



HAL
open science

Frequency causality measures and Vector AutoRegressive (VAR) models: An improved subset selection method suited to parsimonious systems

Christophe Chorro, Emmanuelle Jay, Philippe de Peretti, Thibault Soler

► To cite this version:

Christophe Chorro, Emmanuelle Jay, Philippe de Peretti, Thibault Soler. Frequency causality measures and Vector AutoRegressive (VAR) models: An improved subset selection method suited to parsimonious systems. 2021. halshs-03216938

HAL Id: halshs-03216938

<https://shs.hal.science/halshs-03216938>

Submitted on 4 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CES

Centre d'Économie de la Sorbonne
UMR 8174

**Frequency causality measures and Vector
AutoRegressive (VAR) models: An improved subset
selection method suited to parsimonious systems**

Christophe CHORRO, Emmanuelle JAY,
Philippe DE PERETTI, Thibault SOLER

2021.13



Frequency causality measures and Vector AutoRegressive (VAR) models: An improved subset selection method suited to parsimonious systems

Christophe Chorro^{*1}, Emmanuelle Jay^{2,3}, Philippe De Peretti¹, and Thibault Soler^{1,3}

¹University Paris 1 Panthéon-Sorbonne, Centre d'Economie de la Sorbonne, 106 bd de l'hôpital, 75013, Paris, France

²Quanted & EIF, Palais Brongniart, 28 place de la Bourse, 75002 Paris, France

³Fideas Capital, 21 avenue de l'Opéra, 75001 Paris, France

Abstract

Finding causal relationships in large dimensional systems is of key importance in a number of fields. Granger non-causality tests have become standard tools, but they only detect the direction of the causality, not its strength. To overcome this point, in the frequency domain, several measures have been introduced such as the Direct Transfer Function (DTF), the Partial Directed Coherence measure (PDC) or the Generalized Partial Directed Coherence measure (GPDC). Since these measures are based on a two-step estimation, consisting in i) estimating a Vector AutoRegressive (VAR) in the time domain and ii) using the VAR coefficients to compute measures in the frequency domain, they may suffer from cascading errors. Indeed, a flawed VAR estimation will translate into large biases in coherence measures. Our goal in this paper is twofold. First, using Monte Carlo simulations, we quantify these biases. We show that the two-step procedure results in highly inaccurate coherence measures, mostly due to the fact that non-significant coefficients are kept, especially in parsimonious systems. Based on this idea, we next propose a new methodology (mBTS-TD) based on VAR reduction procedures, combining the modified-Backward-in-Time selection method (mBTS) and the Top-Down strategy (TD). We show that our mBTS-TD method outperforms the classical two-step procedure. At last, we apply our new approach to recover the topology of a weighted financial network in order to identify through the local directed weighted clustering coefficient the most systemic assets and exclude them from the investment universe before allocating the portfolio to improve the return/risk ratio.

Keywords: VAR model, subset selection methods, frequency causality measures, weighted financial networks, portfolio allocation

*Corresponding author: christophe.chorro@univ-paris1.fr

1 Introduction

Vector AutoRegressive (VAR) models are popular models used for analyzing multivariate time series. They have been widely used in many fields, such as macroeconomics [1], finance [2], or even neuroscience [3]. Their use comes from their simplicity and straightforward theoretical framework for understanding the dynamical structure of systems, capturing complex temporal relationships among time series. In dynamical systems admitting a VAR representation, it is often of interest to capture and quantify complex internal dynamics. These complex interactions can be estimated by Granger non-causality tests [4]. Given an information set, non-causality tests check whether or not adding past values of univariate or multivariate series significantly reduces the forecast error variance. Nevertheless, if causality is detected (rejection of the non-causality hypothesis), Granger's tests do not give any information about the strength of this causality. They only assess the existence and direction of causal relationships. But in many applications such as weighted graphs or networks, the quantification of causal strength is of great importance. To deal with this point, the concept of Granger causality has been extended in the frequency domain by considering several indicators measuring the causal strength or coupling similarities. The most popular measures are the Direct Transfer Function measure (DTF) [5, 6], the Partial Directed Coherence measure (PDC) [7, 8], and the Generalized Partial Directed Coherence measure (GPDC) [9]. Such indicators are computed using a two-step approach: first estimate a VAR model of lag p , then switching to the frequency domain using Fourier transform of the estimated VAR coefficients to compute the indicator of interest. In such an approach, it is obvious that a flawed estimation of the VAR model will translate into inaccurate measures of the DTF, PDC, or GPDC. We will refer to this aspect as cascading errors.

A common way to estimate a VAR is to rely on a suitable estimation method, and then to use information criteria such as the Akaike Information Criterion (AIC) [10, 11, 12] or the Bayesian Information Criterion (BIC) [13] to select the correct lag. This procedure raises two issues: the first concerns the ability of the information criteria to correctly approximate the true lag p of the underlying data generating process [2, 14], especially for small samples. The second one refers to the significance of individual VAR parameters. By construction, in VAR models, time series depend on all lagged variables in the system. This assumption is very strong and unrealistic in most applications. Indeed, most systems will admit parsimonious structures with only a few significant coefficients. Said differently, in a multivariate system, it is unusual for all time series to be mutually dependent at each lag. Thus, if information criteria are used solely to estimate p , then when computing coherence measures, non-significant coefficients are likely to be used, thus biasing the causality measures.

Our contribution to the literature on coherence measures is twofold. First, we estimate the impact of cascading errors on the accuracy of computed coherence measures within a standard VAR estimation. We implement Monte Carlo simulations using a system with five time series, with $p = 3$. This system has been analyzed with PDC and DTF methods in [8, 15, 16]. It admits 12% non-zero coefficients with only one on the third lag. This toy model is quite stringent, and likely to be found in real systems such as macroeconomics [17] and finance [18, 19, 20]. Through simulations, we also compute cascading errors for several sample sizes and residual correlation matrices (non-diagonal matrices) to present

a realistic framework. We show that standard VAR estimation leads to highly biased coherence measures, since non-significant coefficients appear in the coherence measures, as mentioned above. Therefore, it is straightforward to investigate if more advanced subset selection method can solve the issue raised. Accordingly, our major second contribution checks whether or not more advanced VAR model selection methods leading to more parsimonious representations could lower or even suppress the cascading errors. In the literature, three types of procedures may be considered. The first type reduces the number of VAR coefficients by adding/deleting parameters using information criteria. These include Top-Down (TD), Bottom-Up (BU) [14], or a more recent method based on the BU approach called modified Backward-in-Time Selection (mBTS) [18, 21]. The second type of procedures is based on hypothesis testing, either at an individual level, i.e. on each coefficient taken separately such as t -test (hereafter TT) or for a group of coefficients such as likelihood ratio and Wald tests [14, 22, 23]. Finally, the third procedure relies on shrinkage methods, such as Lasso (Least Absolute Shrinkage and Selection Operator) [24, 17]. In this latter, selection of variables and VAR coefficients estimation are conducted simultaneously. We evaluate each approach and propose a new and extended method by combining the mBTS method and the TD strategy (mBTS-TD). We show, in the framework of coherence measures, that our approach outperforms the competing ones.

Based on this result, we apply our methodology combining mBTS-TD and GPDC measure to build financial networks. This application emphasizes the advantages of our methodology. Firstly, the mBTS-TD method inherently provides a parsimonious structure without using network dimensionality reduction tools such as a significant threshold, the Minimum Spanning Tree [25] or Planar Maximally Filtered Graphs [26]. Secondly, the strength of the causal relationship allows us to recover the network topology in a more precise way. In such a financial network, we propose a new portfolio allocation method by excluding the most systemic assets, identified using the local directed weighted clustering coefficient [27]. We therefore obtain financial portfolio as performing as possible, where the non-systemic assets are equally allocated [28]. Related performance measures are compared with those obtained using a classical VAR estimation to compute the GPDC or allocating the whole universe.

This paper is set out as follows: in section 2, we introduce the econometric methodology (VAR model and Granger non-causality tests) and coherence measures; in section 3, we estimate the cascading errors and show that the use of standard VAR estimation leads to large errors in the coherence measures; in section 4, we begin by introducing advanced VAR estimation procedures and ascertaining their efficiency when used to compute coherence measures; in section 5, we apply our methodology on financial time series and finally, section 6 concludes and discusses our results.

Notations: Bold capital letters, such as \mathbf{A} , represent matrices and bold letters like \mathbf{v} represent column vectors. If \mathbf{A} is a matrix of size $m \times m$, a_{jk} is the element j,k of \mathbf{A} . $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_2$ are the L_1 -norm and L_2 -norm, respectively. \mathbf{I}_m is the $m \times m$ identity matrix. $\mathbf{0}_m$ and $\mathbf{1}_m$ are $m \times 1$ vectors of zeros and ones. \mathbf{A}' is the transpose of \mathbf{A} . The determinant of \mathbf{A} is denoted by $\det(\mathbf{A})$. For any complex z , z^* is the conjugate of z , and i is the square root of -1 . The notation vec is the column stacking operator. \otimes is the Kronecker product. $|\cdot|$ denotes the absolute value or modulus for complex numbers. $\mathcal{N}(\mu, \sigma^2)$ and $\chi^2(k)$ are the Gaussian and Chi-squared distributions, respectively.

2 Econometric methodology

In this section, we first introduce Vector AutoRegressive models (VAR) and model order identification. Then we present the concept of Granger causality and coherence measures, paying particular attention to Generalized Partial Directed Coherence (GPDC). Furthermore, we discuss the reasons why focusing on coherence measures rather than Granger causality can improve information about causal strength. Finally, we address the accuracy of the coherence measure, since this may present cascading errors due to its first step in the VAR estimation.

2.1 Non-causality test in Vector AutoRegressive Models

Let $\mathbf{x}(t) = (x_1(t), \dots, x_m(t))'$ be a zero-mean m -dimensional stationary process admitting the following VAR(p) representation (see [14] section 2 for classical stability and stationarity conditions):

$$\mathbf{x}(t) = \mathbf{A}_1\mathbf{x}(t-1) + \dots + \mathbf{A}_p\mathbf{x}(t-p) + \boldsymbol{\epsilon}(t), \quad t \in \mathbb{Z} \quad (1)$$

where $\mathbf{A}_1, \dots, \mathbf{A}_p$ are $(m \times m)$ coefficient matrices, p is the model order, and $\boldsymbol{\epsilon}(t) = (\epsilon_1(t), \dots, \epsilon_m(t))'$ is a $(m \times 1)$ vector of white noises with $E[\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}'(s)] = 0$ for $t \neq s$ and $\boldsymbol{\epsilon}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$.

The coefficient matrices $\mathbf{A}_1, \dots, \mathbf{A}_p$ describe the temporal relationships within the m time series in the system. The concept of causality is therefore directly related to these coefficients. These coefficient matrices also play a fundamental role when making forecasts. The structure of $\boldsymbol{\Sigma}_\epsilon$ reveals the contemporaneous or instantaneous effects between the time series.

In this paper, the VAR coefficients are estimated using the Least Squares estimator (LS), either in a multivariate (LS) or in univariate (OLS) environment (equation by equation) [14]. In addition to estimating the VAR coefficients, the model order p must also be estimated. This step is crucial for the accuracy of the VAR estimate. The lag order p is chosen to minimize an information criterion such as AIC [10, 11] or BIC [13]. In this paper, we choose AIC and BIC by using the two estimators of $\boldsymbol{\Sigma}_\epsilon(p)$ to investigate the role of the penalty factor in model order selection. For VAR(p) in (1), AIC and BIC are defined as follows:

$$\begin{aligned} \text{AIC}(p) &= \ln \left(\det \left(\widehat{\boldsymbol{\Sigma}}_\epsilon(p) \right) \right) + \frac{2}{T}pm^2 \\ \text{BIC}(p) &= \ln \left(\det \left(\widehat{\boldsymbol{\Sigma}}_\epsilon(p) \right) \right) + \frac{\ln T}{T}pm^2 \\ \text{AIC}_{un}(p) &= \ln \left(\det \left(\widetilde{\boldsymbol{\Sigma}}_\epsilon(p) \right) \right) + \frac{2}{T}pm^2 \\ \text{BIC}_{un}(p) &= \ln \left(\det \left(\widetilde{\boldsymbol{\Sigma}}_\epsilon(p) \right) \right) + \frac{\ln T}{T}pm^2 \end{aligned}$$

where T is the number of observations, $\widetilde{\boldsymbol{\Sigma}}_\epsilon(p)$ and $\widehat{\boldsymbol{\Sigma}}_\epsilon(p)$ are the unbiased and the maximum likelihood estimates of $\boldsymbol{\Sigma}_\epsilon(p)$ for the VAR(p) in (1), given by

$$\widehat{\boldsymbol{\Sigma}}_\epsilon(p) = \frac{\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}'(t)}{T} \quad (2)$$

$$\tilde{\Sigma}_\epsilon(p) = \frac{\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}'(t)}{T - pm - 1} \quad (3)$$

The main difference between AIC and BIC is the increase in the BIC penalty factor compared to AIC.

After estimating the VAR, the most common way to assess complex interactions is to use Granger non-causality tests. The concept of non-causality defined by Granger [4] is based on the idea that, if a time series $x_k(t)$ causes another time series $x_j(t)$, then the past of $x_k(t)$ will significantly decrease the forecast error in $x_j(t)$. Let $\mathbf{x}(t)$ be a m -dimensional stationary process admitting the VAR(p) representation defined in (1). Granger non-causality can be tested by using a Wald multiple restrictions test [14] on the VAR coefficients. This test, jointly tests whether a set of coefficients are non-significant. For example, if a time series $x_k(t)$ does not *Granger-cause* $x_j(t)$, then $a_{kj}(1) = \dots = a_{kj}(p) = 0$, where $a_{kj}(1), \dots, a_{kj}(p)$ are the elements k, j of the matrices $\mathbf{A}_1, \dots, \mathbf{A}_p$. The general null hypothesis is given by $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{c}$, where \mathbf{C} is a $(q \times m^2p)$ matrix called the restriction matrix of the VAR coefficients (1 for tested coefficients and 0 otherwise). Moreover, q denotes the number of restrictions, $\boldsymbol{\beta}$ is a $(m^2p \times 1)$ vector with $\boldsymbol{\beta} = \text{vec}(\mathbf{A}_1, \dots, \mathbf{A}_p)$, and \mathbf{c} is a $(q \times 1)$ vector with $\mathbf{c} = \mathbf{0}_q$ for Granger non-causality.

The Wald statistic is therefore

$$\Gamma = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{c})' \left[\mathbf{C} \left((\mathbf{X}\mathbf{X}')^{-1} \otimes \tilde{\Sigma}_\epsilon \right) \mathbf{C}' \right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{c}) \quad (4)$$

where

- $\hat{\boldsymbol{\beta}}$ is the $(m^2p \times 1)$ vector of the estimated VAR coefficients $\boldsymbol{\beta}$,
- \mathbf{X} is the $(mp \times T)$ matrix with $\mathbf{X}(t) = \text{vec}(\mathbf{x}(t-1), \dots, \mathbf{x}(t-p))$,
- $\tilde{\Sigma}_\epsilon$ is the unbiased estimator of $\Sigma_\epsilon(p)$ defined in (3).

Under the null hypothesis H_0 : $\Gamma \sim \chi^2(q)$. This result is valid only asymptotically and for the VAR model assumptions defined in (1) with $\Sigma_\epsilon = \sigma^2 \mathbf{I}_m$. H_0 is not rejected (non-causality) for a given probability α if $\Gamma \leq \chi_\alpha^2(q)$, where $\chi_\alpha^2(q)$ is the quantile of the distribution. Nevertheless, if the null hypothesis H_0 is rejected, it means that a time series $x_k(t)$ *Granger-causes* another time series $x_j(t)$. This test does not provide any information about the causal strength. To deal with this aspect, several measures called coherence measures have been proposed in the frequency domain.

2.2 Coherence measures

Coherence measures describe the connectivity between times series in the frequency domain and are often used in the neurosciences to understand functional connectivity patterns between different brain regions. The most popular coherence measures are the Directed Coherence measure (DC) [29], the Partial Directed Coherence measure (PDC) [7, 8], the Direct Transfer Function measure (DTF) [5, 6], and the Generalized Partial Directed Coherence measure (GPDC) [9]. These measures are based on a two-step approach: first the VAR coefficients are estimated, and then the measure is computed using the transfer function matrix or its inverse matrix on the VAR coefficients. The DC, introduced by Saito and Harashima [29] for bivariate cases, describes whether and how two time

series are functionally connected. The three other measures PDC, GPDC, and DTF can be applied to multivariate cases. The PDC introduced by Baccalá and Sameshima [7, 8] provides a frequency domain representation of Granger Causality. It is a generalization to the multivariate case of the DC, based on the Partial Coherence that describes the mutual interaction between two time series when the effects of all others have been subtracted. In other words, it quantifies only the direct connections between time series. Baccalá and Sameshima [9] have extended their measure, called GPDC, by taking into account the variance of white noise, so that it is more accurate with finite time series samples and leads to a scale invariant measure. Finally, the DTF was introduced by Kamiński and Blinowska [5, 6]. It describes the causal influence, but it does not distinguish direct and indirect relationships, whereas PDC and GPDC provide the multivariate relationships from a partial perspective. In this paper we focus on PDC and GPDC measures because in a multivariate environment the distinction between indirect and direct relationships is of great importance.

For two time series $x_j(t)$ and $x_k(t)$ the GPDC is defined so as to exhibit the causality from k to j at each frequency f as follows:

$$\omega_{jk}(f) = \frac{\frac{1}{\sigma_{jj}} \tilde{a}_{jk}(f)}{\sqrt{\sum_{n=1}^m \frac{1}{\sigma_{nn}^2} \tilde{a}_{nk}(f) \tilde{a}_{nk}^*(f)}} \quad (5)$$

where

- f are the discrete frequencies¹ lying in $\left[-\frac{1}{2}; \frac{1}{2}\right]$,
- $\tilde{a}_{jk}(f)$ is the discrete Fourier transform of the coefficients $a_{jk}(1), \dots, a_{jk}(p)$ defined by

$$\tilde{a}_{jk}(f) = \begin{cases} 1 - \sum_{l=1}^p a_{jk}(l) e^{-2i\pi fl}, & \text{if } j = k \\ - \sum_{l=1}^p a_{jk}(l) e^{-2i\pi fl}, & \text{otherwise} \end{cases}$$

- σ_{jj}^2 is the j -th element of the diagonal of Σ_ϵ .

The GPDC $\omega_{jk}(f)$ at frequency f represents the relative strength of interaction with respect to a given signal source. Note that the GPDC is represented as a power spectral density, i.e., $|\omega_{jk}(f)|^2$. Moreover, given that the VAR coefficients $a_{jk}(l)$ are real numbers, their Discrete Transform Fourier has Hermitian symmetry $\tilde{a}_{jk}(f) = \tilde{a}_{jk}(-f)^*$. The spectrum is symmetric at the frequency $f = 0$, i.e., $|\tilde{a}_{jk}(f)| = |\tilde{a}_{jk}(-f)|$, so it is possible to

¹For a discrete time series sampled at frequency f_e , its Fourier Transform will reveal information for frequencies lying in $\left[-\frac{f_e}{2}; \frac{f_e}{2}\right]$. In our case $f_e = 1$, we can therefore choose the interval $\left[-\frac{1}{2}; \frac{1}{2}\right]$ with a step of $\frac{1}{F-1}$, where F is the number of frequencies.

represent only a half-period of the spectrum ($f \in \left[0; \frac{1}{2}\right]$). Finally, the GPDC satisfies the following properties :

$$0 \leq |\omega_{jk}(f)|^2 \leq 1 \quad (6)$$

$$\sum_{n=1}^m |\omega_{nk}(f)|^2 = 1, \forall k = 1, \dots, m. \quad (7)$$

In such an approach, it is obvious that a flawed estimation of the VAR model will translate into inaccurate measures of GPDC, but it is also true for all coherence measures that directly use VAR coefficients in the transfer function matrix or its inverse.

The VAR estimation may be incorrectly performed for many well-known reasons, such as incorrect model order selection, small sample size, or correlated residuals. Moreover, an obvious error that is often omitted when computing coherence measures and also generates cascading errors is the estimation of zero coefficients due to VAR estimation in a multivariate environment. In other words, in a multivariate system it is unusual for all time series to be mutually dependent (parsimonious model). Some coefficients are therefore equal to zero, and the estimation of these zero coefficients (non-causal terms) inevitably biases the non-zero ones (causal terms). This is particularly true for PDC and GPDC, which compute only direct causality, and due to their normalization property (7), the errors made in the non-causal terms have a direct impact on the accuracy of the causal terms. Thus, in the presence of parsimonious VAR models and by combining the well-known estimation errors, high cascading errors and spurious causalities become likely.

3 Impacts of standard VAR estimation on GPDC

In this section, we use Monte Carlo simulations to illustrate the way estimation of VAR model in a multivariate environment can impact GPDC accuracy when the underlying system is parsimonious. The simulations are conducted by varying both the sample size and the residual correlation matrix in order to cover more realistic examples, but also to highlight standard VAR estimation errors on the GPDC. Firstly, the system and settings are presented. Then we focus on the errors in the VAR coefficients, and finally on the GPDC errors by separating the causal and non-causal terms to determine the part that will be most impacted by cascading errors.

3.1 System and error measures

In the simulation study, we analyze five time series generated by the VAR(3) model used in [8, 15, 16]. This system admits a parsimonious structure with 12% of the m^2p coefficients being non-zero. It has only one coefficient on the third lag. Hence, this system is in principle not suitable for a standard VAR estimation where the m^2p coefficients are estimated.

The VAR model [8, 15, 16] is as follows:

$$\begin{aligned}
 x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + \epsilon_1(t) \\
 x_2(t) &= 0.5x_1(t-2) + \epsilon_2(t) \\
 x_3(t) &= -0.4x_1(t-3) + \epsilon_3(t) \\
 x_4(t) &= -0.5x_1(t-2) + 0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \epsilon_4(t) \\
 x_5(t) &= -0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \epsilon_5(t)
 \end{aligned} \tag{S}$$

The causal structure of S is shown in Fig. 1 and the theoretical GPDC in Fig. 2.

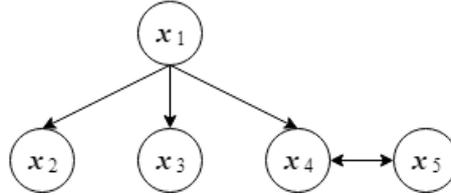


Fig 1: Causal structure of (S). In this system, the time series \mathbf{x}_1 causes \mathbf{x}_2 , \mathbf{x}_3 , and \mathbf{x}_4 , while \mathbf{x}_4 and \mathbf{x}_5 are causing each other.

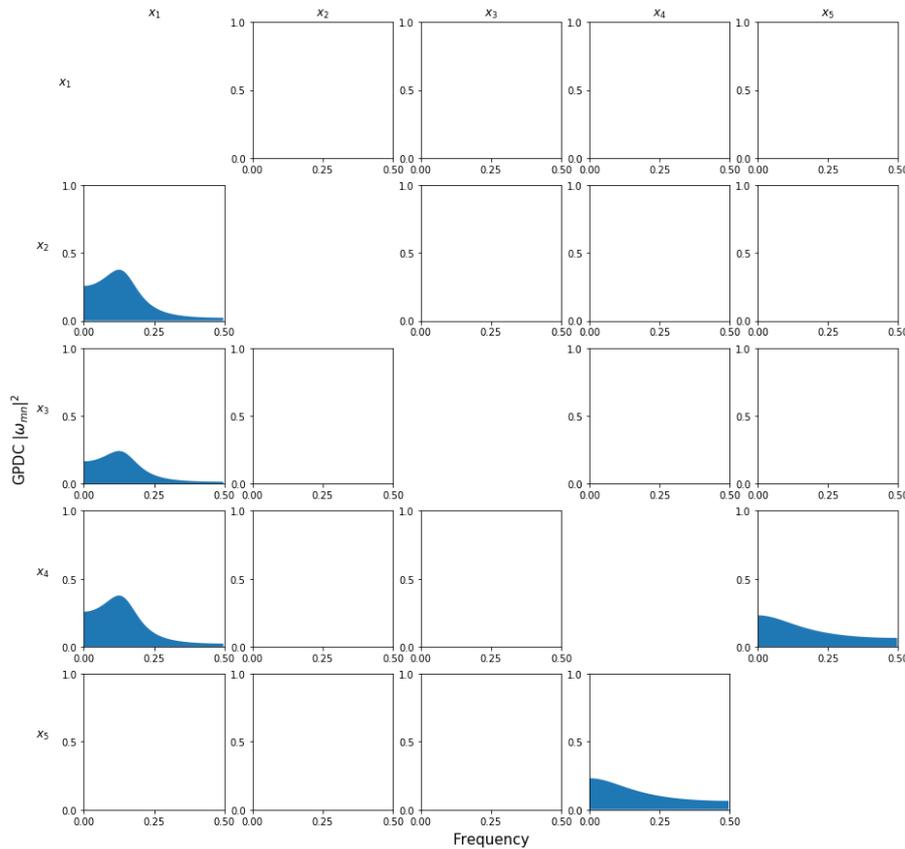


Fig 2: Theoretical GPDC of (S) (column k causes row j). This is computed using the true coefficients and the identity matrix for the residual correlation matrix ($\Sigma_\epsilon = \mathbf{I}_5$). Interpreting the GPDC: \mathbf{x}_1 causes \mathbf{x}_2 , \mathbf{x}_3 , and \mathbf{x}_4 . In contrast, \mathbf{x}_1 causes \mathbf{x}_5 indirectly via \mathbf{x}_4 , but as GPDC quantifies only direct interactions, the causality values for all frequencies are equal to zero. If $j = k$, the GPDC represents the part that is not explained by other signals. Since it is quite difficult to interpret, the diagonal is not reported here.

The impacts of the different VAR models are evaluated through 1000 simulations of (S). For each simulation, (S) is generated by using a sample size T and a multivariate Gaussian distribution for the white noise with $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma_\epsilon)$. The initial values used to generate (S) are set to zero. For T and Σ_ϵ , we use the following settings:

- Four sample sizes: $T = \{128, 256, 512, 1024\}$
- Four Σ_ϵ matrices, the identity and three symmetric Toeplitz matrices with $\rho \in \{0.25, 0.50, 0.75\}$. A $(m \times m)$ symmetric Toeplitz matrix (\mathbf{Tp}) has the form $(\mathbf{Tp})_{jk} = \rho^{|j-k|}$ for $j, k = 1, \dots, m$, with $\rho \in \mathbb{R} : |\rho| < 1$. The four matrices are as follows:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0.25 & 0.06 & 0.02 & 0 \\ 0.25 & 1 & 0.25 & 0.06 & 0.02 \\ 0.06 & 0.25 & 1 & 0.25 & 0.06 \\ 0.02 & 0.06 & 0.25 & 1 & 0.25 \\ 0 & 0.02 & 0.06 & 0.25 & 1 \end{pmatrix}$$

(1) Id

(2) Tp_1

$$\begin{pmatrix} 1 & 0.50 & 0.25 & 0.13 & 0.06 \\ 0.50 & 1 & 0.50 & 0.25 & 0.13 \\ 0.25 & 0.50 & 1 & 0.50 & 0.25 \\ 0.13 & 0.25 & 0.50 & 1 & 0.50 \\ 0.06 & 0.13 & 0.25 & 0.50 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0.75 & 0.56 & 0.42 & 0.32 \\ 0.75 & 1 & 0.75 & 0.56 & 0.42 \\ 0.56 & 0.75 & 1 & 0.75 & 0.56 \\ 0.42 & 0.56 & 0.75 & 1 & 0.75 \\ 0.32 & 0.42 & 0.56 & 0.75 & 1 \end{pmatrix}$$

(3) Tp_2

(4) Tp_3

The increase in ρ indicates that the error processes are more strongly correlated, and allows us to see the estimation impacts when the system approaches a structural VAR model. To build the covariance matrices, we use the symmetric Toeplitz matrix [30, 19] because it provides a flexible framework to generate a positive-definite covariance matrix. The symmetric Toeplitz matrix structure used depends on just one parameter ρ . Moreover, it reflects the stationarity of auto-regressive systems (if $|\rho| < 1$).

To compare and quantify the cascading errors of the different VAR models, the relative L_2 -norm error is used for the VAR coefficients, whereas the L_2 -norm error is computed for GPDC. The L_2 -norm error is used instead of the relative one because, for non-causal terms, the theoretical GPDC is null for all frequencies, resulting in a null denominator for the relative error. The two error metrics are defined as follows:

- Relative L_2 -norm error of VAR coefficients: $\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 / \|\mathbf{A}\|_2$ where $\mathbf{A} = (\mathbf{A}_1, \dots, \mathbf{A}_p)$
- L_2 -norm error of GPDC: $\| |\widehat{\omega}_{jk}|^2 - |\omega_{jk}|^2 \|_2$ on each pair $j \neq k$, where ω_{jk} is the vector containing each value of $\omega_{jk}(f)$ for all discrete frequencies f .

Finally, to assess the errors through the 1000 simulations, we report for the VAR coefficients the median of the relative L_2 -norm error and for the GPDC the sum of the median of the L_2 -norm error on both the causal terms and the non-causal terms.

3.2 Estimation errors

Here, we evaluate the impact of standard VAR estimations through 1000 Monte Carlo simulations of (S) on both the VAR coefficient and the GPDC errors. Fig. 3 shows the first results for the simulated data, focusing on the estimation errors of the VAR coefficients. For this purpose, five VAR models are estimated in a multivariate environment, either by setting the order of the model at $p = 3$ (true model order), or by determining it using the four information criteria defined in section 2. Note that the Tp_1 results are not reported because they are quite similar to the identity matrix, but they are available upon request.

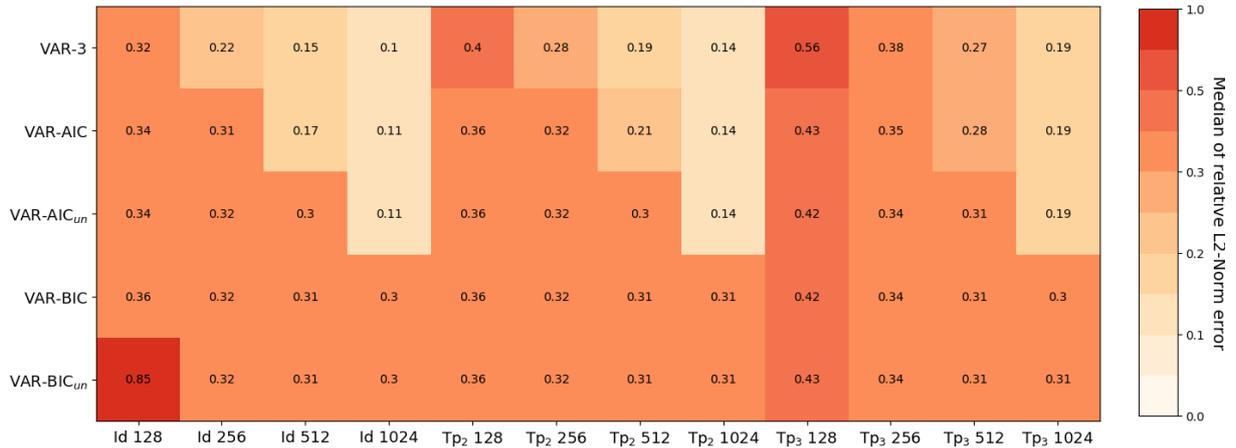


Fig 3: Median over 1000 simulations of relative L_2 -norm error on the coefficients estimated using the five standard VAR estimations. The increase in the median indicates a deterioration in the estimation of coefficients.

On the basis of these first empirical results presented in Fig. 3, the following conclusions can be drawn:

- The accuracy of the estimated coefficients improves significantly when the sample size increases, but this improvement is less prominent when using BIC criteria. For the VAR-3 or the two VAR-AIC, the errors are almost reduced by a factor of three between $T = 128$ and $T = 1024$. Moreover, the performance for all models deteriorates when the residuals are strongly correlated.
- In VAR estimation the selection of the model order is crucial for accuracy. The four information criteria rarely find the true model order for small sample sizes ($T = 128, 256$) and correlated residuals (Tp_2 and Tp_3). For the smallest sample sizes, they select the true lag order $p = 3$ in 7% of cases and otherwise they have a lag order $p = 2$, except for BIC_{un} , which finds a model order $p = 1$ for Id 128 in 70% of simulations. Thus, in these cases, VAR models using information criteria must have higher errors than VAR-3. Nevertheless, for three cases Tp_2 128, Tp_3 128, and Tp_3 256 it is in fact VAR-3 that has the worst results. Indeed, as (S) has only one non-zero coefficient on the third lag, the VAR-3 estimates twenty-four zero coefficients, which significantly increases the error. For a parsimonious system like (S), the true model order can become an over-fitted model, especially for small sample sizes. However, the model order must not be too seriously undervalued, as happens with $VAR-BIC_{un}$ for Id128, otherwise a compensation effect will occur in the values of the coefficients, overestimating them by significantly increasing the errors.

- On this system, the AIC criterion provides better results for the LS estimation, because it is the least restricted and so can find the true model order for large sample sizes ($T = 512, 1024$). However, it can also under-fit the model to yield lower errors than VAR-3 for small sample sizes and correlated residuals.

The following two Figs. 4 and 5 show the GPDC errors on causal (Fig. 4) and non-causal terms (Fig. 5), using the same models and settings as before.



Fig 4: Sum of medians over 1000 simulations of L_2 -norm error on the causal GPDC (5) estimated using the five standard VAR estimations. The increase in the median indicates a deterioration in the estimation of the causal GPDC.



Fig 5: Sum of medians over 1000 simulations of L_2 -norm error on the non-causal GPDC (5) estimated using the five standard VAR estimations. The increase in the median indicates a deterioration in the estimation of the non-causal GPDC.

For causal terms (Fig. 4), the same conclusions can be drawn as for the errors in the coefficients, whereas for non-causal terms, the results are quite different (Fig. 5). For non-causal terms (zero coefficients), VAR-3 presents the worst errors for all settings, except for $T = 1024$, where errors are similar to the other VAR models. These errors in VAR-3 support the idea of an over-fitted model because non-causal terms are directly related to the estimation of zero coefficients. In contrast, VAR models using the most restricted information criteria (AIC_{un}, BIC, and BIC_{un}) become the best models for all

cases, due to the non-estimation of the third lag. Nevertheless, if the model is under-fitted (VAR-BIC_{un}), as noted with the errors in the coefficients, the errors will also be very high (4.22 for Id128) due to compensation effects in the coefficients. We also note that, in the case where the VAR model contains many zero coefficients (88% of zero coefficients), the errors in non-causal terms can be higher than those in the causal terms. Moreover, due to the normalization property defined in (7), the errors in the non-causal terms have a direct impact on the accuracy of the causal strength. Thus, in the presence of parsimonious VAR models, the standard VAR estimation with all possible information criteria is in fact not well suited.

In this section, we have confirmed that standard VAR estimation provides high cascading errors and reveals spurious causalities. Indeed, a VAR estimation in a multivariate environment does not take into account the parsimonious structure of the underlying model, and better suited methods are needed. Thus, in order to correct the cascading errors in the GPDC and to adopt a more parsimonious structure, we use searching procedures with parameter constraints, or shrinkage methods, called subset selection methods.

4 Improving GPDC estimation accuracy

Standard VAR estimation is therefore not well-suited to estimate parsimonious data generating processes, since such processes may have only a few non-zero coefficients. The significance of individual coefficients must then be assessed prior to computing GPDC functions. To deal with this point, it is possible to use subset selection methods. In the literature, at least three procedures have been considered: first, procedures based on information criteria to add or delete coefficients such as Bottom-Up strategy, Top-Down strategy [14] and modified Backward-in-Time Selection [21]; second, procedures based on hypothesis testing, such as t -test for individual coefficients and multivariate tests [14, 22, 23], i.e. that jointly test the coefficients (e.g. likelihood ratio and Wald test); finally, procedures based on shrinkage methods such as Lasso Regression [24]. However, for the second family, the multivariate approach aims at testing the non-significance of a set of coefficients, and is therefore not of particular interest in our setting (such a drawback is also found in the classical Bottom-Up strategy [14]) explaining why we only focus on the individual t -test (TT) in our study. The results are assessed through VAR coefficients errors and GPDC accuracy as in the previous section by comparing our mBTS-TD method to the mBTS method, the TD strategy, the TT procedure (described in Appendix A.1), and the Lasso method (described in Appendix A.2). Finally, we check error distributions and the identification of the true causal structure of (S).

4.1 Proposed method: mBTS-TD

The proposed method first uses the modified Backward-in-Time Selection (mBTS) to estimate the VAR coefficients. The main advantage of the mBTS method is that only terms that improve the prediction of the equation are included, and this allowing us to work with high-dimensional systems like $K = 20$ in [21]. As already shown in [31], the mBTS method dramatically improves GPDC accuracy. Nevertheless, a drawback with the mBTS method is that the maximum lag p_{max} is fixed *a priori*. If it is too small, the internal dynamics of the system are not completely modeled and the model is under-fitted. In this case, the coefficients are over-estimated due to the compensation effects

causing possibly large errors as remarked for example in the last line of Fig. 3. On the other hand, if p_{max} is too large, undesirable lagged variables may appear in the model by revealing spurious causalities. We therefore propose to combine the mBTS method with the Top-Down strategy (TD). Indeed, the advantage of the TD strategy is that all coefficients in each equation are tested, but it is very sensitive to the initial VAR estimation which determines the model order and thus can amplify the errors if the initial VAR estimation is not carried out properly. The two reasons to combine the mBTS method and TD strategy are: firstly, to be less dependent on the choice of p_{max} so that we may set its values high enough to capture all possible connections; and secondly, as a further consequence, so that we may produce a more parsimonious model when p_{max} is set at a high value. Hereafter, we define the mBTS method and the TD strategy.

The modified Backward-in-Time Selection (mBTS) is a Bottom-Up strategy (BU) introduced by Vlachos and Kugiumtzis [21]. It is based on Dynamic Regression models [18], which estimate each equation separately. Unlike the TD strategy, the mBTS method adds progressively lagged variables, starting from the first lag for all variables, and moving backward in time.

First, a maximum order p_{max} is fixed, and this provides the vector $(1 \times mp_{max})$ of all lagged variables for the j -th equation of the VAR(p) model in (1):

$$\mathbf{v} = (x_1(t-1), \dots, x_1(t-p_{max}), \dots, x_m(t-1), \dots, x_m(t-p_{max}))$$

An explanatory vector $\boldsymbol{\vartheta}$ is built from \mathbf{v} by progressively adding only the most significant lagged variable at each step.

For the j -th equation of the VAR(p) model in (1), the mBTS algorithm is as follows:

1. Start with an empty vector $\boldsymbol{\vartheta} = \emptyset$, the information criterion IC^{old} initialized to the variance of the j -th series, and $\boldsymbol{\tau} = (1, \dots, 1)'$ the $(m \times 1)$ lag order vector of the variables.
2. Compute IC_n^{new} relative to the m dynamic regression models formed by the m candidate explanatory vectors $\boldsymbol{\vartheta}_n^{\text{cand}}$, where $\boldsymbol{\vartheta}_n^{\text{cand}} = (\boldsymbol{\vartheta}, x_n(t-\tau_n))$, $\forall n \in \{1, \dots, m\}$.
3. Select the variable according to the IC value:
 - If $\min\{IC^{\text{old}}, IC_1^{\text{new}}, \dots, IC_m^{\text{new}}\} = IC^{\text{old}}$, then $\boldsymbol{\tau} = \boldsymbol{\tau} + \mathbf{1}_m$.
 - If $\min\{IC^{\text{old}}, IC_1^{\text{new}}, \dots, IC_m^{\text{new}}\} = IC_n^{\text{new}}$, then $IC^{\text{old}} = IC_n^{\text{new}}$, $x_n(t-\tau_n)$ is added to the explanatory vector $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}, x_n(t-\tau_n))$ and only τ_n is increased by one.
4. Repeat steps 2 and 3 until $\boldsymbol{\tau} = (p_{max}, \dots, p_{max})'$.

Finally, the Top-Down strategy (TD) tests the VAR coefficients separately in the m equations. The goal is to eliminate the non-significant coefficients for each equation by evaluating the information criterion. The order of the tested terms is arbitrary, but as in [14], the largest lag p is tested first for all variables from $x_m(t-p)$ to $x_1(t-p)$, then the lag $p-1$ with the same order of variables, and the process is iterated until $p=1$.

For the j -th equation obtained with the mBTS algorithm, the TD strategy is applied as follows:

1. Start with the vector $\boldsymbol{\vartheta}$ and the information criterion IC^{old} obtained with the mBTS algorithm.
2. Sort the vector $\boldsymbol{\vartheta}$ from the largest to the smallest lag p and for all series from x_m to x_1 .
3. Compute IC_n^{new} by deleting the n -th element in the vector $\boldsymbol{\vartheta}$, $\boldsymbol{\vartheta}_n^{\text{cand}} = \boldsymbol{\vartheta} \setminus \{\boldsymbol{\vartheta}_n\}$.
4. Delete the variable according to the IC_n^{new} value:
 - If $\min\{IC^{\text{old}}, IC_n^{\text{new}}\} = IC^{\text{old}}$, then $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}$.
 - If $\min\{IC^{\text{old}}, IC_n^{\text{new}}\} = IC_n^{\text{new}}$, then $IC^{\text{old}} = IC_n^{\text{new}}$ and $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_n^{\text{cand}}$.
5. Repeat steps 3 and 4 $\forall (\boldsymbol{\vartheta}_n)_{n \in [1, |\boldsymbol{\vartheta}|]}$, where $|\cdot|$ denotes the cardinality of the vector in this case.

4.2 Comparison with standard VAR

In this part, subset selection methods are compared to each other, and also to the standard VAR estimation presented in the previous section. The impacts are evaluated on the accuracy of both the VAR coefficients and the GPDC through 1000 Monte Carlo simulations on (S). Twenty nine different VAR models are estimated. First, the five VAR models presented in section 3, and then the six subset methods with the four information criteria. For the mBTS and mBTS-TD methods we set $p_{\text{max}} = 6$, but in section 4.3 we will check the robustness of these methods by testing the stability with respect to p_{max} . The t -test procedure (TT) is used for both a significance level of 5% and 1% (see in A.1). The Lasso tuning parameter is estimated using 5-fold cross-validation (see in A.2). Moreover, only the best information criterion is reported for each method, and as previously, we do not report the Tp_1 results. Fig. 6 shows the Monte Carlo simulation results for the estimation errors in the VAR coefficients.

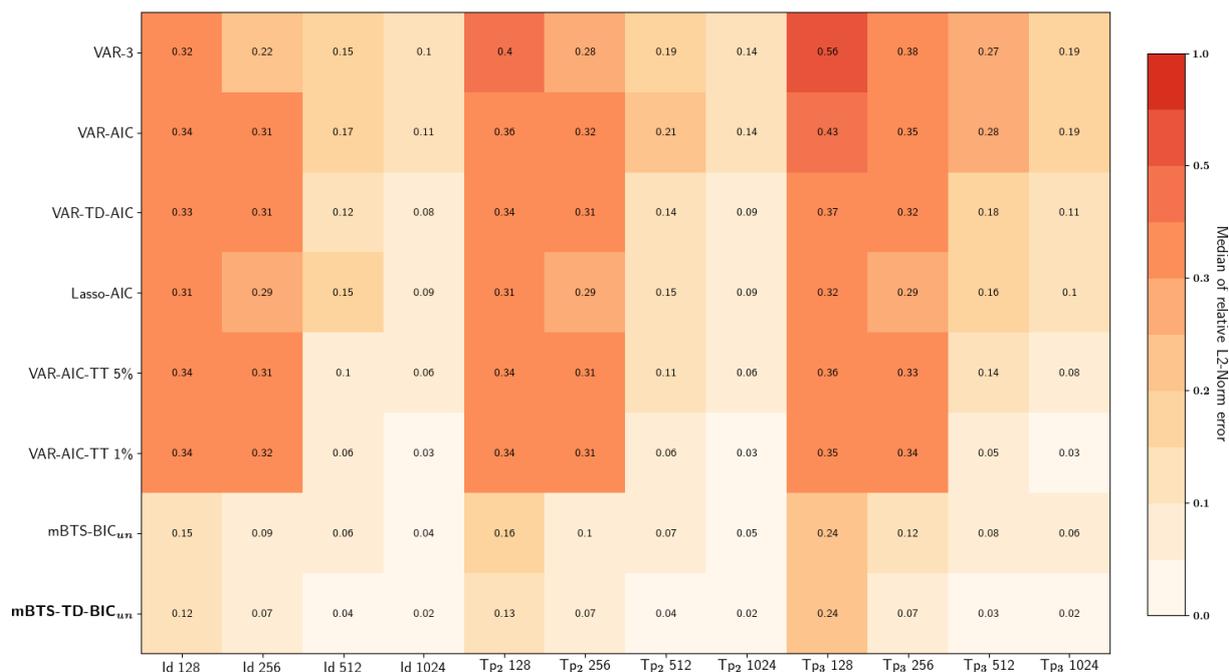


Fig 6: Median over 1000 simulations of the relative L_2 -norm error on the coefficients estimated using subset selection methods and standard VAR estimations. An increase in the median implies a deterioration in the estimate of the coefficients.

For estimation errors on the VAR coefficients, all subset selection methods provide better results than VAR-AIC for all the settings proposed, and they also behave better than VAR-3 for almost all settings. However, the two methods using mBTS, and in particular the combination of mBTS and TD, clearly stand out from the others. There is a considerable gap between these methods and the others. The errors are at least divided by two for small sample sizes and can be divided by five for the largest ones. Moreover, by adding the TD strategy to a first suitable VAR estimation that respects the parsimonious structure, such as mBTS, the improvement can be significant, especially when residuals are correlated. Nevertheless, TD may also confirm its drawback of being very sensitive to the first estimation (VAR-AIC), because it does not provide better results than VAR-AIC, even if it is more stable in the presence of correlated residuals. Lasso-AIC also provides good results compared to VAR-AIC or VAR-3, but its errors are at least twice as great as with mBTS methods. The TT procedures (especially TT 1%) are better than TD and Lasso for $T \geq 512$, but are quite similar for the two smaller sample sizes with higher errors than the mBTS methods. mBTS-TD can ensure highly stable errors across the different covariance matrices, in particular for sample sizes larger than 256, where the errors are the same. For example, with $T = 256$ the errors are equal to 0.07 whatever the covariance matrix. Finally, to conclude regarding these first results, subset selection methods are better suited to modelling parsimonious structures for the estimation of coefficients. When adding (mBTS) or deleting (TD) parameters as in mBTS-TD, the information criterion with the highest penalty factor, viz., BIC_{un} , provides better results than the less restricted ones.

Figs. 7 and 8 show GPDC errors in causal (Fig. 7) and non-causal terms (Fig. 8) of (S), confirming previous results for the VAR coefficients.

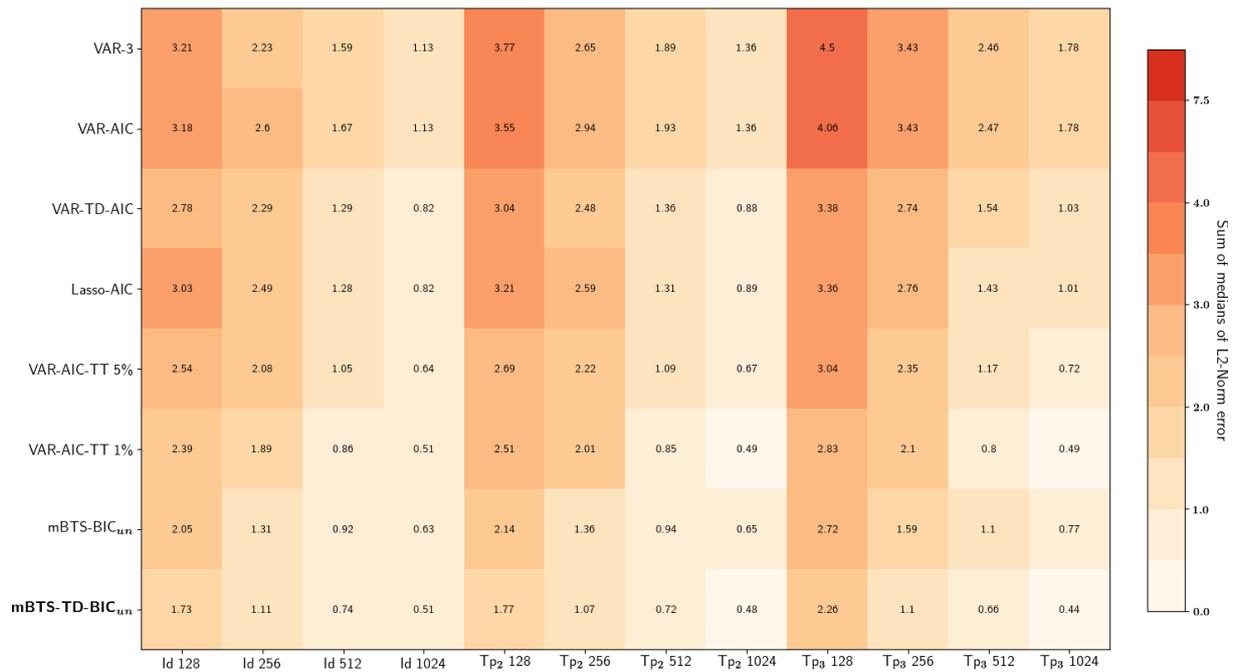


Fig 7: Sum of medians over 1000 simulations of the L_2 -norm error in the causal GPDC (5), estimated using the subset selection methods and standard VAR estimations. An increase in the median implies a deterioration in the estimate of the causal GPDC.

In Fig. 7, the results for the GPDC causal terms are similar to the errors in the coefficients. This confirms the idea that subset selection methods are better suited than standard VAR. mBTS-TD performs better than the others for all settings, and also provides stable errors.

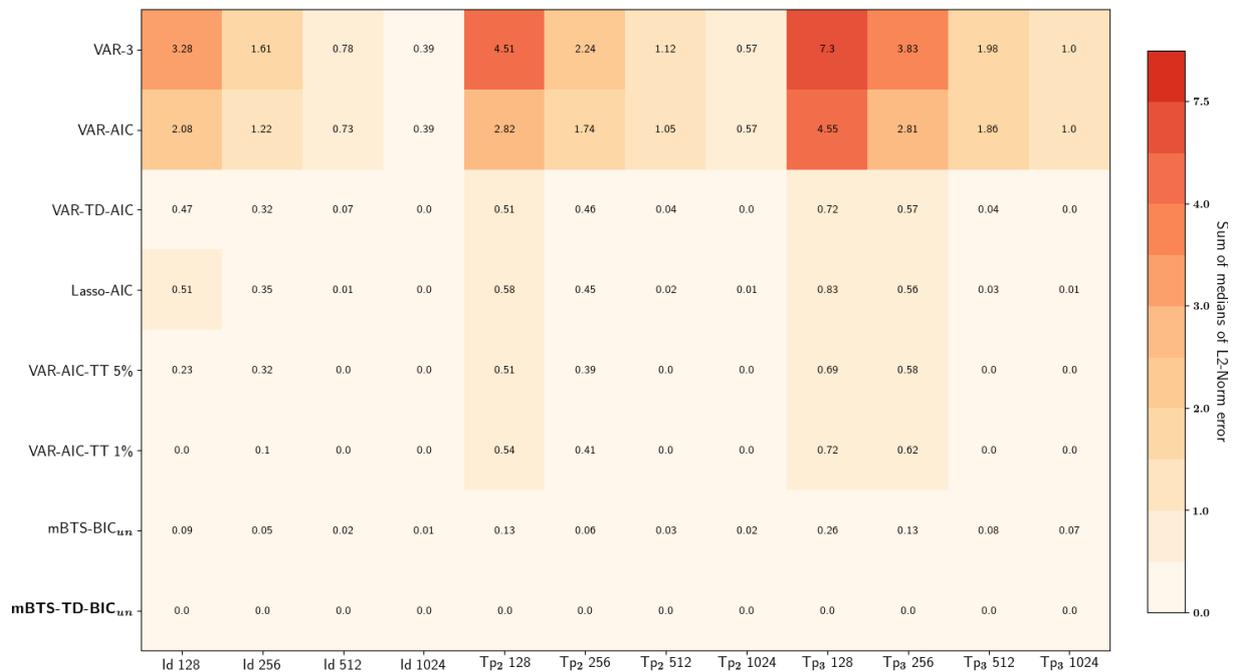


Fig 8: Sum of medians over 1000 simulations of the L_2 -norm error in the non-causal GPDC (5), estimated using the subset selection methods and standard VAR estimations. An increase in the median implies a deterioration in the estimate of the non-causal GPDC.

For the errors in the non-causal terms shown in Fig. 8, the results are clear. Each subset selection method perfectly plays out its role in modelling only the most significant coefficients (causal terms). The six subset selection methods greatly reduce errors compared to VAR-AIC or VAR-3, and can provide a sum of medians of the L_2 -norm errors close to zero for all non-causal terms. However, even in this case, the mBTS-TD method provides the most interesting results. This is the only method that presents a sum of medians close to zero for all settings, and it significantly improves on the mBTS method for sample sizes $T \leq 256$ and correlated residuals.

4.3 Robustness checks: error distributions and causal structure identification

Having used our preliminary analysis in section 4.2 to identify the three methods TT 1%, mBTS and mBTS-TD that seem most suitable for computing the GPDC, we extend here the comparison. In contrast to in section 4.2, where medians of the L_2 -norm error were used to compare the methods, in this section, we first evaluate the two methods through the L_2 -norm error distributions for the causal and non-causal GPDC, then focus on identification of the true causal structure of (S) using the F-Measure (FM) and Hamming Distance (HD) as in [31].

GPDC error distributions

In Table 1, we report the average value and standard deviation of the L_2 -norm error distributions for the causal GPDC and in Fig. 9 we provide an example of the L_2 -norm error distribution with $T = 256$ and $\Sigma_\epsilon = \text{Tp}_2$. Table 2 and Fig. 10 exhibit the same results for the non-causal terms.

	VAR-AIC	VAR-AIC TT 1%	mBTS BIC_{un} 3	mBTS-TD BIC_{un} 3	mBTS BIC_{un} 6	mBTS-TD BIC_{un} 6	mBTS BIC_{un} 9	mBTS-TD BIC_{un} 9
Id 128	0.687 (0.350)	0.544 (0.356)	0.449 (0.303)	0.389 (0.293)	0.480 (0.322)	0.417 (0.312)	0.492 (0.326)	0.429⁺ (0.318)
Id 256	0.546 (0.294)	0.406 (0.294)	0.302 (0.206)	0.251 (0.187)	0.317 (0.217)	0.266 (0.199)	0.327 (0.223)	0.274⁺ (0.206)
Id 512	0.378 (0.217)	0.239 (0.216)	0.203 (0.133)	0.167 (0.123)	0.212 (0.140)	0.175 (0.131)	0.218 (0.144)	0.180⁺ (0.135)
Id 1024	0.238 (0.122)	0.123 (0.095)	0.140 (0.094)	0.113 (0.085)	0.145 (0.098)	0.117 (0.089)	0.147 (0.099)	0.119⁺ (0.091)
Tp_2 128	0.764 (0.404)	0.575 (0.403)	0.475 (0.348)	0.407 (0.347)	0.505 (0.364)	0.439 (0.368)	0.524 (0.371)	0.457⁺ (0.375)
Tp_2 256	0.614 (0.334)	0.423 (0.319)	0.309 (0.216)	0.247 (0.196)	0.327 (0.229)	0.262 (0.208)	0.336 (0.233)	0.271⁺ (0.214)
Tp_2 512	0.447 (0.253)	0.242 (0.231)	0.212 (0.151)	0.168 (0.132)	0.220 (0.156)	0.175 (0.139)	0.226 (0.160)	0.181⁺ (0.143)
Tp_2 1024	0.291 (0.152)	0.119 (0.097)	0.143 (0.099)	0.108 (0.084)	0.147 (0.101)	0.112 (0.088)	0.149 (0.103)	0.114⁺ (0.089)
Tp_3 128	0.882 (0.462)	0.674 (0.480)	0.631 (0.464)	0.565 (0.486)	0.652 (0.466)	0.591 (0.490)	0.665 (0.466)	0.604⁺ (0.491)
Tp_3 256	0.721 (0.390)	0.451 (0.360)	0.370 (0.288)	0.276 (0.273)	0.386 (0.300)	0.291 (0.286)	0.394 (0.303)	0.300⁺ (0.291)
Tp_3 512	0.561 (0.320)	0.239 (0.245)	0.245 (0.185)	0.161 (0.142)	0.254 (0.191)	0.168 (0.149)	0.259 (0.192)	0.173⁺ (0.153)
Tp_3 1024	0.393 (0.223)	0.127 (0.121)	0.178 (0.133)	0.108 (0.092)	0.182 (0.137)	0.111 (0.097)	0.184 (0.138)	0.114⁺ (0.100)

Table 1: Causal GPDC: Average value and standard deviation in parentheses of the L_2 -norm error distribution (1000 simulations) for the causal GPDC, estimated using VAR-AIC, VAR-AIC-TT 1%, mBTS- BIC_{un} , and mBTS-TD- BIC_{un} with $p_{max} = 3, 6, 9$. The lower average error is highlighted for each setting and p_{max} . The superscript symbol $+$ indicates the lowest average error among mBTS-TD- BIC_{un} 9, mBTS- BIC_{un} 3, VAR-AIC-TT 1% to underline the efficiency of the mBTS-TD approach even for a large p_{max} .

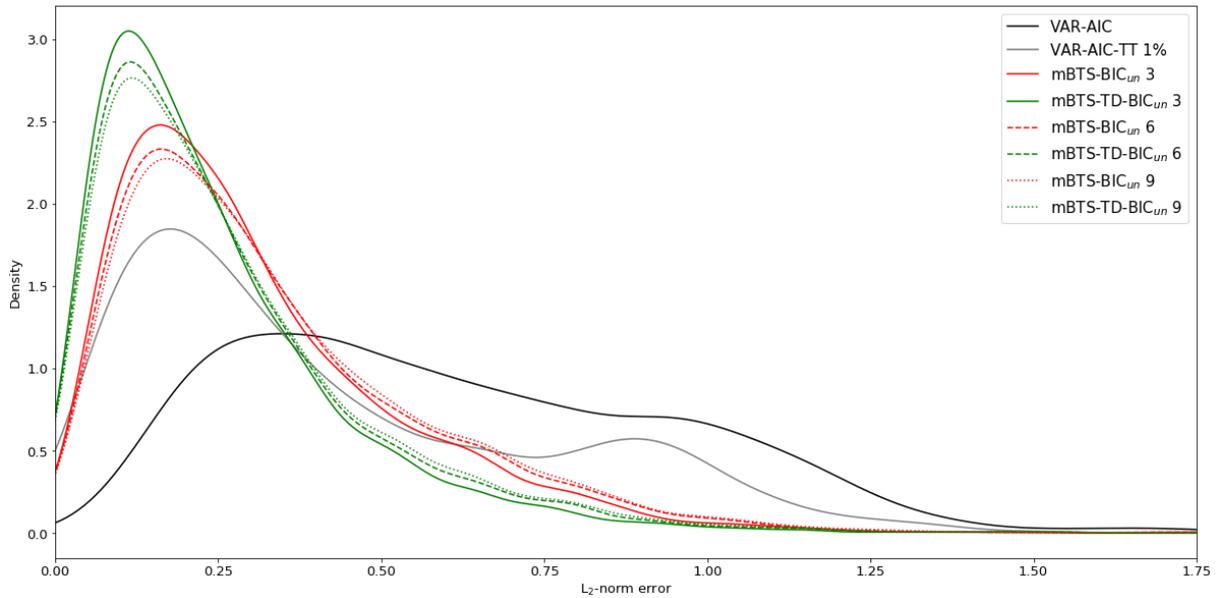


Fig 9: Causal GPDC. L_2 -norm error distribution (1000 simulations) with $T = 256$ and $\Sigma_\epsilon = Tp_2$ for the causal GPDC, estimated using VAR-AIC, VAR-AIC-TT 1%, mBTS- BIC_{un} , and mBTS-TD- BIC_{un} with $p_{max} = 3, 6, 9$.

	VAR-AIC	VAR-AIC TT 1%	mBTS BIC_{un} 3	mBTS-TD BIC_{un} 3	mBTS BIC_{un} 6	mBTS-TD BIC_{un} 6	mBTS BIC_{un} 9	mBTS-TD BIC_{un} 9
Id 128	0.195 (0.202)	0.055 (0.208)	0.035 ⁺ (0.124)	0.020 (0.115)	0.048 (0.150)	0.031 (0.137)	0.058 (0.167)	0.040 (0.156)
Id 256	0.112 (0.115)	0.031 (0.113)	0.014 (0.052)	0.005 (0.041)	0.019 (0.063)	0.009 (0.052)	0.022 (0.069)	0.012⁺ (0.059)
Id 512	0.065 (0.063)	0.011 (0.054)	0.005 (0.022)	0.001 (0.015)	0.007 (0.026)	0.003 (0.020)	0.008 (0.029)	0.004⁺ (0.023)
Id 1024	0.033 (0.028)	0.002 (0.013)	0.002 (0.011)	0 (0.007)	0.003 (0.012)	0.001 (0.008)	0.003 (0.013)	0.001⁺ (0.010)
Tp ₂ 128	0.266 (0.283)	0.071 (0.249)	0.042 ⁺ (0.156)	0.026 (0.145)	0.055 (0.177)	0.037 (0.165)	0.065 (0.190)	0.045 (0.178)
Tp ₂ 256	0.156 (0.171)	0.041 (0.149)	0.016 (0.061)	0.006 (0.048)	0.020 (0.069)	0.010 (0.056)	0.023 (0.075)	0.012⁺ (0.062)
Tp ₂ 512	0.096 (0.095)	0.015 (0.075)	0.008 (0.030)	0.002 (0.019)	0.009 (0.033)	0.003 (0.023)	0.011 (0.036)	0.004⁺ (0.026)
Tp ₂ 1024	0.047 (0.040)	0.002 (0.014)	0.003 (0.013)	0 (0.006)	0.004 (0.015)	0.001 (0.009)	0.005 (0.016)	0.001⁺ (0.010)
Tp ₃ 128	0.404 (0.417)	0.113 (0.367)	0.075 (0.253)	0.050 (0.246)	0.086 (0.265)	0.061 (0.260)	0.095 (0.274)	0.069⁺ (0.269)
Tp ₃ 256	0.256 (0.267)	0.059 (0.216)	0.029 (0.112)	0.012 (0.097)	0.033 (0.116)	0.015 (0.099)	0.036 (0.120)	0.018⁺ (0.104)
Tp ₃ 512	0.162 (0.158)	0.019 (0.104)	0.013 (0.046)	0.002 (0.027)	0.014 (0.049)	0.003 (0.030)	0.016 (0.051)	0.004⁺ (0.033)
Tp ₃ 1024	0.084 (0.077)	0.003 (0.028)	0.008 (0.028)	0 (0.012)	0.009 (0.029)	0.001 (0.013)	0.009 (0.030)	0.001⁺ (0.014)

Table 2: Non-causal GPDC: Average value and standard deviation in parentheses of the L_2 -norm error distribution (1000 simulations) for the non-causal GPDC, estimated using VAR-AIC, VAR-AIC-TT 1%, mBTS- BIC_{un} , and mBTS-TD- BIC_{un} with $p_{max} = 3, 6, 9$. The lower average error is highlighted for each setting and p_{max} . The superscript symbol ⁺ indicates the lowest average error among mBTS-TD- BIC_{un} 9, mBTS- BIC_{un} 3, VAR-AIC-TT 1% to underline the efficiency of the mBTS-TD approach even for a large p_{max} .

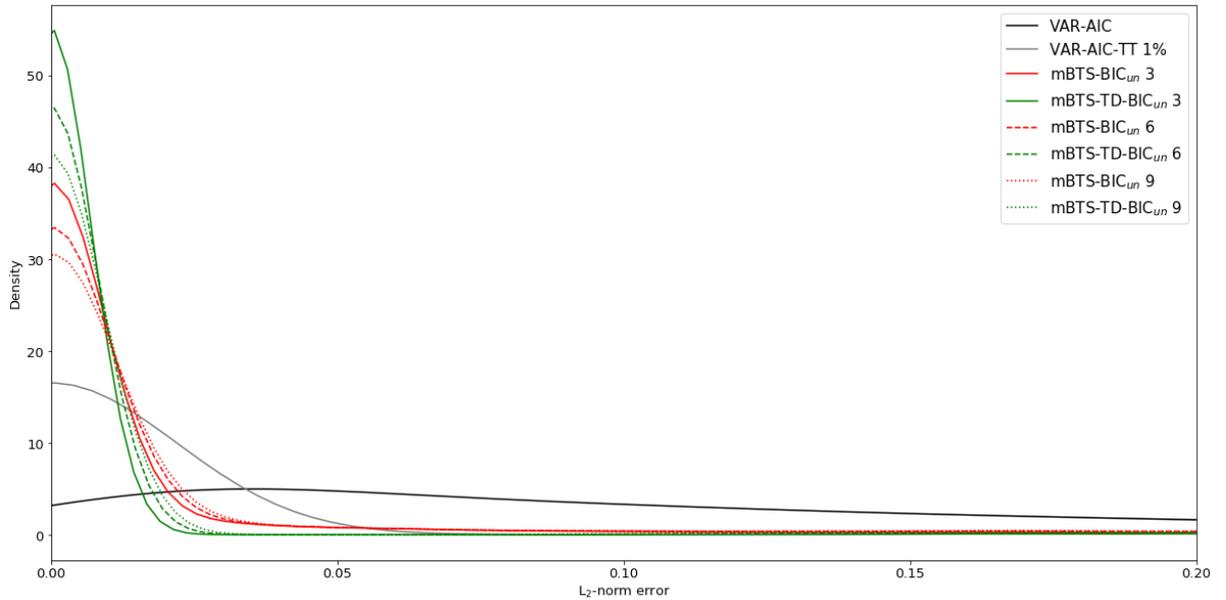


Fig 10: Non-causal GPDC. L_2 -norm error distribution (1000 simulations) with $T = 256$ and $\Sigma_\epsilon = Tp_2$ for the non-causal GPDC, estimated using VAR-AIC, VAR-AIC-TT 1%, mBTS- BIC_{un} , and mBTS-TD- BIC_{un} with $p_{max} = 3, 6, 9$.

The following conclusions can be drawn regarding both causal (Table 1 and Fig. 9) and non-causal (Table 2 and Fig. 10) terms:

- By taking into account only the same p_{max} for the two methods, denoted in bold in Tables 1 and 2, mBTS-TD clearly stands out from mBTS by providing lower average errors for each setting.
- Whatever p_{max} is selected, the mBTS-TD error distributions are more concentrated, with a lower fat tail.
- mBTS-TD-BIC_{un}9 always admits lower errors than mBTS-BIC_{un}3 and VAR-AIC-TT 1% for the causal GPDC, and only in two cases exhibits higher errors than mBTS-BIC_{un}3 for the non-causal terms denoted with a superscript symbol ⁺ in Tables 1 and 2.

Causal structure identification

To identify the true causal structure, we use the F-Measure (FM) and the Hamming Distance (HD) discussed in [31]. FM focuses on the identification of pairs of true causality, whereas HD focuses on the identification of all pairs. We consider the existence of causality between two time series $x_j(t)$ and $x_k(t)$ if $|\widehat{\omega}_{jk}|^2 > 0.01$ at least at one frequency f . FM and HD are defined as follows:

$$FM = \frac{2TP}{2TP + FN + FP}$$

$$HD = FN + FP$$

where TP are the True Positives (causality correctly identified), FN are the False Negatives (causality not identified), and FP are the False Positives (wrongly identified causality). FM ranges from 0 to 1. If FM = 1, then there is a perfect identification of the pairs of true causality, whereas if FM = 0, then no true causality is detected. HD ranges from 0 to $m(m - 1)$. If HD = 0, there is a perfect identification, whereas if HD = $m(m - 1)$, all pairs are misclassified.

Tables 3 and 4 report the average value of FM (Table 3) and HD (Table 4) for each setting.

	VAR-AIC	VAR-AIC TT 1%	mBTS $BIC_{un} 3$	mBTS-TD $BIC_{un} 3$	mBTS $BIC_{un} 6$	mBTS-TD $BIC_{un} 6$	mBTS $BIC_{un} 9$	mBTS-TD $BIC_{un} 9$
Id 128	0.500	0.881	0.873	0.941	0.840	0.909	0.818	0.885⁺
Id 256	0.568	0.884	0.925	0.968	0.898	0.943	0.881	0.928⁺
Id 512	0.663	0.928	0.970	0.985	0.958	0.977	0.950	0.971⁺
Id 1024	0.851	0.991	0.994	0.997	0.992	0.996	0.990	0.995⁺
Tp2 128	0.475	0.863	0.861	0.929	0.830	0.897	0.808	0.874⁺
Tp2 256	0.530	0.878	0.916	0.965	0.894	0.945	0.879	0.930⁺
Tp2 512	0.581	0.924	0.952	0.980	0.941	0.971	0.930	0.965⁺
Tp2 1024	0.751	0.990	0.988	0.997	0.984	0.995	0.981	0.994⁺
Tp3 128	0.446	0.819	0.794	0.878	0.771	0.853	0.754	0.833⁺
Tp3 256	0.479	0.873	0.875	0.954	0.857	0.936	0.842	0.921⁺
Tp3 512	0.500	0.932	0.932	0.986	0.921	0.978	0.912	0.971⁺
Tp3 1024	0.596	0.977	0.958	0.995	0.953	0.993	0.950	0.992⁺

Table 3: Average value of the F-measure (FM) over 1000 simulations of the GPDC, estimated using VAR-AIC, VAR-AIC-TT 1%, mBTS- BIC_{un} , and mBTS-TD- BIC_{un} with $p_{max} = 3, 6, 9$. FM ranges from 0 to 1. If FM = 1 there is perfect identification of the pairs of true causality, whereas if FM = 0 no true causality is detected. The lower average value is highlighted for each setting and p_{max} . The superscript symbol ⁺ indicates the lowest average error among mBTS-TD- $BIC_{un} 9$, mBTS- $BIC_{un} 3$, VAR-AIC-TT 1% to underline the efficiency of the mBTS-TD approach even for a large p_{max} .

	VAR-AIC	VAR-AIC TT 1%	mBTS $BIC_{un} 3$	mBTS-TD $BIC_{un} 3$	mBTS $BIC_{un} 6$	mBTS-TD $BIC_{un} 6$	mBTS $BIC_{un} 9$	mBTS-TD $BIC_{un} 9$
Id 128	10.014	1.325	1.451	0.626	1.893	0.997	2.215	1.291⁺
Id 256	7.596	1.311	0.808	0.333	1.134	0.605	1.351	0.779⁺
Id 512	5.076	0.774	0.306	0.150	0.439	0.237	0.529	0.302⁺
Id 1024	1.750	0.088	0.063	0.028	0.085	0.037	0.099	0.046⁺
Tp2 128	10.995	1.540	1.586	0.756	2.011	1.128	2.340	1.412⁺
Tp2 256	8.877	1.394	0.920	0.363	1.186	0.582	1.381	0.758⁺
Tp2 512	7.226	0.820	0.499	0.202	0.626	0.294	0.751	0.368⁺
Tp2 1024	3.320	0.106	0.118	0.033	0.165	0.050	0.194	0.063⁺
Tp3 128	12.309	2.070	2.392	1.279	2.740	1.579	3.018	1.846⁺
Tp3 256	10.877	1.444	1.406	0.472	1.643	0.678	1.852	0.848⁺
Tp3 512	9.989	0.727	0.729	0.146	0.861	0.229	0.968	0.300⁺
Tp3 1024	6.771	0.240	0.434	0.050	0.495	0.070	0.527	0.077⁺

Table 4: Average value of Hamming Distance (HD) over 1000 simulations of the GPDC, estimated using VAR-AIC, VAR-AIC-TT 1%, mBTS- BIC_{un} , and mBTS-TD- BIC_{un} with $p_{max} = 3, 6, 9$. HD ranges from 0 to $m(m - 1)$, where $m = 5$. If HD = 0 there is perfect identification, whereas if HD = 20 all pairs are misclassified. The lower average value is highlighted for each setting and p_{max} . The superscript symbol ⁺ indicates the lowest average error among mBTS-TD- $BIC_{un} 9$, mBTS- $BIC_{un} 3$, VAR-AIC-TT 1% to underline the efficiency of the mBTS-TD approach even for a large p_{max} .

For the two measures FM (Table 3) and HD (Table 4), mBTS-TD provides on average a better identification of the true causal structure than mBTS for each setting by taking into account only the same p_{max} highlighted in boldface in Tables 3 and 4. As previously for the error distributions, mBTS-TD- $BIC_{un} 9$ outperforms mBTS- $BIC_{un} 3$ and VAR-AIC-TT 1%, denoted with a superscript symbol ⁺ in Tables 3 and 4. Note that VAR-AIC provides the worst results, whatever measure is taken into account.

We end this simulation study by comparing the computational efficiency of mBTS and mBTS-TD on (S) across the different sample sizes. The computation times of the two methods for one realization ($T = 1024$ and Tp_3) are quite similar, with 0.166 seconds for mBTS and 0.174 seconds for mBTS-TD. The computations are carried out using Python 3.7 with 2.70GHz CPU (Intel Xeon E-2176M) and 32Gb RAM.

To conclude regarding the accuracy of the GPDC results, subset selection methods are well suited when the underlying model admits a parsimonious structure. Each subset method improves the GPDC accuracy, and when we combine mBTS and TD methods, the cascading errors on GPDC are drastically reduced for both causal and non-causal terms. We can also add that, with $p_{max} = 9$, for mBTS-TD the errors on the GPDC are lower than for mBTS starting with the true lag $p_{max} = 3$. mBTS-TD therefore reduces dependence on the choice of p_{max} and produces a more parsimonious model when p_{max} is large. Fig. 11 compares examples of the GPDC estimated using mBTS-TD, VAR-3, and VAR-AIC with the theoretical GPDC.

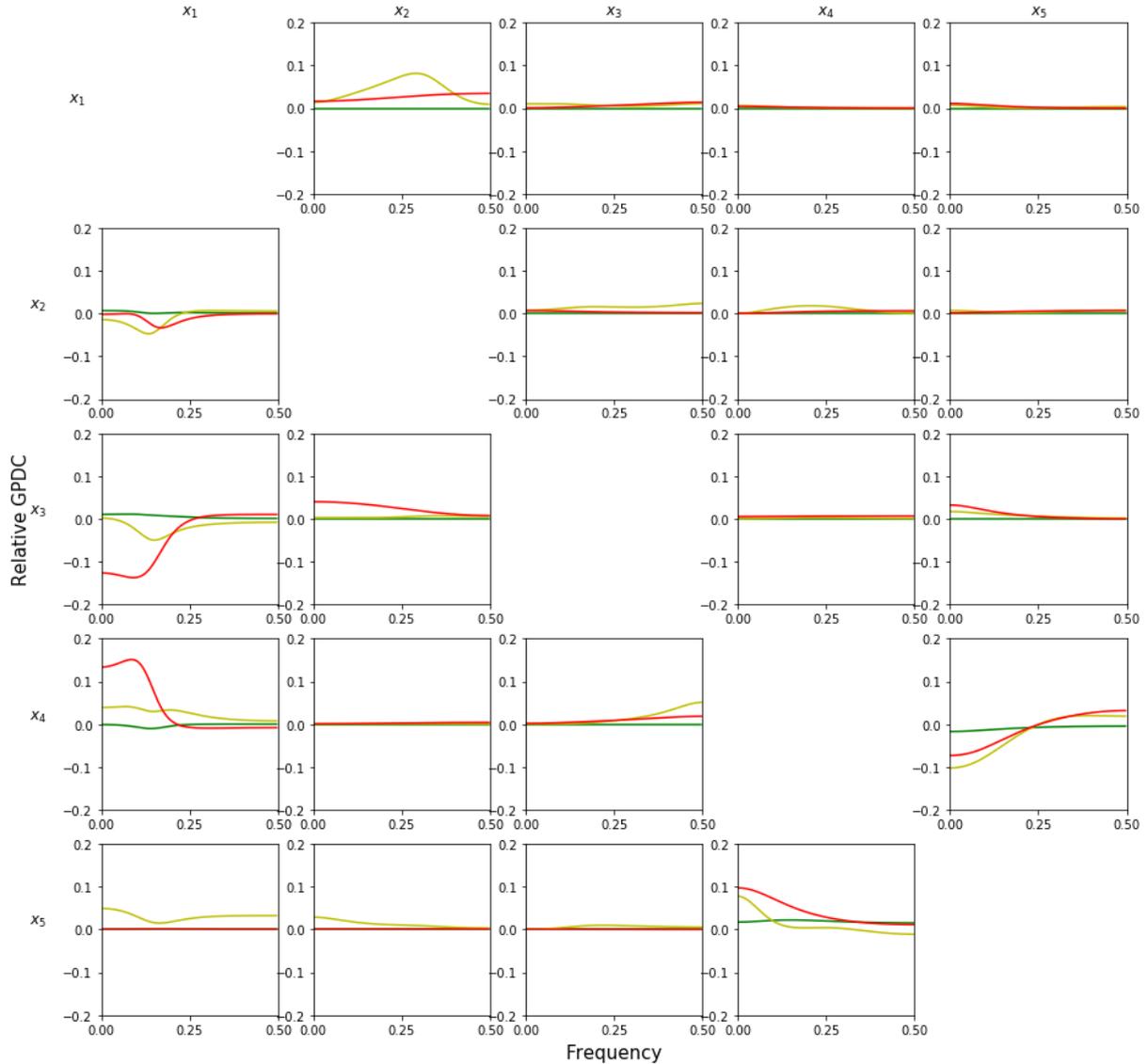


Fig 11: Relative GPDC compared to the theoretical GPDC of (S) for $T = 256$ and $\Sigma_\epsilon = \text{Tp}_2$: VAR-2 in red, VAR-3 in yellow, and mBTS-TD- BIC_{un} in green. Relative GPDC: $|\hat{\omega}_{jk}|^2 - |\omega_{jk}|^2$ for each pair $j \neq k$, where ω_{jk} is the vector containing each value of $\omega_{jk}(f)$ for all discrete frequencies f .

As found with Monte Carlo simulations, the differences in the causal part between the mBTS-TD method and the theoretical GPDC are very small. Moreover, in this example, none of the zero-coefficients are estimated by mBTS-TD, which makes it possible to have a GPDC equal to zero, like the theoretical GPDC for non-causal terms.

5 Financial application

In the literature, several classical approaches exist to model the links between the assets of a financial universe. The correlation matrix is often used [25, 32, 33, 34] to build weighted or binary networks. Unfortunately, this methodology suffers from two major drawbacks. First, these networks are undirected, only highlighting the existence of the relationship between assets not their directions. Second, the network dimension must also be reduced (using methods such as the Minimum Spanning Tree [25] or Planar Maximally Filtered Graphs [26]) otherwise the network is complete and difficult to use in practice for portfolio allocation. In a symmetric way, methods based on Granger non-causality tests in VAR models as in [35, 36] allow to retrieve a directed but unweighted network, remaining very sensitive to the underlying VAR processes. In [37], an alternative directed causal network is built, beyond VAR modeling, but focuses only on very short-dynamics.

In this section, we make use of the GPDC measure, estimated with our mBTS-TD method, to modeling financial markets dependency structures. This approach provides not only a precise network topology (taking into account both the direction and the strength of the relationship between assets via the GPDC) but also solve the dimensionality puzzle, via the mBTS-TD estimation process that intrinsically produces parsimonious causal structures. In a second step, we study on real data the empirical performances of financial portfolios obtained excluding the most systemic nodes of the incomplete GPDC financial network.

5.1 Building a GPDC financial network to identify systemic assets

A network $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ is a set of objects with \mathbf{V} the set of nodes and \mathbf{E} the set of edges between nodes. The edge (j, k) connects a pair of nodes j and k . The mathematical representation of a directed weighted network is given by the $m \times m$ adjacency matrix $\mathbf{Z} = (z_{jk})$, $z_{jk} \in \mathbb{R}^+$ if $(j, k) \in \mathbf{E}$ and 0 otherwise. In the sequel, let $\mathbf{Z}^{(1)} = (z_{jk}^{(1)})$ be the adjacency matrix built using the GPDCs in which are plugged the VAR coefficients, estimated using our mBTS-TD procedure, and $\mathbf{Z}^{(2,\gamma)} = (z_{jk}^{(2,\gamma)})$ the adjacency matrix built using the GPDCs based on VAR-AIC models. In this latter case, note that $\mathbf{Z}^{(2,\gamma)}$ depends on a threshold parameter γ , defined hereafter. Since our mBTS-TD procedure pre-filters the VAR by removing unnecessary coefficients, $z_{jk}^{(1)}$ can be directly defined as follows:

$$z_{jk}^{(1)} = \begin{cases} \max |\boldsymbol{\omega}_{jk}|^2, & \text{if } j \neq k \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where $\boldsymbol{\omega}_{jk}$ is the vector containing each value of $\omega_{jk}(f)$ for all discrete frequencies f ($f \in [0, \frac{1}{2}]$) as defined in (5). We use the maximal value in the $\boldsymbol{\omega}_{jk}$ vector in order to take into account the most relevant information between the two assets, i.e. based on short (high-frequency) and long-term (low frequency) relationships.

When a classical VAR-AIC is used to compute the GPDCs, all VAR coefficients are involved, returning, in general, non-null GPDCs. The resulting complete weighted network is useless in practice. We thus apply a filter to each component of the associated ω_{jk} vector and compute the vector ω_{jk}^γ whose coordinates are given by:

$$\omega_{jk}^\gamma(f) = \begin{cases} \omega_{jk}(f), & \text{if } |\omega_{jk}(f)| \geq \gamma \\ 0, & \text{otherwise} \end{cases}$$

with $\gamma \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$.

Thus, an element of the adjacency matrix $\mathbf{Z}^{(2,\gamma)}$, is defined as:

$$z_{jk}^{(2,\gamma)} = \begin{cases} \max |\omega_{jk}^\gamma|^2, & \text{if } j \neq k \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Once built, the previous incomplete financial networks can help us to improve asset allocation strategies using the classical tools of network theory such as centrality measures or clustering coefficients as in [32, 33, 38, 39, 40, 41]. In order to show the potential of using both the mBTS-TD method and the GPDC to build a financial network, we propose to identify the most systemic assets with the local directed weighted clustering coefficient. Indeed, the local directed weighted clustering coefficient allows to identify the most embedded assets in the network and thus the most systemic ones. This tool introduced by Clemente et al. in [27] measures how a node is embedded into the network by quantifying its number of triangles out of all its possible triangles. Furthermore, it takes into account the strength of a node in the normalization factor (see also [42]). Starting from a directed weighted network with adjacency matrix \mathbf{Z} , we obtain the associated directed unweighted network with adjacency matrix \mathbf{Z}^u defining $z_{jk}^u = 1$ if $z_{jk} \neq 0$, and 0 otherwise. Thus, the local directed weighted clustering coefficient for the asset (node) j is defined as follows:

$$h_j = \frac{1}{2} \frac{[(\mathbf{Z} + \mathbf{Z}')(\mathbf{Z}^u + \mathbf{Z}^{u'})^2]_{jj}}{s_j(d_j - 1) - 2s_j^{\leftrightarrow}}$$

where $d_j = (\mathbf{Z}^{u'} + \mathbf{Z}^u)_j \mathbf{1}_m$ and $s_j = (\mathbf{Z}' + \mathbf{Z})_j \mathbf{1}_m$ are respectively the total degree (total number of edges) and the total strength (case of weighted graph) of the asset j . $s_j^{\leftrightarrow} = \frac{(\mathbf{Z}\mathbf{Z}^u + \mathbf{Z}^u\mathbf{Z})_{jj}}{2}$ is the strength of bilateral edges between j and k . Note that h_j belongs to $[0, 1]$, a high value indicating that the asset j is heavily embedded in the network, and captures in particular in and out diffusion processes, and therefore spillover and feedback effects.

5.2 Building a diversified Equally Weighted portfolio

For a given investment universe, the above methodology allow us to identify the most systemic nodes of the GPDC network as the one associated with the greatest local directed weighted clustering coefficients computed from adjacency matrices $\mathbf{Z}^{(1)}$ or $\mathbf{Z}^{(2,\gamma)}$. To build and compare financial portfolios we only base our financial strategy on m non-systemic assets and we basically allocate all of them with the same weight $\frac{1}{m}$ resulting in an Equally Weighted (EW) portfolio. The interest of such an approach in our framework is to focus solely on the improvement resulting from the asset selection process. It does not require any additional estimation procedure (covariance matrix) nor complex optimization issues. What is more, the authors in [28] have shown that this method can even provide higher performances than more advanced ones.

5.3 Dataset description and Empirical performances

We consider national financial markets, each market (node) being represented by the MSCI ACWI (All Country World Index)². We first apply two filters. First we remove the less liquid ones (Argentina, Czech Republic, Egypt, Greece, Hungary, Pakistan), and then those with no quotes since 2001 (Qatar, Saudi Arabia, United Arab Emirates). This universe of 40 assets (see Table 7 in Appendix A.3) allows us to take both in account the differences in time delay between areas (feedback effects) as well as local discrepancy (e.g. macroeconomic differences). We use asset returns computed on a daily basis from January 2001, the 18th to October 2019, the 25th to build, every four weeks, a temporal network, using a rolling window of $T = 256$ working days, with a rebalancing period of four weeks. Since financial assets returns exhibit heteroskedasticity, we normalize each time series using a Generalized Auto-Regressive Conditional Heteroskedastic (GARCH) [43] filter (see Appendix A.4) to estimate the corresponding VAR process used to compute the GPDC measure and to identify the non-systemic assets.

The asset exclusion procedure using mBTS-TD for the VAR estimation and the GPDC measure (“mBTS-TD GPDC”) is compared with those obtained using a classical VAR estimation (“VAR-AIC GPDC”) with several thresholds $\gamma \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$ and on the whole universe (“EW”), i.e. without any asset selection. In order to assess the potential of our methodology, we report several portfolio statistics computed over the whole period: the annualized return, the annualized volatility, the ratio between the annualized return and the annualized volatility and the maximum drawdown (largest decline in portfolio value). The portfolio generates better performances if it provides an higher return/volatility ratio and a lower maximum drawdown. The (EW) portfolios’ performances are computed in USD currency, because if the asset returns are kept into local currency, hedging costs (selling the currency forward) have to be considered.

In Table 5, we provide the portfolios’ results when ten assets are excluded representing 25% of the initial universe and we only report the best threshold γ for the classical VAR estimation. For this case, “mBTS-TD GPDC” shows a significant improvement with respect to the other methods. It provides significant higher annualized return, similar annualized volatility (higher return/volatility ratio) and also a similar drawdown compared to (EW) without exclusion.

²Data are available upon request.

EW Portfolios 10 excluded assets	Annualized Return	Annualized Volatility	Ratio Return/Volatility	Max Drawdown
mBTS-TD GPDC	10.46%	16.63%	0.63	62.03%
VAR-AIC GPDC 0.03	9.51%	16.84%	0.56	62.11%
<i>EW</i>	9.35%	16.76%	0.56	61.90%
Non-selected assets (VAR-AIC GPDC 0.03)	8.70%	17.64%	0.49	61.48%
Non-selected assets (mBTS-TD GPDC)	5.89%	18.28%	0.32	61.63%

Table 5: Performance indicators for EW portfolios with 10 excluded assets from January 2002 to October 2019. The results are ranked in descending order according to the ratio (Return / Volatility)

Given these results, we can consider that our methodology “mBTS-TD GPDC” succeeds in identifying the less performing/riskiest assets. To reinforce this aspect, Table 6 provides the first four order moments of systemic assets return distribution for our proposed methodology and the “VAR-AIC GPDC 0.03”.

10 Non-Selected Assets return distribution	mBTS-TD GPDC	VAR-AIC GPDC 0.03
Mean	0.0003	0.0004
Standard Deviation	0.0156	0.0159
Skewness	-0.0180	-0.0114
Kurtosis	10.3154	10.8294

Table 6: Moments of out-of-sample asset return distribution for ten non-selected assets for “mBTS-TD GPDC” and “VAR-AIC GPDC 0.03” from January 2002 to October 2019.

We observe that the assets return distribution in the non-selected universe obtained using “mBTS-TD GPDC” provides better figures, in particular for the mean and the skewness. Indeed, the assets non-selected by the “mBTS-TD GPDC” have a lower average return and more negative skewness than in the “VAR-AIC GPDC 0.03” case, which confirms that “mBTS-TD GPDC” better identifies the less performing/riskiest assets than the standard VAR estimation. Regarding portfolios’ performances, the “mBTS-TD GPDC” exclusion process takes full advantage of the related precise network topology combined with an intrinsic parsimonious causal structure. This paves the way for interesting results in the case of more complex assets allocation processes that will be the objective of a forthcoming study.

6 Conclusion

Retrieving complex interactions in multivariate systems admitting a VAR representation is of key importance in a number of fields. To this end, coherence measures have been introduced to quantify causal strength between variables. Nevertheless, we prove in this paper, through careful Monte Carlo simulations, that applying a naive approach first estimating a VAR model using LS, and then computing coherence measures, is highly inefficient especially when the underlying data generating processes are parsimonious. To overcome this problem, we apply classical subset selection methods, and show that they do improve coherence measures but not sufficiently. We therefore introduce a new subset selection method, namely the mBTS-TD one, and, still using Monte Carlo simulations, prove that it clearly outperforms its natural competitors, and allows us to dramatically reduce to so-called cascading errors in both the causal and non-causal structure of the system. Last, we have implemented our procedure in the financial domain making use of the GPDC measure estimated with the mBTS-TD strategy to model financial markets dependency structures. This approach provides us not only with a precise network topology (taking into account both the direction and the strength of the relationship between assets) but also solves the network dimension puzzle producing a parsimonious causal structure. We take advantage of this financial network identifying, via the local directed weighted clustering coefficient, the most systemic assets to exclude them, with profit, from our investment universe.

A Appendix

A.1 Testing procedure (TT)

Alternative procedures to information criteria are based on hypothesis testing. Indeed, the significant coefficients can be chosen with the individual t -ratio, i.e. by excluding all the smallest absolute values of t -ratios until all absolute t -ratios are greater to a threshold η . This procedure called Testing Procedure (TT) [22, 23] is a similar approach to the TD strategy, but the coefficients are deleted using the individual t -ratios. This strategy is much faster than the TD strategy because it immediately identifies which variable is deleted in the next step, whereas in the TD strategy each coefficient has to be retested. Nonetheless, the threshold η has to be fixed *a priori*. In general, the value is fixed to 2 which corresponds roughly to the 5% significant level or alternatively using the quantile of the t -distribution $\mathcal{T}_\nu^{\alpha/2}$ for a given probability α and ν degrees of freedom.

For the j -th equation obtain from the full VAR model (1) estimated in a multivariate environment, the TT procedure is applied as follows:

1. Compute the t -ratios associated to $\hat{\mathbf{b}}_j$ the $(q \times 1)$ vector of the estimated coefficients for the j -th equation, where q is the number of lagged variables (for the first step $q = mp$). For the coefficient $(\hat{\mathbf{b}}_j)_n$, the t -ratio is defined as follows:

$$\varphi_{n,j} = \frac{(\hat{\mathbf{b}}_j)_n}{\left(\frac{\hat{\boldsymbol{\epsilon}}_j' \hat{\boldsymbol{\epsilon}}_j}{T - q - 1} (\mathbf{X} \mathbf{X}')_{nn}^{-1} \right)^{1/2}} \quad (10)$$

where $n \in \{1, \dots, q\}$, $\boldsymbol{\epsilon}_j$ is the $(T \times 1)$ vector of residuals $(\mathbf{y}_j - \hat{\mathbf{b}}_j' \mathbf{X})$ with $\mathbf{y}_j = (x_j(1), \dots, x_j(T))'$ the $(T \times 1)$ vector of observations, \mathbf{X} the $(q \times T)$ matrix defined in (4) and $(\mathbf{X} \mathbf{X}')_{nn}^{-1}$ the n -th element of the diagonal of $(\mathbf{X} \mathbf{X}')^{-1}$.

2. Delete the coefficient n with the lowest absolute t -ratio, if and only if $|\varphi_{n,j}| < \mathcal{T}_{T-q}^{\alpha/2}$.
3. Re-estimate the coefficients for the j -th equation by removing the n -th row from the \mathbf{X} matrix and recompute the t -ratios again from (10) with the new residuals obtained and decreasing the number of lagged variables q by one.
4. Repeat steps 2 and 3 until $|\varphi_{n,j}| \geq \mathcal{T}_{T-q}^{\alpha/2} \forall n \in \{1, \dots, q\}$.

A.2 Lasso

The Lasso method (Least Absolute Shrinkage and Selection Operator) was introduced by Tibshirani [24, 17], and is the Least Square (LS) method with the L_1 -norm constraint on the VAR coefficients. The Lasso estimator for the VAR(p) model in (1) is defined as follows:

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_2^2 + \lambda \|\mathbf{A}\|_1$$

where

- $\mathbf{Y} = (\mathbf{x}(1), \dots, \mathbf{x}(T))$ is a $(m \times T)$ matrix of observations,
- \mathbf{A} is the $(m \times mp)$ matrix defined in section 3,
- \mathbf{X} is the $(mp \times T)$ matrix defined in (4),
- $\lambda \in \mathbb{R}^+$ is the tuning parameter.

If $\lambda = 0$, the Lasso method coincides with the OLS estimate. If $\lambda > 0$, the least significant coefficients in $\hat{\mathbf{A}}$ are shrunk to zero. In typical cases, a cross-validation procedure may be used, as Tibshirani [24] suggested, to choose both the lag p and the tuning parameter λ . In this paper we only select λ by considering that the lag p is predetermined by the VAR(p) model order estimation, as outlined in the section 2

The Lasso method has the advantage of estimating coefficients and selecting variables simultaneously, but it also adds complexity in the estimation of the parameter λ .

A.3 Country indices dataset

Index	Currency
Australia	AUD
Austria	EUR
Belgium	EUR
Brazil	BRL
Canada	CAD
Chile	CLP
China	HKD
Colombia	COP
Denmark	DKK
Finland	EUR
France	EUR
Germany	EUR
Hong Kong	HKD
India	INR
Indonesia	IDR
Ireland	EUR
Israel	ILS
Italy	EUR
Japan	JPY
Korea	KRW
Malaysia	MYR
Mexico	MXN
Netherlands	EUR
New Zealand	NZD
Norway	NOK
Peru	PEN
Philippines	PHP
Poland	PLN
Portugal	EUR
Russia	RUB
Singapore	SGD
South Africa	ZAR
Spain	EUR
Sweden	SEK
Switzerland	CHF
Taiwan	TWD
Thailand	THB
Turkey	TRY
United Kingdom	GBP
United States	USD

Table 7: Country equity indices in the MSCI ACWI (All Country World Index)

A.4 GARCH model

Let $x(t)$ be a zero-mean stationary process admitting the following GARCH(1,1) representation:

$$x(t) = \sqrt{h(t)} \epsilon(t)$$
$$h(t) = \beta_0 + \beta_1 x^2(t-1) + \beta_2 h(t-1)$$

where $h(t)$ is the conditional variance, β_0 , β_1 and β_2 are the coefficients and $\epsilon(t)$ is the white noise with $\epsilon(t) \sim \mathcal{N}(0, 1)$. Moreover, the parameters are estimated by maximizing the conditional log-likelihood.

In order to remove the heteroskedasticity, $x(t)$ is standardized as follows:

$$\tilde{x}(t) = \frac{x(t)}{\sqrt{h(t)}}$$

References

- [1] C. A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.
- [2] R. S. Tsay. *Analysis of financial time series*. Wiley series in probability and statistics. Wiley-Interscience, 2nd edition, 2005.
- [3] P. Valdés-Sosa, J. M. Sanchez, A. Castellanos, M. Vega-Hernández, J. Bosch, L. Melie-Garcia, and E. Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360:969–981, 2005.
- [4] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.
- [5] M. J. Kamiński and K. J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, 65(3):203–210, 1991.
- [6] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 85(2):145–157, 2001.
- [7] K. Sameshima and L. A. Baccalá. Using partial directed coherence to describe neuronal ensemble interactions. *Journal of Neuroscience Methods*, 94(1):93–103, 1999.
- [8] L. A. Baccalá and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84(6):463–474, 2001.
- [9] L. A. Baccalá, K. Sameshima, and D. Y. Takahashi. Generalized partial directed coherence. In *2007 15th International Conference on Digital Signal Processing*, pages 163–166, 2007.
- [10] H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, 1973.

- [11] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [12] E. J. Bedrick and C.-L. Tsai. Model selection for multivariate regression in small samples. *Biometrics*, 50(1):226–231, 1994.
- [13] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
- [14] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer, 2005.
- [15] L. A. Baccalá, C. S. N. de Brito, D. Y. Takahashi, and K. Sameshima. Unified asymptotic theory for all partial directed coherence forms. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371:20120158, 2013.
- [16] K. Sameshima and Baccalá L. A. *Methods in brain connectivity inference through multivariate time series analysis*. CRC Press, 2014.
- [17] N.-J. Hsu, H.-L. Hung, and Y.-M. Chang. Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, 52(7):3645–3657, 2008.
- [18] I. Vlachos and D. Kugiumtzis. Backward-in-Time Selection of the Order of Dynamic Regression Prediction Model. *Journal of Forecasting*, 32(8):685–701, 2013.
- [19] F. Han, H. Lu, and H. Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150, 2015.
- [20] H. Qiu, S. Xu, F. Han, H. Liu, and B. Caffo. Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. *Proceedings of the International Conference on Machine Learning. International Conference on Machine Learning*, 37:1843–1851, 2015.
- [21] E. Siggiridou and D. Kugiumtzis. Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, 64:1759–1773, 2016.
- [22] R. Brüggemann and H. Lütkepohl. Lag selection in subset var models with an application to a u.s. monetary system. Econometric Society World Congress 2000 Contributed Papers 0821, Econometric Society, 2000.
- [23] R. Brüggemann. *Model reduction methods for vector autoregressive processes*. Springer, 01 2004.
- [24] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.
- [25] R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems*, 11:193–197, 1999.
- [26] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, 2005.

- [27] G.P. Clemente and R. Grassi. Directed clustering in weighted networks: A new perspective. *Chaos, Solitons & Fractals*, 107:26 – 38, 2018.
- [28] V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, May 2009.
- [29] Y. Saito and H. Harashima. Tracking of information within multichannel EEG record - causal analysis in EEG. In N. Yamaguchi and K. Fujisawa, editors, *Recent Advances in EEG and EMG Data Processing*, pages 133–146. Elsevier, New York, 1981.
- [30] S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.
- [31] E. Siggiridou, V. K. Kimiskidis, and D. Kugiumtzis. Dimension reduction of frequency-based direct granger causality measures on short time series. *Journal of Neuroscience Methods*, 289:64–74, 2017.
- [32] F. Pozzi, T. Di Matteo, and T. Aste. Spread of risk across financial markets: better to invest in the peripheries. *Scientific Reports*, 3, 2013.
- [33] G. Peralta and A. Zareei. A network approach to portfolio selection. *Journal of Empirical Finance*, 38:157 – 180, 2016.
- [34] Y. Li, X.-F. Jiang, Y. Tian, S.-P Li, and B. Zheng. Portfolio optimization based on network topology. *Physica A: Statistical Mechanics and its Applications*, 515:671 – 681, 2019.
- [35] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535 – 559, 2012.
- [36] A. Papan, C. Kyrtsov, D. Kugiumtzis, and C. Diks. Financial networks based on granger causality: A case study. *Physica A: Statistical Mechanics and its Applications*, 482:65 – 73, 2017.
- [37] H. Gatfaoui and P. de Peretti. Flickering in information spreading precedes critical transitions in financial markets. *Scientific Reports*, 9, 2019.
- [38] F. Ren, Y.-N. Lu, S.-P. Li, X.-F Jiang, L.-X Zhong, and T. Qiu. Dynamic portfolio strategy using clustering approach. *PLOS ONE*, 12(1):1–23, 01 2017.
- [39] X. Guo, Z. Xue, H. Zhang, and T. Tian. Development of stock correlation networks using mutual information and financial big data. *PLOS ONE*, 13(4):1–16, 04 2018.
- [40] G. Clemente, R. Grassi, and A. Hitaj. Asset allocation: new evidence through network approaches. *Annals of Operations Research*, pages 1–20, 2019.
- [41] G. P. Clemente, R. Grassi, and A. Hitaj. Smart network based portfolios. *arXiv preprint*, 2019.
- [42] G. Fagiolo. Clustering in complex directed networks. *Physical review. E*, 76:026107, 2007.

- [43] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.