariance Estimator and Related Quantities

Recall that $\rho_{i,j}$ is the rank correlation between $W_{t,i}$ and $W_{t,j}$. By stationarity, this does not depend on t. We have the following.

Lemma 8 *Under the Regularity Conditions, for n large enough, there is a finite constant c_0 such that*

$$\Pr\left(\max_{i,j\leq K}|\hat{\rho}_{i,j}-\rho_{i,j}|\geq c_0\sqrt{\frac{\ln K}{n}}\right)\leq K^{-1}.$$

Proof. This follows from the inequality in Lemma 7. There, we set $x^2=$

$5\ln(K)/(c_2 n)$ to deduce that for $c_0 = \left(\sqrt{5/c_2} + c_1\right)$,

$$\Pr\left(\max_{i,j\leq K} |\hat{\rho}_{i,j} - \rho_{i,j}| \geq c_0\sqrt{\frac{\ln K}{n}}\right) \leq \exp\left\{-\frac{5(\ln K) - 2(1+\epsilon)\ln K}{1+\epsilon}\right\}$$

for $\epsilon = \sqrt{5\ln K/(c_2 n)}\,(\ln n)(\ln\ln(4n))$. Under the conditions of the lemma, for n large enough, $\epsilon \leq 1$. Substituting in the above display we find that the r.h.s. is bounded above by K^{-1} and this proves the lemma. ■

We now show that the correlation matrix obtained from Spearman's rho converges.

Lemma 9 *Under the Regularity Conditions, for n large enough, there is a constant c_0 (the same as in Lemma 8), such that,*

$$\Pr\left(\max_{i,j\leq K}\left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| \geq \frac{3c_0}{\pi}\sqrt{\frac{\ln K}{n}}\right) \leq K^{-1}.$$

Proof. Recalling the link between Spreaman's rho and the correlation coefficient for the Gaussian copula (e.g. Liu et al., 2012), we have that $\hat{\Sigma}_{i,j} - \Sigma_{i,j} = 2\sin\left(\frac{\pi}{6}\hat{\rho}_{i,j}\right) - 2\sin\left(\frac{\pi}{6}\rho_{i,j}\right)$. Given that $\sin(x)$ is Lipschitz with constant one, the result follows from Lemma 8. ■

Lemma 10 *Suppose that the Regularity Conditions hold. Then, there is a constant $c_3 > 0$, such that, for n large enough,*

$$\max_{i,j\leq K}\Pr\left(\left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| \geq z\right) \leq \exp\left\{-nc_3 z^2\right\}$$

for any z satisfying $zn^{1/2} \to \infty$ and $z(\ln n)(\ln\ln n) \to 0$.

Proof. By the proof of Lemmas 8 and 9,

$$\Pr\left(\left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| \geq \frac{3}{\pi}\left(x + c_1 n^{-1/2}\right)\right) \leq \exp\left\{-\frac{nc_2 x^2}{1 + x(\ln n)(\ln\ln(4n))}\right\}$$

where the constants are from those lemmas. Set $z = \frac{3}{\pi}\left(x + c_1 n^{-1/2}\right)$. Then, $x = (\pi/3)\,z - c_1 n^{-1/2}$. Substituting in the above display, the probability is bounded above by

$$\exp\left\{-\frac{nc_2\left[(\pi/3)\,z - c_1 n^{-1/2}\right]^2}{1+\epsilon}\right\}$$

where $\epsilon = \left[(\pi/3)\, z - c_1 n^{-1/2}\right] (\ln n)\, (\ln \ln (4n))$. By the restriction on z, as in the statement of the lemma, for n large enough, we have that $\epsilon \leq 1$, and that there is a constant $c_3 > 0$ such that the above display is less than $\exp\{-nc_3 z^2\}$. ■

A.1.6 Lemmas for the Control of the Precision Matrix Estimator

The following result for the control of the operator norm will be used in the proofs.

Lemma 11 *Suppose that $\hat{Q}$ and Q are symmetric matrices such that Q has eigenvalues bounded away from zero an infinity. If $\left|\hat{Q} - Q\right|_{\text{op}} = \epsilon$, then $\left|\hat{Q}^{-1} - Q^{-1}\right|_{\text{op}} = O\left(|Q^{-1}|_{\text{op}}^2\, \epsilon\right)$ as long as $|Q^{-1}|_{\text{op}} < \epsilon^{-1}$.*

Proof. With the present notation, Lemma 4 Le and Zhong (2021) says that

$$\left|\hat{Q}^{-1} - Q^{-1}\right|_{\text{op}} \leq \left|Q^{-1}\right|_{\text{op}} \frac{\left|Q^{-1}\left(\hat{Q} - Q\right)\right|_{\text{op}}}{1 - \left|Q^{-1}\left(\hat{Q} - Q\right)\right|_{\text{op}}}. \tag{A.3}$$

Then, the result follows from the fact that $\left|Q^{-1}\left(\hat{Q} - Q\right)\right|_{\text{op}} \leq |Q^{-1}|_{\text{op}} \left|\hat{Q} - Q\right|_{\text{op}}$ together with the condition of the lemma to ensure that the denominator is greater than zero. ■

The operator norm can be bounded by the uniform norm of the elements using the following.

Lemma 12 *Suppose that $\hat{Q}$ and Q are symmetric matrices. Then, $\left|\hat{Q} - Q\right|_{\text{op}} \leq \left|\hat{Q} - Q\right|_{0,\infty} \left|\hat{Q} - Q\right|_{\infty}$.*

Proof. First, note that $\left|\hat{Q} - Q\right|_{\text{op}} \leq \left|\hat{Q} - Q\right|_{1,\infty}$ because $\hat{Q} - Q$ is symmetric. This is well known because, for any matrix A (not to be confused with the autoregressive matrix in (1)), $A'Ax = \sigma^2 x$ where σ^2 is the maximum eigenvalue of $A'A$ and x is the corresponding eigenvector. Hence, $\sigma^2 |x|_{\infty} = |A'Ax|_{\infty}$. By a special case of Holder inequality, $|A'Ax|_{\infty} \leq |A'|_{\infty,1} |A|_{\infty,1} |x|_{\infty}$. This implies that $\sigma^2 = |A|_{\text{op}}^2 \leq |A|_{1,\infty} |A|_{\infty,1}$. Then, using the fact that, in our case, $A = \hat{Q} - Q$ is symmetric, we deduce the inequality at the start of the proof. Moreover, $\left|\hat{Q} - Q\right|_{1,\infty} \leq \left|\hat{Q} - Q\right|_{0,\infty} \left|\hat{Q} - Q\right|_{\infty}$

because $\left|\hat{Q}-Q\right|_{0,\infty}$ is the maximum number of nonzero elements across the columns of $\hat{Q}-Q$. ■

Define the event

$$E := \left\{1_{\{\hat{\Theta}_{i,j}>0\}} = 1_{\{\Theta_{i,j}>0\}}\right\} \tag{A.4}$$

We shall derive a number of results conditional on such event. The event E means that $\left\{\hat{B}_i : i \in [2K]\right\}$ in Algorithm 4 correctly identifies all the nonzero entries in Θ. The next result can be found in the proof of Theorem 3 in Le and Zhong (2021).

Lemma 13 *Suppose that the Regularity Conditions hold. On the event (A.4), there is a constant c_4 such that*

$$\Pr\left(\left|\hat{\Theta}-\Theta\right|_\infty \geq z\right) \leq 2K \Pr\left(\left|\hat{\Sigma}-\Sigma\right|_\infty \geq z c_4\right). \tag{A.5}$$

We can now use the lemmas from Section A.1.5.

Lemma 14 *Suppose that the Regularity Conditions hold. On the event (A.4), there is a constant $c_5 > 0$, such that, for n large enough,*

$$\Pr\left(\left|\hat{\Theta}-\Theta\right|_\infty \geq z\right) \leq 2\exp\left\{-nc_5 z^2 + 3\ln K\right\}$$

for any z satisfying $zn^{1/2} \to \infty$ and $z(\ln n)(\ln\ln n) \to 0$. Moreover, $\left|\hat{\Theta}-\Theta\right|_\infty = O_P\left(\sqrt{\frac{\ln K}{n}}\right)$.

Proof. We bound the r.h.s. in the display of Lemma 13 using Lemma 10 and the union bound. We can then deduce that the r.h.s. of (A.5) is bounded above by $2K^3\exp\{-nc_3c_4^2z^2\}$. Defining $c_5 := c_3c_4^2$ and rearranging we deduce the first statement. The second statement follows by choosing z large enough and proportional to a quantity $O\left(\sqrt{\frac{\ln K}{n}}\right)$ so that the first statement immediately gives that $\left|\hat{\Theta}-\Theta\right|_\infty = O_P\left(\sqrt{\frac{\ln K}{n}}\right)$. Such choice of z is consistent with the constraint given in the lemma. ■

We also need an exponential inequality for $\hat{\Theta}_{11}^{-1} - \Theta_{11}^{-1}$. For simplicity, we state the result for $\hat{\Theta}^{-1}$ rather than $\hat{\Theta}_{11}^{-1}$.

Lemma 15 *Suppose that the Regularity Conditions hold and that* $s\sqrt{\ln K/n} = o(1)$. *On the event (A.4), there is a constant* $c_6 > 0$ *such that, for* n *large enough,*

$$\Pr\left(\left|\hat{\Theta}^{-1} - \Theta^{-1}\right|_{\infty} \geq z\right) \leq 2\exp\left\{-ns^{-2}c_6 z^2 + 3\ln K\right\}$$

for any z *satisfying* $zn^{1/2} \to \infty$ *and* $z(\ln n)(\ln\ln n) \to 0$.

Proof. First, we note that for any symmetric matrix Q, $|Q|_{\infty} \leq |Q|_{\text{op}}$. This is because $|Q|_{\text{op}} = \max_{x,y} x'Qy$ where the maximum is over vectors with unit Euclidean norm. By this remark and (A.3) we deduce that the set $\left\{\left|\hat{\Theta}^{-1} - \Theta^{-1}\right|_{\infty} \geq z\right\}$ is contained in the set

$$\left\{ \left|\Theta^{-1}\right|_{\text{op}} \frac{\left|\Theta^{-1}\left(\hat{\Theta} - \Theta\right)\right|_{\text{op}}}{1 - \left|\Theta^{-1}\left(\hat{\Theta} - \Theta\right)\right|_{\text{op}}} \geq z \right\}.$$

For arbitrary events A and B, we shall use the trivial decomposition $A = \{A \cap B\} \cup \{A \cap B^c\} \subseteq \{A \cap B\} \cup B^c$, where B^c is the complement of B. Then, we deduce that the event in the above display is contained in the event

$$\left\{\left|\Theta^{-1}\left(\hat{\Theta} - \Theta\right)\right|_{\text{op}} \geq 1/2\right\} \cup \left\{\left|\Theta^{-1}\right|_{\text{op}} \left|\Theta^{-1}\left(\hat{\Theta} - \Theta\right)\right|_{\text{op}} \geq z/2\right\} \tag{A.6}$$

For $z/\left|\Theta^{-1}\right|_{\text{op}} \to 0$, the above union of two events is contained in the second event. This is the case because the eigenvalues of Θ are bounded away from zero and infinity by Lemma 3. Hence, it is sufficient to bound the latter. Using a standard inequality for operator norms, and then Lemma 12, we deduce that

$$\left|\Theta^{-1}\left(\hat{\Theta} - \Theta\right)\right|_{\text{op}} \leq \left|\Theta^{-1}\right|_{\text{op}} \left|\left(\hat{\Theta} - \Theta\right)\right|_{0,\infty} \left|\left(\hat{\Theta} - \Theta\right)\right|_{\infty}.$$

On the event E in (A.4), $\left|\left(\hat{\Theta} - \Theta\right)\right|_{0,\infty} \leq |\Theta|_{0,\infty} \leq s$. We assume E holds without making it explicit in the notation. In consequence, recalling that, by Lemma 3, $\sigma_{\max}$ is the largest singular value of $\Theta^{-1} = \Sigma$, which is bounded uniformly in K, we have that

$$\Pr\left(\left|\Theta^{-1}\right|_{\text{op}} \left|\Theta^{-1}\left(\hat{\Theta} - \Theta\right)\right|_{\text{op}} \geq z/2\right) \leq \Pr\left(\left|\left(\hat{\Theta} - \Theta\right)\right|_{\infty} \geq z/\left(2\sigma_{\max}^2 s\right)\right).$$

By Lemma 14 and the conditions of the present lemma, the r.h.s. is bounded above by $2\exp\left\{-nc_5z^2/\left(2\sigma_{\max}^2 s\right)^2 + 3\ln K\right\}$. Setting $c_6 = c_5/\left(4\sigma_{\max}^4\right)$, which is strictly positive, gives the result. ■

The following result will be used in due course.

Lemma 16 *Suppose that U, V_1, V_2 and $\hat{U}$, $\hat{V}_1$, $\hat{V}_2$ are random variables. Then, the event $\left\{\left|\frac{\hat{U}}{\hat{V}_1\hat{V}_2} - \frac{U}{V_1V_2}\right| \geq x\right\}$ is contained in the union of the following three events: $\left\{\left|\frac{\hat{U}\left(\hat{V}_1 - V_1\right)}{\hat{V}_1V_1V_2}\right| \geq x/4\right\}$, $\left\{\left|\frac{\hat{U}\left(\hat{V}_2 - V_2\right)}{\hat{V}_1\hat{V}_2V_2}\right| \geq x/4\right\}$ and $\left\{\left|\frac{\hat{U}-U}{V_1V_2}\right| \geq x/2\right\}$.*

Proof. Add and subtract $\frac{\hat{U}}{V_1V_2}$ to find that

$$\frac{\hat{U}}{\hat{V}_1\hat{V}_2} - \frac{U}{V_1V_2} = \left(\frac{\hat{U}}{\hat{V}_1\hat{V}_2} - \frac{\hat{U}}{V_1V_2}\right) + \left(\frac{\hat{U}}{V_1V_2} - \frac{U}{V_1V_2}\right).$$

The first term on the r.h.s. can be written as

$$\left(\frac{\hat{U}}{\hat{V}_1\hat{V}_2} - \frac{\hat{U}}{V_1V_2}\right) = \left(\frac{\hat{U}}{\hat{V}_1\hat{V}_2V_1V_2}\right)\left[\hat{V}_2\left(\hat{V}_1 - V_1\right) + V_1\left(\hat{V}_2 - V_2\right)\right].$$

We can then deduce the statement of the lemma by basic set inequalities. ■

Let $\hat{\Xi}_{i,j} = \hat{\Sigma}_{\varepsilon,i,j}/\sqrt{\hat{\Sigma}_{\varepsilon,i,i}\hat{\Sigma}_{\varepsilon,j,j}}$ and similarly for $\Xi_{i,j}$ using Σ_ε in place of $\hat{\Sigma}_\varepsilon$. These are estimated and population correlation coefficients between $\varepsilon_{t,i}$ and $\varepsilon_{t,j}$.

Lemma 17 *Suppose that the Regularity Conditions hold. There is a constant $c_7 > 0$, such that, for n large enough,*

$$\max_{i,j\leq K}\Pr\left(\left|\hat{\Xi}_{i,j} - \Xi_{i,j|k}\right| \geq z\right) \leq 16\exp\left\{-ns^{-2}c_7z^2 + 3\ln K\right\}$$

for any z satisfying $zn \to \infty$ and $z\left(\ln n\right)\left(\ln\ln n\right) \to 0$.

Proof. We apply Lemma 16 to deduce that we need to bound the following probabilities

$$\Pr\left(E_1\right) := \Pr\left(\left|\frac{\hat{\Sigma}_{\varepsilon,i,j}\left(\hat{\Sigma}_{\varepsilon,i,i} - \Sigma_{\varepsilon,i,i}\right)}{\sqrt{\hat{\Sigma}_{\varepsilon,i,i}\Sigma_{\varepsilon,i,i}\Sigma_{\varepsilon,j,j}}}\right| \geq z/4\right),$$

$$\Pr\left(E_2\right) := \Pr\left(\left|\frac{\hat{\Sigma}_{\varepsilon,i,j}\left(\hat{\Sigma}_{\varepsilon,j,j} - \Sigma_{\varepsilon,j,j}\right)}{\sqrt{\hat{\Sigma}_{\varepsilon,i,i}\hat{\Sigma}_{\varepsilon,j,j}\Sigma_{\varepsilon,j,j}}}\right| \geq z/4\right)$$

and

$$\Pr(E_3) := \Pr\left(\left|\frac{\hat{\Sigma}_{\varepsilon,i,j}\left(\hat{\Sigma}_{\varepsilon,i,j}-\Sigma_{\varepsilon,i,j}\right)}{\sqrt{\Sigma_{\varepsilon,i,i}\Sigma_{\varepsilon,j,j}}}\right| \geq z/2\right).$$

We further define the following events: $E_4 := \left\{\max_{i,j\leq K}\left|\hat{\Sigma}_{\varepsilon,i,j}\right| \leq 3/2\right\}$, and $E_5 := \left\{\min_{i\leq K}\hat{\Sigma}_{\varepsilon,i,i} \geq \sigma_{\min}/2\right\}$ where $\sigma_{\min} > 0$ is the minimum eigenvalue of Σ, by Lemma 3. Then, $\Pr(E_1) \leq \Pr(E_1 \cap E_4 \cap E_5) + \Pr(E_4^c) + \Pr(E_5^c)$ where, as usual, the superscript c is used to denote the complement of a set. Before bounding each term separately, we note that by the Cauchy interlacing theorem (Bhatia, 1996, Corollary III. 1.5), the smallest eigenvalue of Σ_ε is no smaller than $\sigma_{\min}$. Moreover, $\Sigma_{\varepsilon,i,i} \geq \sigma_{\min}$. To see this note that the l.h.s. is equal to $e_i'\Sigma_\varepsilon e_i$, where e_i is the vector with i^{th} entry equal to one and all other entries equal to zero. On the other hand the r.h.s. is smaller than $\min_{x:x'x=1} x'\Sigma_\varepsilon x$ by the definition of minimum eigenvalue and the Cauchy's interlacing theorem. Now,

$$\begin{aligned}\Pr(E_1 \cap E_4 \cap E_5) \leq & \Pr\left(\left|3\sigma_{\min}^{-3/2}\left(\hat{\Sigma}_{\varepsilon,i,i}-\Sigma_{\varepsilon,i,i}\right)\right| \geq z/4\right) \\ \leq & 2\exp\left\{-ns^{-2}12^{-2}\sigma_{\min}^3 c_6 z^2 + 3\ln K\right\} \qquad \text{(A.7)}\end{aligned}$$

using the bounds implied by the events E_4 and E_5, the aforementioned remarks on $\Sigma_{\varepsilon,i,i}$, and then Lemma 15. Noting that $\hat{\Sigma}_{\varepsilon,i,j} \leq \Sigma_{\varepsilon,i,j} + \left|\hat{\Sigma}_{\varepsilon,i,j}\text{-}\Sigma_{\varepsilon,i,j}\right|$ and that $|\Sigma_{\varepsilon,i,j}| \leq 1$ because ε_t is the innovation of the variable Z_t with entries having variance one, we deduce that $\Pr(E_4^c) \leq \Pr\left(\left|\hat{\Sigma}_{\varepsilon,i,j}\text{-}\Sigma_{\varepsilon,i,j}\right| \geq 1/2\right)$ and this probability is eventually bounded by (A.7) as long as $z \to 0$. By the same argument used to bound $\Pr(E_4^c)$, we deduce that $\Pr(E_5^c)$ is eventually less than (A.7). Hence, $\Pr(E_1)$ is bounded by three times the r.h.s. of (A.7) for n large enough. By similar arguments, we also note that $\Pr(E_2)$ and $\Pr(E_3)$ are bounded by three and two times, respectively, the r.h.s. of (A.7). Putting everything together, and setting $c_7 := 12^{-2}\sigma_{\min}^3 c_6$, the result follows. ■

For any set $\mathbf{k} \subset [K]$ we let $\hat{\Xi}_{i,j|\mathbf{k}}$ be the correlation of $\varepsilon_{t,i}$ with $\varepsilon_{t,j}$ conditioning on $\{\varepsilon_{t,l} : l \in \mathbf{k}\}$.

Lemma 18 *Under the Regularity Conditions, there is a constant $c_7 > 0$ (same as in*

Lemma 17), such that, for n large enough,

$$\max_{i,j\leq K,\mathbf{k}\in\mathcal{K}_{i,j}} \Pr\left(\left|\hat{\Xi}_{i,j|\mathbf{k}} - \Xi_{i,j|\mathbf{k}}\right| \geq z\right) \leq 16\exp\left\{-\left(n-m\right)s^{-2}c_7 z^2 + 3\ln K\right\}$$

for $\mathcal{K}_{i,j} \subseteq \{[K]\setminus\{i,j\}\}$ of cardinality m and z satisfying

$$z\left(n-m\right) \to \infty \text{ and } z\left(\ln\left(n-m\right)\right)\left(\ln\ln\left(n-m\right)\right) \to 0.$$

Proof. By Lemma 2 in Kalisch and Bühlmann (2007) if the distribution of the sample correlation coefficient is $f\left(x;n\right)$ where n is the sample size, the distribution of the partial correlation coefficient is the same with n replaced by $n-m$, i.e. $f\left(x;n-m\right)$. Hence, we can use Lemma 10 with n replaced by $n-m$ everywhere and the lemma is proved. ■

The next is a trivial variation of lemma 3 in Kalisch and Bühlmann (2007) adapted to our inequalities.

Lemma 19 *Suppose that the Regularity Conditions hold. Define $L := 1/\left(1-2^{-2}\left[1+\bar{\sigma}\right]^2\right)$ where $\bar{\sigma}$ is as in Lemma 3. For $g\left(x\right) = 2^{-1}\ln\left(\frac{1+x}{1-x}\right)$, $x\in(-1,1)$, there is a constant $c_7 > 0$ (same as the one in Lemma 18), such that, for n large enough,*

$$\max_{i,j\leq K,\mathbf{k}\in\mathcal{K}_{i,j}} \Pr\left(\left|g\left(\hat{\Xi}_{i,j|\mathbf{k}}\right) - g\left(\Xi_{i,j|\mathbf{k}}\right)\right| \geq z\right) \leq 32\exp\left\{-\left(n-m\right)s^{-2}c_8\left(z/L\right) + 3\ln K\right\}$$

for $\mathcal{K}_{i,j} \subseteq \{[K]\setminus\{i,j\}\}$ of cardinality m and for z satisfying $z\left(n-m\right)\to\infty$ and $z\left(\ln\left(n-m\right)\right)\left(\ln\ln\left(n-m\right)\right)\to 0$.

Proof. By the mean value theorem $g\left(x\right) - g\left(y\right) = \partial g\left(\tilde{y}\right)\left(x-y\right)$ for $\tilde{y}$ is in the convex hull of $\{x,y\}$, $x,y\in(-1,1)$; here, $\partial g\left(\tilde{y}\right) = 1/\left(1-\tilde{y}^2\right)$ is the derivative of g evaluated at $\tilde{y}$. Suppose $\left|x-y\right| \leq \left(1-\bar{\sigma}\right)/2$ and $y\in[-\bar{\sigma},\bar{\sigma}]$ for some $\bar{\sigma} < 1$. Note that $\tilde{y}^2 \leq \left(y+\left|x-y\right|\right)^2$, so that $\partial g\left(\tilde{y}\right) \leq L$ and substituting the aforementioned upper bound for y and $\left|x-y\right|$ in terms of $\bar{\sigma}$, and using the definition of L. Set $V := \hat{\Xi}_{i,j|\mathbf{k}} - \Xi_{i,j|\mathbf{k}}$ and $U := \partial g\left(\tilde{\Xi}_{i,j|\mathbf{k}}\right)$ where $\tilde{\Xi}_{i,j|\mathbf{k}}$ is in the convex hull of $\left\{\hat{\Xi}_{i,j|\mathbf{k}}, \Xi_{i,j|\mathbf{k}}\right\}$. The event $\{UV\geq z\}$ is contained in the union of the events $\{V\geq z/L\}$ and $\{U > L\}$. From Lemma 18 we have that $\Pr\left(V\geq z/L\right) \leq 16\exp\left\{-\left(n-m\right)s^{-2}c_7\left(z/L\right) + 3\ln K\right\}$ for z satisfying the conditions of that lemma. The lemma then follows if we show that $\{U\geq L\}\subseteq\{V\geq z/L\}$ for $z\to 0$, as in the

statement of the lemma. To this end, note that $\{U \geq L\}$ is contained in the union of the events $\{U > L, V \leq (1-\bar{\sigma})/2\}$ and $\{V > (1-\bar{\sigma})/2\}$. The latter event is eventually contained in $\{V \geq z/L\}$ when $z \to 0$. Finally, the event $\{U > L, V \leq (1-\bar{\sigma})/2\}$ has probability zero because, by the remarks at the beginning of the proof, we know that $U \leq L$ when $V \leq (1-\bar{\sigma})/2$ and $\left|\Xi_{i,j|\mathbf{k}}\right| \leq \bar{\sigma}$, which is the case by Lemma 3, uniformly in K, for any $\mathbf{k} \in \mathcal{K}_{i,j}$. Hence, the lemma is proved. ■

A.1.7 Technical Lemmas for Lasso

For $S \subseteq [2K]$ and some constant $L > 0$, recall that the square of the compatibility constant is $\phi^2_{\text{comp}}(L, S, \Sigma) := \min\left\{\frac{sb'\Sigma b}{|b|_1^2} : b \in \mathcal{R}(L, S)\right\}$ where $\mathcal{R}(L, S) := \{b : |b_{S^c}|_1 \leq L\,|b_S|_1 \neq 0\}$ (van de Geer and Bühlmann, 2009) . Here S^c is the complement of S in $[2K]$. Throughout this section, the notation is as in Algorithm 2 and Section 5.5.1 and $\sigma_{\min}$ is as in Lemma 3. We have the following.

Lemma 20 *Under the regularity Conditions, for any $S \subseteq [2K]$ of cardinality s, and $L > 0$, $\phi_{\text{comp}}\left(L, S, \hat{\Sigma}\right) \geq \sigma_{\min}^{1/2} - (L+1)\sqrt{s\left|\hat{\Sigma} - \Sigma\right|_\infty}$.*

Proof. Note that the square root of the minimum eigenvalue of a matrix is a lower bound for the compatibility constant. To see this, note that $sb'\Sigma b/|b_S|_1^2 \geq s\sigma_{\min}|b|_2^2/|b_S|_1^2 \geq \sigma_{\min}$ because $s|b|_2^2 \geq s|b_S|_2^2 \geq |b_S|_1^2$. Then, the lemma is special case of Corollary 10.1 in van de Geer and Bühlmann (2009). ■

We now derive a basic bound for the Lasso procedure computed across $2K$ response variables, one at the time, using the sufficient statistic $\hat{\Sigma}$.

Lemma 21 *Define*

$$\lambda_0 = 2\left(1 + \max_{i\in[2K]} \sum_{j\in[2K]:j\neq i} |\Theta_{i,j}/\Theta_{i,i}|\right)\left|\hat{\Sigma} - \Sigma\right|_\infty. \tag{A.8}$$

Under the Regularity Conditions, on the event $E_{\text{Lasso}} := \{\lambda \geq 2\lambda_0\}$, we have that $\max_{i\in[K]}\left|\hat{\beta}^{(i)} - \beta^{(i)}\right|_1 = O_P(s\lambda/\sigma_{\min})$.

Proof. We prove first the result for a fixed i. We shall then see that the bound is uniform in $i \in [K]$. To avoid notational complexities, we use a notation that is only local to this proof. Set $\Gamma = \Sigma_{-i,-i}$, $\gamma = \Sigma_{-i,i}$, $b = \beta^{(i)}_{-i}$ and $\hat{b} = \hat{\beta}^{(i)}_{-i}$. Note

that $b = \Gamma^{-1}\gamma$ by definition. As in the text we use the hat for estimators of various quantities. Write $\delta = \hat{b} - b$. Given that the Lasso estimator minimises the Lasso objective function we have that

$$-2\hat{\gamma}'\hat{b} + \hat{b}'\hat{\Gamma}\hat{b} + \lambda\left|\hat{b}\right|_1 \leq -2\hat{\gamma}'b + b'\hat{\Gamma}b + \lambda\left|b\right|_1.$$

This can be rearranged to give the following inequality

$$\delta'\hat{\Gamma}\delta \leq 2\left(\hat{\gamma}' - b'\hat{\Gamma}\right)\delta + \lambda\left(\left|b\right|_1 - \left|\hat{b}\right|_1\right)$$

(Loh and Wainwright, 2012, eq. 5.1). Adding and subtracting $b'\Gamma$, we write $\left(\hat{\gamma}' - b'\hat{\Gamma}\right) = (\hat{\gamma}' - b'\Gamma) + b'\left(\Gamma - \hat{\Gamma}\right)$. Given that $b'\Gamma = \gamma'$, by definition of γ and $\hat{\gamma}$, we have that $\left|\hat{\gamma} - \Gamma b\right|_\infty \leq \left|\hat{\Sigma} - \Sigma\right|_\infty$. By definition of Γ and $\hat{\Gamma}$ and a basic inequality, $\left|\left(\Gamma - \hat{\Gamma}\right)b\right|_\infty \leq \left|b\right|_1\left|\hat{\Sigma} - \Sigma\right|_\infty$. However, $\left|b\right|_1 = \sum_{j\in[2K]:j\neq i}\left|\Theta_{i,j}/\Theta_{i,i}\right|$ because the regression coefficients can be obtained from the precision matrix: $\beta_j^{(i)} = -\Theta_{i,j}/\Theta_{i,i}$. Hence, by definition of λ_0 as in the statement of the lemma and the last display, we deduce that $\delta'\hat{\Gamma}\delta \leq \lambda_0\left|\delta\right|_1 + \lambda\left(\left|b\right|_1 - \left|\hat{b}\right|_1\right)$. This is in the form of the basic inequality in van de Geer and Bühlmann (2009, last display on p.1387). On the set $\{\lambda \geq 2\lambda_0\}$, the r.h.s. of the previous inequality is bounded above by $2^{-1}\lambda\left|\delta\right|_1 + \lambda\left(\left|b\right|_1 - \left|\hat{b}\right|_1\right)$. Then, by arguments in van de Geer and Bühlmann (2009, second and third display on p.1388, replacing λ_0 with $2^{-1}\lambda$ in their definition of L, so that here $L = 3$), we deduce that

$$\left|\delta\right|_1 \leq 4\sqrt{s\delta'\hat{\Gamma}\delta/\hat{\phi}^2_{\text{comp}}}$$

where $\hat{\phi}_{\text{comp}} := \phi_{\text{comp}}\left(L, S, \hat{\Sigma}\right)$ is the compatibility constant, which we shall show to be strictly positive. Lemma 11.2 in van de Geer and Bühlmann (2009) says that $\sqrt{\delta'\hat{\Gamma}\delta} = O\left(\frac{\lambda\sqrt{s}}{\hat{\phi}_{\text{comp}}}\right)$ once we replace λ_0 with $\lambda/2$ in their lemma. By Lemmas 20 and 9, $\hat{\phi}_{\text{comp}} = \sigma^{1/2}_{\min} - O_P\left(\sqrt{s\frac{\ln K}{n}}\right)$ choosing $L = 3$ in Lemma 20. We also have that $\sqrt{s\frac{\ln K}{n}} = o\left(\sigma^{1/2}_{\min}\right)$. By these remarks and the above display, we deduce $\left|\delta\right|_1 = O_P\left(\frac{s\lambda}{\sigma_{\min}}\right)$. The bound is uniform in $i \in [K]$ because Lemma 3. Hence, the result follows. ■

Lemma 22 *Suppose that the Regularity Conditions hold. Then, for λ_0 is as in (A.8),*

$\lambda_0 = O_P\left((\omega/\nu^2)\sqrt{\frac{\ln K}{n}}\right)$ *where* ν *is as in Lemma 3.*

Proof. Under the Regularity Conditions, an upper bound for (A.8) is given by $2(1+\omega/\nu^2)\left|\hat{\Sigma}-\Sigma\right|_\infty$. This is $O_P\left((\omega/\nu^2)\sqrt{\frac{\ln K}{n}}\right)$ using Lemma 9. Hence, the result follows. ■

A.1.8 Proof of Theorem 1

This follows from Lemma 9.

A.1.9 Proof of Theorem 2

An upper bound for (A.8) is given by $2(1+\omega/\nu^2)\left|\hat{\Sigma}-\Sigma\right|_\infty$. Then, in Lemma 21, the set $\Pr(E_{\text{Lasso}}) \to 1$ as $K \to \infty$, for $\lambda = 4(1+\omega/\nu^2) \times \frac{3c_0}{\pi}\sqrt{\frac{\ln K}{n}}$, by Lemma 9. Therefore, by Lemma 21, $\max_{i\in[K]}\left|\hat{\beta}^{(i)}-\beta^{(i)}\right|_1 = O_P\left(\omega s\sqrt{\frac{\ln K}{n}}\right)$ and we can choose $c = 12(1+\nu^{-2})c_0/\pi$ in the statement of the theorem. Hence, the result follows.

A.1.10 Proof of Theorem 3

Note that $\theta_{\min}$ is a lower bound on $\min_{i,j}\left\{\left|\beta_j^{(i)}\right| : \left|\beta_j^{(i)}\right| > 0\right\}$. This is because $\left|\beta_j^{(i)}\right| = |\Theta_{i,j}/\Theta_{i,i}|$. Note that $-\Theta_{i,i}$ is the variance of $Z_{t,i}$ conditioning on all other covariates. Hence, $|\Theta_{i,i}| \leq 1$ because $Var(Z_{ti}) = 1$. Then, the event in the probability of the theorem is contained in the event $\max_{i\in[K]}\left|\hat{\beta}^{(i)}-\beta^{(i)}\right|_1 > \tau$, because $\tau = o(\theta_{\min})$. The latter event has probability going to zero according to Theorem 2.

A.1.11 Proof of Theorem 4

By Theorem 6 in Cai et al. (2011), $\left|\hat{\Omega}-\Theta\right|_\infty \leq 4|\Theta|_{1,\infty}\lambda_n$, on the event $E_{\text{Clime}} := \left\{\lambda_n \geq |\Theta|_{1,\infty}\left|\hat{\Sigma}-\Sigma\right|_\infty\right\}$. Choosing $\lambda_n = \omega\left(\frac{3c_0}{\pi}\sqrt{\frac{\ln K}{n}}\right)$, by Lemma 9, $\Pr(E_{\text{Clime}}) \to 1$ as $K \to \infty$.

A.1.12 Proof of Theorem 5

Due to the fact that $|\Theta_{i,j}| \in \{0\} \cup [\theta_{\min}, \infty)$ and $\left|\hat{\Omega}_{i,j}\right| \in \{0\} \cup [\tau, \infty)$ uniformly in $i, j \in [2K]$, the event in the probability of the theorem is eventually contained in

$\left\{\left|\hat{\Omega}-\Theta\right|_\infty \geq \tau\right\}$. This goes to zero by Theorem 4 because τ is of larger order of magnitude than $\left|\hat{\Omega}-\Theta\right|_\infty$.

A.1.13 Proof of Theorem 6

Under the event E in (A.4), we are within the framework of the results in Le and Zhong (2021). When such event is true, the result follows from Theorem 3 in Le and Zhong (2021). The proof of their result requires a bound in probability for $\left|\hat{\Sigma}-\Sigma\right|_\infty$; see the third display on their page 12. In their proof this is denoted by the symbol $|W_{X,nj}|_\infty$. We control this quantity using Lemma 9. To finish the proof note that $\Pr(E) \to 1$ using either Theorem 3 or Theorem 5.

A.1.14 Proof of Theorem 7

From Lemma 1, recall that $\Sigma_\varepsilon = \Theta_{11}^{-1}$ and $A = -\Theta_{11}^{-1}\Theta_{12}$. By Lemmas 11 and 12, the Regularity Conditions and Theorem 6, we deduce that $\left|\hat{\Theta}_{11}^{-1}-\Theta_{11}^{-1}\right|_{\text{op}} = O_P\left(s\sqrt{\frac{\ln K}{n}}\right)$ on the event E in (A.4); note that $|\Theta_{11}|_{0,\infty} \leq s$. The event E has probability going to one by either Theorem 3 or Theorem 5. This proves the first bound in the theorem. To prove the convergence of the autoregressive matrix estimator, we note that $A-\hat{A} = \hat{\Theta}_{11}^{-1}\hat{\Theta}_{12} - \Theta_{11}^{-1}\Theta_{12}$. The r.h.s. can be rewritten as $\hat{\Theta}_{11}^{-1}\left(\hat{\Theta}_{12}-\Theta_{12}\right) + \left(\hat{\Theta}_{11}^{-1}-\Theta_{11}^{-1}\right)\Theta_{12}$. The first term in the sum is equal to $\Theta_{11}^{-1}\left(\hat{\Theta}_{12}-\Theta_{12}\right) + \left(\hat{\Theta}_{11}^{-1}-\Theta_{11}^{-1}\right)\left(\hat{\Theta}_{12}-\Theta_{12}\right)$. Then, by standard inequalities and the previous bounds, it is not difficult to deduce that its operator norm is $O_P\left(s\sqrt{\frac{\ln K}{n}}\right)$. The same follows for the operator norm of $\left(\hat{\Theta}_{11}^{-1}-\Theta_{11}^{-1}\right)\Theta_{12}$. This concluded the proof of the theorem.

A.1.15 Proof of Theorem 8

The assumptions in Kalisch and Bühlmann (2007) are satisfied by our Regularity Conditions together with the faithfulness condition stated in the theorem. In particular, from Kalisch and Bühlmann (2007, proof of Lemma 4), it is sufficient to bound the probability of a Type I and Type II error, as given by the following

$$\Pr\left(\left|g\left(\hat{\Xi}_{i,j|\mathbf{k}}\right)-g\left(\Xi_{i,j|\mathbf{k}}\right)\right| \geq z\right) \leq 32\exp\left\{-(n-m)\,s^{-2}c_7\,(z/L)^2 + 3\ln K\right\}$$

where m is the cardinality of $\mathbf{k}$, g is as defined in Lemma 19, and setting $z = c_n$ where c_n as in Kalisch and Bühlmann (2007): $c_n \asymp n^{-\eta_c}$. Choosing m equal to the maximal number of adjacent nodes, there are $O(K^m)$ hypotheses to test. By Lemma 5 in Kalisch and Bühlmann (2007), we can assume $m \leq s$ with probability going to one. By this remark and the union bound we need the following to go to zero: $K^s 32 \exp\left\{-(n-s)\, s^{-2} c_7 \left(c_n/L\right)^2 + 3 \ln K\right\}$. By the Regularity Conditions, $s = O(n^{\eta_s}) = o\left(n^{1/2}\right)$ and $K^s = O(n^{s\eta_K})$ for some finite η_K. Hence we must have $n^{\eta_s} \ln n = o\left(n^{1-2(\eta_s+\eta_c)}\right)$. This is the case if $2\eta_c + 3\eta_s < 1$, as stated in the theorem. The theorem is then proved following the steps in the proof of Lemma 4 in Kalisch and Bühlmann (2007).

A.1.16 Proof of Theorem 9

Define the set $E_G := \left\{\hat{G} = G\right\}$, where $\hat{G}$ is the PCDAG estimated using Algorithm 5 and G is the true PCDAG. Hence, on E_G we have that that $\hat{\mathcal{V}}(i) = \mathcal{V}(i)$. By Theorem 8, the event E_G has probability going to one. Hence, in what follows, we shall replace $\hat{\mathcal{V}}(i)$ with $\mathcal{V}(i)$. By the assumption of the present theorem, G has all edges that are directed. Let

$$\hat{\Psi} := \begin{bmatrix} \hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(1),\hat{\mathcal{V}}(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(2),\hat{\mathcal{V}}(2)} & \ddots & \vdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(K),\hat{\mathcal{V}}(K)} \end{bmatrix}$$

and

$$\hat{\Phi} := \begin{bmatrix} \hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(1),1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(2),2} & \ddots & \vdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(K),K} \end{bmatrix};$$

where the symbol $\mathbf{0}$ denotes a generic conformable matrix of zeros. Then, the nonzero consecutive entries in the i^{th} column of $\hat{\Psi}^{-1}\hat{\Phi}$ is equal to $\hat{d}_i$ as defined in Algorithm 6. Here, we shall define the population version of the above by Ψ and Φ. We define a matrix R such that $\Delta = \left(R\hat{\Psi}^{-1}\hat{\Phi}\right)'$. The matrix R reshapes $\hat{\Psi}^{-1}\hat{\Phi}$ so that we can

find Δ. We write such matrix R as

$$R := \begin{bmatrix} R_1^{(1)} & R_1^{(2)} & \cdots & R_1^{(K)} \\ R_2^{(1)} & R_2^{(2)} & \cdots & R_2^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ R_K^{(1)} & R_K^{(2)} & \cdots & R_K^{(K)} \end{bmatrix},$$

where $R_k^{(i)}$ is a $1 \times \mathcal{V}(i)$ vector defined as follows. If $k \notin \mathcal{V}(i)$, then, $R_k^{(i)}$ is a row vector of zeros; for example $R_k^{(k)} = 0$, $k \in [K]$. If $k \in \mathcal{V}(i)$, $R_k^{(i)}$ will have a one in the position such that $R_k^{(i)} \hat{d}_i' \varepsilon_{t,\mathcal{V}(i)} = \hat{d}_{i,j} \varepsilon_{t,k}$, where j is the position of the element in $\mathcal{V}(i)$ that is equal to k; $\hat{d}_{i,j}$ is the estimated regression coefficient of $\varepsilon_{t,k}$ in the regression of $\varepsilon_{t,i}$ on $\varepsilon_{t,\mathcal{V}(i)}$. This also means that the number of ones in the k^{th} row of R is equal to the number of direct descendants of the variable $\varepsilon_{t,k}$. We denote such number by κ_k. Now, note that $\left|R\hat{\Psi}^{-1}\hat{\Phi} - R\Psi^{-1}\Phi\right|_{\text{op}} \leq |R|_{\text{op}} \left|\hat{\Psi}^{-1}\hat{\Phi} - \Psi^{-1}\Phi\right|_{\text{op}}$. Then, $|R|_{\text{op}}^2$ is the maximum eigenvalue of RR' and the latter matrix is diagonal with (k,k) entry equal to κ_k. It is easy to see that RR' is diagonal because the positions for two different parents cannot overlap, i.e. $R_k^{(i)} \left(R_l^{(i)}\right)' = 0$ when $k \neq l$. Then, $|R|_{\text{op}} = \kappa^{1/2}$, where $\kappa := \max_k \kappa_k$, as defined in the theorem. Hence, it remains to bound $\left|\hat{\Psi}^{-1}\hat{\Phi} - \Psi^{-1}\Phi\right|_{\text{op}}$; note that the singular values of a matrix are invariant of transposition. Adding and subtracting $\Psi^{-1}\hat{\Phi}$, using the triangle inequality, and a basic norm inequality,

$$\left|\hat{\Psi}^{-1}\hat{\Phi} - \Psi^{-1}\Phi\right|_{\text{op}} \leq \left|\hat{\Psi}^{-1} - \Psi^{-1}\right|_{\text{op}} \left|\hat{\Phi}\right|_{\text{op}} + \left|\Psi^{-1}\right|_{\text{op}} \left|\hat{\Phi} - \Phi\right|_{\text{op}}. \tag{A.9}$$

By Lemma 11, $\left|\hat{\Psi}^{-1} - \Psi^{-1}\right|_{\text{op}} \leq |\Psi^{-1}|_{\text{op}}^2 \left|\hat{\Psi} - \Psi\right|_{\text{op}}$. The maximum singular value of a block diagonal matrix is the maximum of the singular values of each of the blocks. By Cauchy's interlacing theorem, $\left|\hat{\Psi} - \Psi\right|_{\text{op}} \leq \left|\hat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right|_{\text{op}}$ and the latter is $O_P\left(s\sqrt{\frac{\ln K}{n}}\right)$ by Theorem 7. Using again Cauchy's interlacing theorem, we deduce that the largest singular value of Ψ^{-1} is bounded above by the largest singular value of Θ, which is finite. Moreover, $\left|\hat{\Phi}\right|_{\text{op}} \leq |\Phi|_{\text{op}} + \left|\hat{\Phi} - \Phi\right|_{\text{op}}$. The maximum singular value of Φ is just the maximum of $\Sigma'_{\varepsilon,\hat{\mathcal{V}}(i),i}\Sigma_{\varepsilon,\hat{\mathcal{V}}(i),i}$ w.r.t. $i \in [K]$. It is increasing in the cardinality of $\hat{\mathcal{V}}(i)$. Hence, $\Sigma'_{\varepsilon,\hat{\mathcal{V}}(i),i}\Sigma_{\varepsilon,\hat{\mathcal{V}}(i),i} \leq \Sigma'_{\varepsilon,\cdot,i}\Sigma_{\varepsilon,\cdot,i}$, recalling the notation at the start of Section 4. The latter is bounded above by $\max_{x'x\leq 1} x'\Sigma'_\varepsilon\Sigma_\varepsilon x = |\Sigma_\varepsilon|_{\text{op}}^2$, which is bounded, by

the Regularity Conditions. By the same argument as before, the maximum singular value of $\hat{\Phi} - \Phi$ is the square root of the largest, w.r.t. $i \in [K]$, of the maximum eigenvalue of

$$\left(\hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(i),i} - \Sigma_{\varepsilon,\mathcal{V}(i),i}\right)' \left(\hat{\Sigma}_{\varepsilon,\hat{\mathcal{V}}(i),i} - \Sigma_{\varepsilon,\mathcal{V}(i),i}\right)$$

where on E_G, $\hat{\mathcal{V}}(i) = \mathcal{V}(i)$. This quantity is increasing in the cardinality of $\mathcal{V}(i)$ so that the square root of the above display is bounded above by $\left|\hat{\Sigma}_{\varepsilon} - \Sigma_{\varepsilon}\right|_{\text{op}}$, which is $O_P\left(s\sqrt{\frac{\ln K}{n}}\right)$ by Theorem 7. Using the derived upper bounds, it is easy to deduce that (A.9) is $O_P\left(s\sqrt{\frac{\kappa \ln K}{n}}\right)$.

From Lemma 2, deduce that $\Pi\varepsilon_t = D\Pi\varepsilon_t + \xi_t$. This can be rewritten as $\varepsilon_t = \Pi^{-1}D\Pi\varepsilon_t + \Pi^{-1}\xi_t$. Hence, $\varepsilon_t = \Delta\varepsilon_t + \Pi^{-1}\xi_t$, where $\Delta = \Pi^{-1}D\Pi$. Now, note that on the event E_G, as defined at the start of the proof, any permutation matrix $\hat{\Pi}$ that makes $\hat{\Pi}\hat{\Delta}\hat{\Pi}^{-1}$ lower triangular, with diagonal entries equal to zero, also satisfies (4) when we replace Π with it. According to Algorithm 6 we choose the one that requires the least number of row permutations of the identity, which is unique. Then, on E_G, $\hat{\Pi} = \Pi$ because also Π is unique. Therefore, on E_G, $\hat{D} := \hat{\Pi}\Delta\hat{\Pi}^{-1}$ converges to $D := \Pi\Delta\Pi^{-1}$. This shows the first statement of the theorem. The convergence rate of $\hat{H} - H$ to zero can be deduce from the first statement of the theorem together with Lemma 11, and Cauchy's interlacing theorem and the definition $\Sigma_{\varepsilon} = H\left(\mathbb{E}\xi_t\xi_t'\right)H'$ in order to bound the singular values of $H^{-1} := (I - D)$.

A.2 Choice of Tuning Parameters

Algorithms 2 and 3 require to choose the penalty parameter λ and the threshold τ. As shown in Theorems 3 and 5 we need $\tau > \lambda$. The exact values can be chosen by crossvalidation (CV). CV may not be suitable for time series problems. However, it has been shown to work for prediction problems in the case of autoregressive process of finite order (Burmann and Nolan, 1992). To this end, we divide the sample data into n_{CV} nonoverlapping blocks of equal size each. Each block is a test sample. Given the i^{th} test sample, we use the remaining data as i^{th} estimation sample. Compute $\hat{\Theta}$ on the i^{th} estimation sample and denote this by $\hat{\Theta}_{\text{est}}(\lambda, \tau, i)$ to make the dependence on the parameters and block explicit. Compute the scaling matrix $\hat{\Sigma}$ on the i^{th} test sample using Algorithm 1 and denote it by $\Sigma_{\text{test}}(i)$ to make the dependence explicit.

We minimize the negative loglikelihood:

$$\frac{1}{n_{\text{CV}}}\sum_{i=1}^{n_{\text{CV}}}\left[\text{Trace}\left(\hat{\Sigma}_{\text{test}}\left(i\right)\hat{\Theta}_{\text{est}}\left(\lambda,\tau,i\right)\right)-\ln\det\left(\hat{\Theta}_{\text{est}}\left(\lambda,\tau,i\right)\right)\right]$$

w.r.t. $(\lambda,\tau) \in \mathcal{T}$ where $\mathcal{T} \subset (0,\infty)^2$. Here, for any matrix A, $\text{diag}\,(A)$ a diagonal matrix with same diagonal entries as A.

In our simulations and empirical analysis, the parameter τ is fixed to 2λ, and we select λ employing CV with $n_{CV} = 5$. Starting with a penalization equal to $\lambda = 0.10$, we first search (by dividing iteratively by two) a value for the minimum λ such that all off-diagonal elements of $\hat{\Theta}_{11}$ are zero (precisely smaller than 1e-6). We denote this value as λ_0. Then we search for the optimal λ in $\{\lambda_0/2, \lambda_0/(2^2), \ldots, \lambda_0/(2^5)\}$.

Computing both optimal parameters and a causal graph from the PC algorithm can be time consuming over many simulations. Hence, in our simulations, we employ an additional simplification. Rather than carrying out CV for each simulation, we use two separate simulation samples to compute two values of λ according to the aforementioned procedure. We then use the average of these two values as tuning parameter λ in all simulations with the same design.

A.3 Finite Sample Analysis via Simulations

We assess the finite sample performance of the different estimators and evaluate their asymptotic properties for various degrees of time series persistence and cross-sectional dimension. We compare our results to naive methods that either do not account for sparsity in Θ or ignore the time series structure of the data.

A.3.1 The True Model

To generate the time series of equation (1) the K variables are divided into $\widetilde{K}$ independent clusters. Each cluster is composed by N variables and shares the same causal structure as well as the autoregressive matrix. We denote with $\widetilde{A}$ and $\widetilde{H}$ the related coefficients of equation (1) for each cluster. The matrix $\widetilde{H}$ is the matrix which relates ε_t with the associated structural shocks ξ_t of a selected cluster. For the sake of simplicity, for each cluster, the variables' order coincides with the topological order so that the matrix Π in Lemma 2 can be set equal to the identity.

We consider $N = 3$ and $N = 4$. When $N = 3$ the three basic causal structures are selected for each cluster, i.e., the causal chain, common cause and v-structure. Given three variables X, Y and Z, if $X \to Y \to Z$, the causal structure is called causal chain while if $X \leftarrow Y \to Z$ it is termed common cause. The causal relation is named v-structure or immorality if $X \to Y \leftarrow Z$. We also consider two additional structures when $N = 4$: diamond 1 and diamond 2. These are defined as $X \to Y \leftarrow Z, X \to U \leftarrow Z$, and $X \to Y \leftarrow Z, Y \to U$, respectively.

The PC algorithm cannot distinguish between causal chain and common cause, since these structures are in the same Markov equivalence class. Then, the PC algorithm will provide the same graph with undirected edges: $X - Y - Z$. Conversely, the v-structure, diamond 1 and diamond 2 can be identified by the PC algorithm. In this case, the PC algorithm will return the causal graph with edges correctly oriented.

To monitor the persistence of the time series, for each cluster, the autoregressive matrix $\widetilde{A}$ is equal to a lower triangular matrix with all elements (including the diagonal) equal to a constant a, which describes the persistence of the series. The matrix $\widetilde{H}$ is a function of the selected causal structure. For the v-structure

$$\widetilde{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

which is related to the causal structure $\varepsilon_{t,1} \to \varepsilon_{t,3} \leftarrow \varepsilon_{t,2}$. Each variable causes itself, but may also affect other variables. Finally, for simplicity, we suppose that the data have Gaussian marginals. In this case, simulation of (1) reduces to simulation of a VAR(1) together with some linear transformations to ensure that all the covariates have variance equal to one. The details are given in Algorithm 7.

A.3.2 Simulation Results

To study the effect of time series persistent, three values of such parameter a are considered: 0.25, 0.5 and 0.75. These values of a produce a wide range of time series dependence. For example, Figure A.1 shows the autocorrelation function of a cluster for a v-structure. To analyze the relevance of sparsity in our approaches, we select $\widetilde{K} = 3, 30, 50$ clusters. We investigate the finite sample properties of our estimator by considering a sample size $n = 1000, 5000$.

Algorithm 7 Simulation of the Gaussian Copula VAR in (1) when the Marginals are Gaussian.

Set $N \times N$ matrices $\widetilde{A}$ and $\widetilde{H}$ s.t. $\widetilde{H}$ is full rank.
For $k = 1, 2, ...\widetilde{K}$:
Simulate i.i.d. $N \times 1$ dimensional Gaussian vectors $\left(e_t^{(k)}\right)_{t\in[n]}$ with mean zero and identity covariance matrix.
Compute $X_t^{(k)} = \widetilde{A}X_{t-1}^{(k)} + \widetilde{H}e_t^{(k)}$, $t \in [n]$.
End of For.
Define the K-dimensional VAR(1) $X_t = A_{\text{block}}X_{t-1} + H_{\text{block}}e_t$, where $X_t = \left(\left(X_t^{(1)}\right)', \left(X_t^{(2)}\right)', \ldots, \left(X_t^{(\widetilde{K})}\right)'\right)'$ and similarly for e_t, $t \in [n]$; a fortiori, A_{block} and $H_{\text{block}}$are block diagonal matrices, where each block equals $\widetilde{A}$ and $\widetilde{H}$, respectively.
Define $S = \left[\text{diag}\left(Var\left(X_t\right)\right)\right]^{-1}$ where $\text{diag}\left(\cdot\right)$ is the diagonal matrix with diagonal equals to its argument.
Set $A = SA_{\text{block}}S^{-1}$, $\Sigma_\varepsilon = SH_{\text{block}}H'_{\text{block}}S'$.
Define the latent $K \times 1$ vector $Z_t = SX_t$, $t \in [n]$.

We use Algorithms 2 and 3 find the moral graph. Recall that the moral graph is defined from the nonzero entries in $\hat{\Theta}$ as in Algorithm 4. We then follow Algorithms 5 and 6 to estimate any remaining parameters. The tuning parameters for Algorithms 2 and 3 are chosen by CV as described in Section A.2. This means only choosing λ. We denote the estimated parameter by λ_{CV}. We use 250 simulations to compute the performance of our methodology.

We also test the performance of the PC algorithm when we impose the restrictions provided by Lasso and CLIME. The elements of $\hat{\Theta}_{11}$ which are equal to zero represent those edges which we exclude from the skeleton. These restrictions can be embedded in the PC using the appropriate *fixedGaps* command, which guarantees that will be no edge between nodes j and i if the element of $\hat{\Theta}_{11}$ in position (i, j) is equal to zero. We obtain improved compute time performance of the PC algorithm in this case. This is particularly relevant in the high dimensional case. As discussed in the main text, when we impose the restrictions from the zeros of $\hat{\Theta}_{11}$, it is advisable to use a tuning parameter λ smaller than the one suggested by CV. This is because the PC algorithm can only delete edges, but not add them back. Hence, it is more important to avoid false negatives than false positives when imposing restrictions. To corroborate this claim, we also report results for $\lambda_{CV}/2$ and $\lambda_{CV}/4$.

We compare our results with two benchmarks. One does not account for sparsity

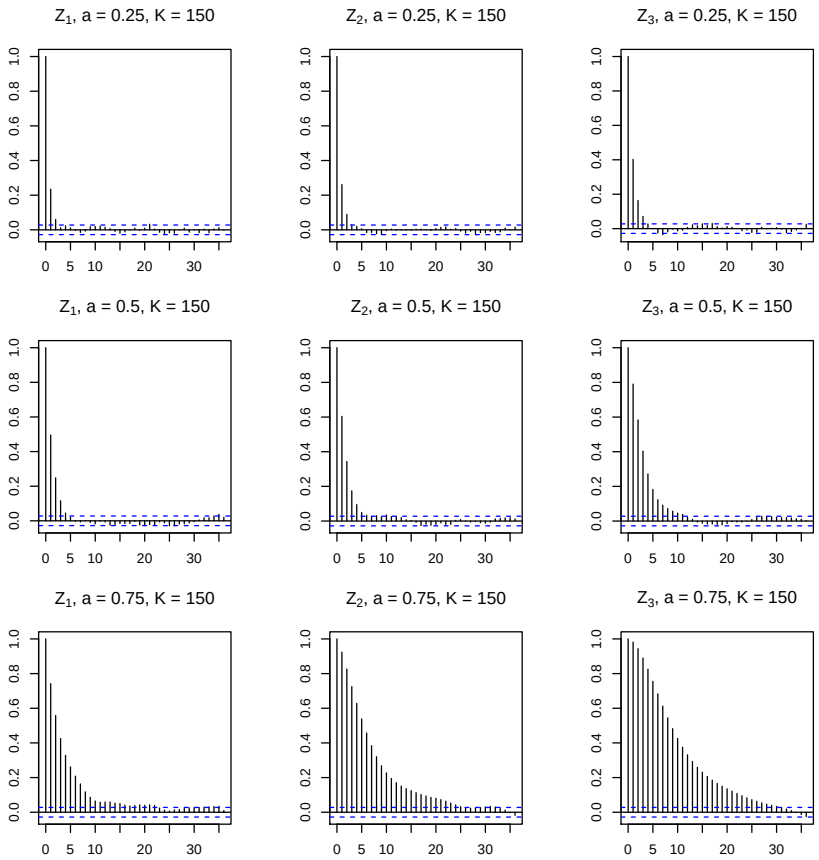


Figure A.1: Autocorrelation functions of the variable Z_t of a cluster where the contemporaneous causal relations are generated by a v-structure.

and is essentially equivalent to choosing $\lambda = 0$ in the estimation. The second does not account for time series dependence and is equivalent to assuming (1) with $A = 0$. In the simulations, we refer to the two benchmarks as $\lambda = 0$ and $A = 0$, respectively. The case $\lambda = 0$ should produce sensible results in the low-dimensional case. The case $A = 0$ might be appropriate when the time series persistence is low. In this case, the procedure is usually biased, but incurs a lower estimation error.

All approaches are compared on their performance to estimate the contemporaneous causal structure. To achieve this, we report the average structural Hamming distance (SHD) of the estimated causal graph to the true (Acid and de Campos, 2003, Tsamardinos et al., 2006). The SHD between two partially directed acyclic graphs counts how many edge types do not coincide. For instance, estimating a non-edge instead of a directed edge contributes an error of one to the overall distance. We remark

that the PC algorithm estimates the Markov equivalence class of a given graph, i.e., the related CPDAG, and some causal structure, as common cause and causal chain, shares the same class, i.e., the same CPDAG, (e.g., for the v-structure the Markov class coincides with the related DAG). Therefore, as the true causal structure in SHD analysis we consider the (block) equivalence class attained by the PC, with a sufficiently high significance level, 1-1e-13, to obtain a deterministic estimate performed on the theoretical correlation matrix of each cluster.

Tables 2 and 3 display the average SHD and standard errors computed over 250 simulations for all approaches. For the sake of conciseness we only report results for the v-structure for the persistency parameter $a \in \{0.25, 0.75\}$ and the number of clusters $\widetilde{K} \in \{3, 50\}$[3]. Our approach produces estimators with superior finite sample performance, relatively to the benchmarks, regardless of the considered causal structures. While not reported here, we note that for both the causal chain and common cause, the performance of the PC Algorithm deteriorates when we impose the a priori restrictions from the zeros of $\hat{\Theta}_{1,1}$ even if we undersmooth.

The discrepancy among the contemporaneous causal structure is also investigated by computing the number of non-zero elements of Θ_{11}. Indeed, we recall that non-zero elements of Θ_{11} correspond to possible edges between variables of the corresponding row and column. We also compute the number of false positive and negative between the estimated and true Θ_{11} of non-zero elements[4]. Tables 4 and 5 summarize the results for the high and low dimensional case, respectively. We only report the results for the v-structure, as we can draw similar conclusions for the other causal structures.

Finally, in Tables 6 and 7, we assess the finite sample performance of the estimators of A and Σ_ε and analyse their asymptotic properties stated in Theorem 7. We compute the average distance from the true matrices, where the distance is measured in terms of the operator's norm: the largest singular value. These statistics are compared only to the case $\lambda = 0$.

[3]The complete results are available upon request.

[4]We say that an element of Θ_{11} is a false positive, if it is estimated as non-zero element while it is zero. Vice versa, it is a false negative, if it is estimated as zero element while it is different from zero.

Table 2: Structural Hamming Distance for a Causal V-Structure. Expected value approximated using 250 Monte Carlo simulations (standard errors in parenthesis) for the SHD between the Lasso and CLIME estimators in Algorithms 2 and 3, and the true one. The contemporaneous causal structure is a v-structure with $K = 150$ variables with $\widetilde{K} = 50$ clusters. Results are reported for different values of λ , where λ_{CV} is the value obtained using cross-validation and denoted by λ_{CV}. The columns labelled *NR* reports the SHD obtained when no restrictions provided by Lasso and CLIME procedures, respectively, are used in the initialization step of the PC. The columns $\lambda = 0$ and $A = 0$ refer to the benchmarks that do not account for sparsity and time series dependence, respectively.

				Lasso					
		λ_{CV}		$\lambda_{CV}/2$		$\lambda_{CV}/4$		$\lambda = 0$	$A = 0$
n	a		NR		NR		NR		
1000	0.25	9.208	9.212	58.032	58.160	66.080	71.616	40.424	45.628
		(0.28)	(0.28)	(0.48)	(0.48)	(0.49)	(0.52)	(0.39)	(0.57)
	0.75	1.960	95.888	4.464	4.488	29.596	29.556	131.060	225.212
		(0.14)	(0.2)	(0.2)	(0.2)	(0.35)	(0.34)	(0.93)	(0.48)
5000	0.25	3.124	3.124	44.700	44.776	31.092	32.504	22.144	144.944
		(0.16)	(0.16)	(0.43)	(0.43)	(0.37)	(0.38)	(0.29)	(0.29)
	0.75	0	99.904	2.496	2.496	2.780	2.780	51.696	230.704
		(0)	(0.03)	(0.17)	(0.17)	(0.17)	(0.17)	(0.47)	(0.62)
				CLIME					
		λ_{CV}		$\lambda_{CV}/2$		$\lambda_{CV}/4$			
n	a		NR		NR		NR		
1000	0.25	27.700	27.740	53.496	53.604	78.984	83.340		
		(0.51)	(0.51)	(0.47)	(0.47)	(0.51)	(0.51)		
	0.75	100.012	100.012	56.776	105.104	12.880	96.220		
		(0.01)	(0.01)	(0.53)	(0.19)	(0.33)	(0.32)		
5000	0.25	2.488	2.488	41.744	41.892	39.896	41.104		
		(0.15)	(0.15)	(0.45)	(0.45)	(0.37)	(0.38)		
	0.75	119.440	138.064	3.192	4.392	6.348	6.348		
		(0.5)	(0.18)	(0.19)	(0.21)	(0.23)	(0.23)		

Table 3: Structural Hamming Distance for a Causal V-Structure. Expected value approximated using 250 Monte Carlo simulations (standard errors in parenthesis) for the SHD between the Lasso and CLIME estimators in Algorithms 2 and 3, and the true one. The contemporaneous causal structure is a v-structure with $K = 9$ variables with $\widetilde{K} = 3$ clusters. Results are reported for different values of λ , where λ_{CV} is the value obtained using cross-validation and denoted by λ_{CV}. The columns labelled *NR* reports the SHD obtained when no restrictions provided by Lasso and CLIME procedures, respectively, are used in the initialization step of the PC. The columns $\lambda = 0$ and $A = 0$ refer to the benchmarks that do not account for sparsity and time series dependence, respectively.

				Lasso					
		λ_{CV}		$\lambda_{CV}/2$		$\lambda_{CV}/4$		$\lambda = 0$	$A = 0$
n	a		NR		NR		NR		
1000	0.25	0.184	0.184	0.16	0.172	0.156	0.16	0.16	2.756
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.16)
	0.75	0.144	5.64	0.184	0.184	0.244	0.244	0.372	9.156
		(0.04)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.04)
5000	0.25	0.18	0.18	0.272	0.272	0.244	0.256	0.224	8.632
		(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.07)
	0.75	0	6	0.144	0.144	0.136	0.136	0.332	8.712
		(0)	(0)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
				CLIME					
		λ_{CV}		$\lambda_{CV}/2$		$\lambda_{CV}/4$			
n	a		NR		NR		NR		
1000	0.25	0.256	0.256	0.184	0.188	0.16	0.152		
		(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)		
	0.75	6.016	6.016	4.776	6.592	0.196	5.244		
		(0.01)	(0.01)	(0.12)	(0.05)	(0.05)	(0.08)		
5000	0.25	0.18	0.18	0.264	0.264	0.256	0.26		
		(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)		
	0.75	0.084	5.028	0.184	0.192	0.264	0.28		
		(0.03)	(0.09)	(0.05)	(0.05)	(0.05)	(0.05)		

Table 4: False Positives and Negatives for a Causal V-Structure. Expected number of true plus false positives (TP+FP), false positives (FP) and false negatives (FN) for the off-diagonal terms of Θ_{11} approximated using 250 Monte Carlo simulations (standard errors in parenthesis). The contemporaneous causal structure is a v-structure with $K = 150$ variables with $\widetilde{K} = 50$ clusters. The number of true positives P in the population is 300, where P+N= 22350. Results are reported for different values of λ , where λ_{CV} is the value obtained using cross-validation and denoted by λ_{CV}. The column $\lambda = 0$ refers to the benchmark that does not account for sparsity.

						Lasso							
		λ_{CV}			$\lambda_{CV}/2$			$\lambda_{CV}/4$			$\lambda = 0$		
n	a	TP+FP	FP	FN	TP+FP	FP	FN	TP+FP	FP	FN	TP+FP	FP	FN
1000	0.25	313.44	13.44	0	2197.832	1897.8	0	9711.5	9411.5	0	22350	22050	0
		(0.33)	(0.33)	(0)	(4.24)	(4.24)	(0)	(7.83)	(7.83)	(0)	(0)	(0)	(0)
	0.75	210.344	4.72	94.376	549.104	249.1	0.024	2109.6	1809.6	0	22350	22050	0
		(0.3)	(0.19)	(0.23)	(1.31)	(1.31)	(0.01)	(3.35)	(3.35)	(0)	(0)	(0)	(0)
5000	0.25	302.52	2.52	0	1472.928	1172.9	0	8488.2	8188.2	0	22350	22050	0
		(0.15)	(0.15)	(0)	(3.14)	(3.14)	(0)	(7.81)	(7.81)	(0)	(0)	(0)	(0)
	0.75	200.096	0	99.904	300	0	0	343.08	43.08	0	22350	22050	0
		(0.03)	(0)	(0.03)	(0)	(0)	(0)	(0.59)	(0.59)	(0)	(0)	(0)	(0)
						CLIME							
		λ_{CV}			$\lambda_{CV}/2$			$\lambda_{CV}/4$					
n	a	TP+FP	FP	FN	TP+FP	FP	FN	TP+FP	FP	FN			
1000	0.25	300.928	0.928	0	1189.848	889.8	0	6638.7	6338.7	0			
		(0.09)	(0.09)	(0)	(3.18)	(3.18)	(0)	(6.96)	(6.96)	(0)			
	0.75	106.144	0	193.856	187.472	13.248	125.7	1024	807.4	83.424			
		(0.21)	(0)	(0.21)	(0.63)	(0.36)	(0.5)	(2.75)	(2.71)	(0.33)			
5000	0.25	300.56	0.56	0	760.752	460.7	0	4570.88	4270.8	0			
		(0.06)	(0.06)	(0)	(2.45)	(2.45)	(0)	(6.59)	(6.59)	(0)			
	0.75	235.48	0.032	64.552	318.344	19.544	1.2	764.216	464.2	0			
		(0.5)	(0.02)	(0.5)	(0.4)	(0.39)	(0.1)	(2)	(2)	(0)			

Table 5: False Positives and Negatives for a Causal V-Structure. Expected number of true plus false positives (TP+FP), false positives (FP) and false negatives (FN) for the off-diagonal terms of Θ_{11} approximated using 250 Monte Carlo simulations (standard errors in parenthesis). The contemporaneous causal structure is a v-structure with $K = 9$ variables with $\widetilde{K} = 3$ clusters. The number of true positives P in the population is 18, where P+N= 72. Results are reported for different values of λ , where λ_{CV} is the value obtained using cross-validation and denoted by λ_{CV}. The column $\lambda = 0$ refers to the benchmark that does not account for sparsity.

		Lasso											
		λ_{CV}			$\lambda_{CV}/2$			$\lambda_{CV}/4$			$\lambda = 0$		
n	a	TP+FP	FP	FN	TP+FP	FP	FN	TP+FP	FP	FN	TP+FP	FP	FN
1000	0.25	22.968	4.968	0	41.584	23.584	0	57.48	39.48	0	72	54	0
		(0.24)	(0.24)	(0)	(0.41)	(0.41)	(0)	(0.34)	(0.34)	(0)	(0)	(0)	(0)
	0.75	12.504	0.016	5.512	18.456	0.456	0	22.168	4.168	0	72	54	0
		(0.06)	(0.01)	(0.06)	(0.06)	(0.06)	(0)	(0.16)	(0.16)	(0)	(0)	(0)	(0)
5000	0.25	18	0	0	20.808	2.808	0	38.04	20.04	0	72	54	0
		(0)	(0)	(0)	(0.16)	(0.16)	(0)	(0.35)	(0.35)	(0)	(0)	(0)	(0)
	0.75	12	0	6	18	0	0	18.12	0.12	0	72	54	0
		(0)	(0)	(0)	(0)	(0)	(0)	(0.03)	(0.03)	(0)	(0)	(0)	(0)
		CLIME											
		λ_{CV}			$\lambda_{CV}/2$			$\lambda_{CV}/4$					
n	a	TP+FP	FP	FN	TP+FP	FP	FN	TP+FP	FP	FN			
1000	0.25	19.256	1.256	0	30.792	12.792	0	46.072	28.072	0			
		(0.11)	(0.11)	(0)	(0.34)	(0.34)	(0)	(0.37)	(0.37)	(0)			
	0.75	6.16	0	11.84	9.456	0.016	8.56	14.104	1.152	5.048			
		(0.05)	(0)	(0.05)	(0.12)	(0.01)	(0.11)	(0.12)	(0.1)	(0.08)			
5000	0.25	18	0	0	19.056	1.056	0	28.872	10.872	0			
		(0)	(0)	(0)	(0.1)	(0.1)	(0)	(0.29)	(0.29)	(0)			
	0.75	13.096	0.04	4.944	19.088	1.096	0.008	22.408	4.424	0.016			
		(0.09)	(0.02)	(0.09)	(0.09)	(0.09)	(0.01)	(0.17)	(0.17)	(0.01)			

Table 6: Average distance between A and $\hat{A}$, Σ_ε and $\hat{\Sigma}_\varepsilon$, respectively, computed over 250 simulations (standard errors in round brackets) when the contemporaneous causal structure is a v-structure for $K = 150$ variables with $\widetilde{K} = 50$ clusters. For each method we report the results obtained also when undersmoohting is performed, i.e., columns $\lambda_{CV}/2$ and $\lambda_{CV}/4$. The column $\lambda = 0$ refers to the benchmark that does not account for sparsity.

		$\left\|A - \hat{A}\right\|_{\text{op}}$						
		Lasso			CLIME			
n	a	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	$\lambda = 0$
1000	0.25	0.567	1.25	2.354	0.684	1.082	2.639	314.10
		(0.003)	(0.003)	(0.006)	(0.006)	(0.003)	(0.012)	(3.106)
	0.75	4.265	1.093	1.095	0.798	3.290	3.812	>1000
		(0.048)	(0.006)	(0.003)	(0.016)	(0.083)	(0.054)	(-)
5000	0.25	0.131	0.369	0.722	0.297	0.307	0.644	24.610
		(0.001)	(0.001)	(0.001)	(0.005)	(0.001)	(0.001)	(0.061)
	0.75	3.604	0.925	0.135	3.555	1.404	0.250	>1000
		(0.016)	(0.001)	(0.001)	(0.101)	(0.048)	(0.001)	(-)
		$\left\|\Sigma_\varepsilon - \hat{\Sigma}_\varepsilon\right\|_{\text{op}}$						
		Lasso			CLIME			
n	a	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	$\lambda = 0$
1000	0.25	0.258	1.394	1.803	0.292	1.143	2.531	0.916
		(0.002)	(0.007)	(0.009)	(0.002)	(0.007)	(0.019)	(0.001)
	0.75	0.430	0.119	0.228	0.118	0.335	0.616	0.331
		(0.006)	(0.001)	(0.0014)	(0.003)	(0.007)	(0.006)	(0.003)
5000	0.25	0.081	0.367	0.511	0.076	0.316	0.582	0.395
		(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
	0.75	0.314	0.039	0.043	0.390	0.184	0.052	0.109
		(0.001)	(0.001)	(0.001)	(0.002)	(0.010)	(0.001)	(0.001)

Table 7: Average distance between A and $\hat{A}$, Σ_ε and $\hat{\Sigma}_\varepsilon$, respectively, computed over 250 simulations (standard errors in round brackets) when the contemporaneous causal structure is a v-structure for $K = 9$ variables with $\widetilde{K} = 3$ clusters. For each method we report the results obtained also when undersmoohting is performed, i.e., columns $\lambda_{CV}/2$ and $\lambda_{CV}/4$. The column $\lambda = 0$ refers to the benchmark that does not account for sparsity.

		$\left\|A - \hat{A}\right\|_{\text{op}}$						
		MB			CLIME			
n	a	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	$\lambda = 0$
1000	0.25	0.243	0.331	0.345	0.307	0.311	0.343	12.716
		(0.004)	(0.004)	(0.004)	(0.007)	(0.004)	(0.004)	(0.122)
	0.75	2.835	0.845	0.182	0.673	1.774	2.186	>1000
		(0.039)	(0.005)	(0.004)	(0.012)	(0.079)	(0.055)	(-)
5000	0.25	0.079	0.097	0.140	0.078	0.091	0.129	10.495
		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.049)
	0.75	2.963	0.853	0.050	3.143	0.840	0.100	>1000
		(0.017)	(0.002)	(0.001)	(0.027)	(0.010)	(0.009)	(-)
		$\left\|\Sigma_\varepsilon - \hat{\Sigma}_\varepsilon\right\|_{\text{op}}$						
		Lasso			CLIME			
n	a	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	λ_{CV}	$\lambda_{CV}/2$	$\lambda_{CV}/4$	$\lambda = 0$
1000	0.25	0.149	0.176	0.169	0.125	0.185	0.179	0.158
		(0.004)	(0.003)	(0.002)	(0.004)	(0.003)	(0.002)	(0.002)
	0.75	0.262	0.051	0.057	0.086	0.180	0.397	0.062
		(0.005)	(0.001)	(0.001)	(0.002)	(0.007)	(0.008)	(0.001)
5000	0.25	0.035	0.060	0.082	0.035	0.050	0.084	0.072
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	0.75	0.257	0.022	0.022	0.415	0.024	0.028	0.027
		(0.002)	(0.001)	(0.001)	(0.004)	(0.001)	(0.002)	(0.001)