Model selection and loss functions for structural time series: networks, spatial models, causality measures, and misspecified or redundant moment conditions

by

Daniela Scidá

B.A., Universidad Nacional de Tucumán, Tucumán, Argentina, 2006
M.A., Centro de Estudios Monetarios y Financieros, Madrid, Spain, 2010
M.A., Brown University, Providence, Rhode Island, USA, 2011

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
Department of Economics at Brown University

Providence, Rhode Island
May 2016

This dissertation by Daniela Scidá is accepted in its present form
by the Department of Economics as satisfying the
dissertation requirements for the degree of Doctor of Philosophy.

Date _____ _____

Eric Renault, Advisor

Recommended to the Graduate Council

Date _____ _____

Adam McCloskey, Reader

Date _____ _____

Mardi Dungey, Reader

Approved by the Graduate Council

Date _____ _____

Peter Weber, Dean of the Graduate School

Vita

Daniela Scidá was born on February 24th, 1982, in Tucumán, Argentina. She studied Economics at Universidad Nacional de Tucumán (UNT) in Tucumán, Argentina. She graduated summa cum laude from UNT in 2006 and obtained her Bachelor's degree. In her final year at UNT she received a gold medal for the highest GPA and a recognition from the Argentine Federation of University Women for the best B.A. in Economic Sciences. After earning her Bachelor's degree, she worked for a year in a joint project between UNICEF and the Government of Tucumán, and for two years for the Finance Ministry of Tucumán as an Economic Consultant. In 2008, she decided to continue her studies and moved to Madrid, Spain, to pursue a Master in Economics and Finance at Centro de Estudios Monetarios y Financieros (CEMFI). During her time at CEMFI, she received a full fellowship from Fundación Carolina and obtained her Master's degree in June 2010. Later in 2010, she moved to Providence, Rhode Island, to commence her doctoral studies in Economics at Brown University. During her time at Brown University, she received a Merit Dissertation Fellowship and was awarded a Graduate Teaching Award. She also received a Core Curriculum Development Grant from Continuing Education at Brown University, and the Info-Metrics Institute Graduate Student Summer Fellowship from the American University in DC. She obtained her second Master's degree in 2011 and her Ph.D in 2016.

Acknowledgments

I owe many thanks to my dissertation committee, Eric Renault, Adam McCloskey, and Mardi Dungey. I am deeply grateful to my advisor, Eric, for his words of wisdom, his guidance, and his constant encouragement. His level of commitment as an advisor is truly admirable and made a big difference in my path at Brown University. I am very grateful to Adam, who in the last years of my Ph.D. was always there to listen and support me. He always listened with no judgment and gave me invaluable advice for both research and life. I am indebted to Mardi, who since our first meeting at Brown University supported my research with a high level of enthusiasm and continued to do so even from the far lands of Australia. I also thank Andriy Norets and Susanne Schennach for valuable feedback and comments.

I would also like to thank Nickolai Riabov, Philipp Ketz, Hyojin Han and David Frazier for their helpful feedback and comments. I would like to thank Rachel Toncelli for making me a better English writer. I cannot thank her enough for the many hours she devoted to sit down with me and read my Econometrics papers in spite of not even being an Economist. I also owe many many thanks to Emily Oster. She was an amazing Job Market coordinator.

I would also like to thank my roommates, classmates, and friends Angelica Meinhofer and Maria Jose Boccardi for six great years of love, adventures and support. I would like to thank my great friend Desislava Byanova for our many discussions, for always listening, and for always being there for me. I also owe many thanks to Angelica Vargas who has been there for me unconditionally since the day I started the Ph.D. She always had the answer to the many questions I had and made my life at Brown much easier.

I am very thankful to all my friends at Brown University. They became my family during these years of the Ph.D. and made Providence a home. I would particularly like

to thank my officemates and friends Desislava Byanova, Michelle Marcus, Alexandra Effenberger, Sanjay Singh, and Kofi Acquah for making the office a great place to work. Thank you Michelle Marcus for answering all of our English questions, for spending so much time reading the abstracts and introductions of my papers, and for your constant help and support. I would also like to thank Morgan Hardy, Svetoslava Milusheva, Tim Squires, Philipp Ketz, Nickolai Riabov, and Rawa Harati for always being there for me.

I am very grateful to my fiance Jeremy David Kaufman for always being so supportive of my career and encouraging me each step of the way. He always made the effort to understand my work, in spite of not having heard of a matrix before in his life. He made me laugh in stressful times and help me whenever I needed it. Thank you mi amor. Lastly, I would like to thank my family for their unconditional love and support in all these years of my graduate education. Thank you all of you for always being there for me in spite of the distance.

Preface

This dissertation is comprised of the following three chapters: (1) "Causality and Markovianity: Information Theoretic Measures" (joint work with Eric Renault), my job market paper entitled (2) "Structural VAR and Financial Networks: A Minimum Distance Approach to Spatial Modeling," and my third year paper entitled (3) "GMM with Minimum Mean Squared Error." All three chapters are about econometric methodology for time series, with a particular focus on model selection and loss functions. They include both theoretical and empirical developments about Information Theory in Econometrics, Generalized Method of Moments, Networks, and Spatial Modeling. The common feature of all the econometric methodologies developed in this dissertation is the applicability to financial econometrics.

Chapter 1 contributes to the Information Theory in Econometrics literature. Many Information Theoretic Measures have been proposed for a quantitative assessment of causality relationships. While Gouriéroux, Monfort, and Renault (1987) had introduced the so-called Kullback Causality Measures, extending Geweke's (1982) work in the context of Gaussian VAR processes, Schreiber (2000) has set a special focus on Granger causality and dubbed the same measure "transfer entropy." Both papers measure causality in the context of Markov processes. One contribution of this chapter is to set the focus on the interplay between measurement of (non)-markovianity and measurement of Granger causality. Both can be framed in terms of prediction of how much the forecast accuracy is deteriorated when some relevant conditioning information is forgotten. In this chapter we argue that this common feature between (non)-markovianity and Granger causality has led people to overestimate the amount of causality because what they consider as a causality measure may also convey a measure of the amount of (non)-markovianity. We set a special focus on the design of measures that properly disentangle these two components. Furthermore, this disen-

tangling leads us to revisit the equivalence between the Sims and Granger concepts of non-causality and the Log-Likelihood Ratio tests for each of them. We argue that a proper assessment of Granger causality implies testing for non-nested hypotheses. For the sake of illustration, we provide some quantitative assessment of the overestimation of Granger causality due to a confusion with non-markovianity. The numerical evidence shows that the amount of overestimation can be serious in certain scenarios.

Chapter 2 fits the network literature, and, more precisely, it focuses on the estimation of financial networks. Spatial models are a natural way to analyze spillovers or network effects. However, these models only provide an estimate of the overall network influence parameter and require the researcher to know a-priori how individuals are connected to each other. Since data on network ties are seldom available, researchers have mostly relied on ad-hoc network structures. In this paper, I put forth a methodology to estimate both the network matrix and the overall network influence parameter in a time series framework. I show that a time series spatial model is a constrained Structural Vector Autoregressive (SVAR) model. Based on these restrictions, the main theoretical contribution of this paper is to propose a two-step minimum distance approach to estimate the (row-standardized) network matrix and the overall network influence parameter from the SVAR estimates. I discuss machine learning methods, based on the PC-algorithm, as one possible identification strategy of SVAR models. To assess the restrictiveness of the constraints imposed by the spatial model on the SVAR model, I develop a Wald-type test. The methodology is illustrated through an application to financial integration among countries based on daily realized volatility data for the 2003-2015 period. I discuss different network measures and graphical tools to interpret the network effects. I find that the overall network influence was the highest during the financial crisis in 2008. After the crisis, the network influence decreased, though it did not return to pre-crisis levels.

Chapter 3 contributes to the generalized method of moments literature. In this

chapter I revisit the concepts of redundancy and partial redundancy when moments are unbiased, i.e. equal to zero, but their estimators are asymptotically biased. I show that, under these circumstances, redundancy and partial redundancy should be understood in terms of Asymptotic Mean Squared Error (AMSE) instead of efficiency. For this purpose, I first reassess what an optimal weighting matrix for a Generalized Method of Moments (GMM) estimator would be under the presence of asymptotic bias in the estimator of the moment conditions. Next, based on this new definition of optimal weighting matrix, I show that adding valid moment conditions does not hurt in terms of AMSE. As a result, this framework allows us to keep the standard point of view that additional moment conditions cannot increase asymptotic MSE under AMSE-efficient GMM. Then, using these results, I derive and reinterpret the necessary and sufficient conditions for redundancy and partial redundancy as originally defined by Breusch, Qian, Schmidt, and Wyhowski (1999). I also provide a discussion about the role, in terms of redundancy, of the variance bias trade-off.

Contents

List of Tables

List of Figures

Chapter 1

Causality and Markovianity: Information Theoretic Measures

1.1 Introduction

As first raised by Geweke (1982, p.304), using measures of causality is important, beyond testing for independence or unidirectional causality, because "in the typical case in which the hypothesis of independence ... is not literally entertained," it requires that one be able to measure the actual degree of causality. We argue in this paper that this measurement issue should be adjusted according to the goal of the study. On the one hand, following Sims (1972, p.540), one may want to question "the practice of making causal interpretations of distributed lag regressions of income on money" or, more generally, to make an economic interpretation of the evidence of some degree of causality. While, on the other hand, one may simply want to choose a forecasting formula, in the spirit of Hoel (1947). The former objective entails a measure of causality that may be different from the one needed for the latter objective.

According to this duality of objectives, we emphasize that two alternative measures of causality may be relevant, depending upon the question at stake. The dif-

ference is actually due to a new point of view on causality measurement. The main contribution of our new approach is to disentangle the measure of Granger causality with a measure of non-Markovianity.

Geweke's (1982) measure of linear feedback has been path-breaking and led to a very general Kullback causality measure proposed by Gouriéroux, Monfort, and Renault (1987) (GMR throughout) for a general process under the maintained assumption that (X, Y) is a Markov process of order p. This Kullback causality measure has been more recently dubbed "transfer entropy" by Schreiber (2000). Even though this framework has a very general validity, we question in this paper a maintained point of view shared by Geweke (1982), GMR, as well as Schreiber (2000), and summarized as follows by GMR (p.390): "despite the fact that the process X does not have a marginal autoregressive representation of order p, the marginal regression has to be performed after a truncation at **this order** p." We introduce the bold notation for the definition of "**this order** p" because it encapsulates our difference with the former literature.

We stress that obviously if (X, Y) is jointly Markov of order p, it is also Markov of order $(p+r)$ for any positive number r. While this non-uniqueness of Markov order is immaterial for statistical inference about the joint process (X, Y), it is crucial when stating as GMR (p.390) that "the marginal regression has to be performed after a truncation at this order p." We note that when the marginal regression is performed after a truncation at order $(p+r)$, increasing r could dramatically reduce the Granger causality measure. Of course, when Y does not cause X in the Granger sense, X is itself Markov of order p. This is the reason why Geweke (1982, p.309) notes that the auto-regression of X on p of its own lags is constrained maximum likelihood (constrained by the non-causality hypothesis) so that the proposed causality measure corresponds to the likelihood ratio test.

However, we argue that, when measuring causality, we precisely have in mind a

case where causality is at play, so that X is not marginally a Markov process of order p. There is no reason to perform a regression of X on only p (or even $p+r$) of its own lags. Increasing r decreases the causality measure and it is only for r tending to infinity that we rightly assess the degree of causality. The overestimated causality measure obtained by a marginal regression on $(p+r)$ lags encapsulates not only the right causality measure, but also a measure of "non-Markovianity." That is, how far the process X is from being Markov of order $(p+r)$. While, generalizing Geweke (1982), GMR prove asymptotic equivalence between their Kullback causality measure and the aforementioned likelihood ratio test statistic, this equivalence is valid only under the null hypothesis of non-causality. In contrast, we consider, like Geweke (1982, p.304), that it is precisely in the typical case in which the hypothesis of non-causality is "not literally entertained" that the causality measure is useful. Hence, one needs to revisit the measure proposed by Geweke (1982) and generalized by GMR.

The main innovation of this paper is to revisit the issue of causality measure in a framework of non-nested hypotheses. We argue that, when the process (X, Y) is assumed to be Markov of order p, assessing the degree of Granger causality from Y to X is not equivalent to comparing the probability distributions of X given p lagged values, including and not including lagged values of Y. When lagged values of Y are not included, the probability distribution of X given its own lagged values should rather include $(p+r)$ lags, with a large r. Since the Kullback contrast is not endowed with the symmetry property of a genuine distance, considering two non-nested models leads to two possible candidates for a causality measure: either the DGP (Data Generating Process) corresponds to the model where having only p lags of X is compensated by including p lags of Y, or it corresponds to the model where not including lagged values of Y is compensated by having $p+r$, instead of p, lagged values of X. In the former case, the causality measure is given by the Kullback contrast between the DGP and the closest candidate distribution without causality, evaluated

at the DGP. In the latter case, one works the other way round by maintaining the assumption that the process X alone is Markov of order $(p+r)$. That is, one wonders what is the probability distribution the closest to the DGP, when allowing for only p lags on the distribution of X given the past but adding past p values of Y to the conditioning information. The Kullback contrast computed in this manner provides an alternative causality measure.

These two alternative causality measures correspond to the two possible strategies for the Cox procedure of Modified Likelihood Ratio (MLR) principle for testing non-nested hypotheses. The first approach tests whether the null hypothesis "(X, Y) Markov of order p" can be rejected by considering the performance of a forecasting equation where X can be forecasted from $(p + r)$ of its own lags. We argue that this first point of view is better suited for addressing the issue of the choice of a forecasting formula, in the spirit of Hoel (1947). When the Kullback measure takes a large value, it means that a large number of lagged values of X is not sufficient to mimic the actual predicting distribution of X given p lagged values of X and Y jointly. Following Hoel's (1947) terminology, the multivariate framework allowing for the forecast of X by using not only its own past but also the past of other variables Y can be seen as "the new forecasting formula," while working only with the past of X is the "old formula." Then, it is natural to make the null hypothesis correspond to the new formula. Typically we have in mind circumstances in which the new formula has some advantages in terms of parsimony because $[p.\dim(Y)] < [r.\dim(X)]$.

The second approach tests whether the null hypothesis "X Markov of order $(p+r)$" can be rejected by considering the performance for forecasting X of an alternative model for (X, Y) that is jointly Markov of order p. Rejecting the null hypothesis in this context provides evidence of Granger causality from Y to X that is even more compelling than the standard likelihood ratio test computed with the constrained version of a parametric model for a process (X, Y) jointly Markov of order p. In fact,

in this situation, the Cox MLR test statistic tells us that not only the p lagged values of Y have significant coefficients when the forecasting formula includes p lagged values of X, but also they are significant if one adds a large number r of lagged values of X. Therefore, our second Kullback causality measure and the associated Cox test is better suited when the question at stake is really the theoretical question of non-causality from Y to X (see, e.g., the discussion in Sims, 1972).

As extensively discussed by Pesaran and Weeks (2001), the dual aspect of tests for non-nested hypotheses leads to two different interpretations: model choice for the purpose of decision making or testing for the empirical validity of a theoretical prediction. The former interpretation, and thus our first causality measure, is more about decision making: choosing the best forecasting formula. By contrast, the second causality measure is really assessing the amount of causality from Y to X, in the sense of a theoretical statement.

Besides the aforementioned existing causality measures, as well as the application in econometrics of the Cox principle for MLR (see, e.g., Fisher and McAleer, 1981; Pesaran, 1982; and Gouriéroux and Monfort, 1994), the methodology set forth in this paper is also related to the concept of "asymmetric VAR" as put forward by Hsiao (1979, 1981) and Keating (2000). The idea of asymmetric VAR is to identify different lag orders for different variables entering a multivariate autoregressive process. The latter authors have precisely stated the importance of this approach for causality testing. For instance Keating (2000, p.2) emphasizes that "each equation of a bivariate VAR could have 3 lags of output and 6 lags of money," which is analogous to us having a larger number $(p + r)$ of lags for the candidate exogenous variables X.

Finally, it is worth noting that the mutual implications between non-Markovianity and causality that has obliged us to cautiously derive a strategy of non-nested hypotheses is immaterial in the context of Sims non-causality. It has been documented in the literature that Sims and Granger non-causality are not fully equivalent for

at least two reasons: role of initial variables (Chamberlain, 1982 and Florens and Mouchart, 1982) as well as the impact of a third variable Z included in the forecasting environment (Dufour and Tessier, 1993). While the introduction of a third variable (or a set of variables) Z in the environment is beyond the scope of this paper, we revisit the role of initial values in the context of causality measures. It sheds more light on the interpretation of the non-equivalence pointed out by Chamberlain (1982) and Florens and Mouchart (1982), but also on the role of the Markov assumption in discussing this equivalence (see also Florens, Mouchart, and Rolin, 1993).

This paper is organized as follows. In the second section, we revisit the Kullback causality measures for Sims and Granger non-causality (from Y to X) as already introduced by GMR. However, in contrast with these authors, we do not maintain any Markov assumption. This allows us to characterize the difference between the two concepts and their measures through a concept and a measure of initiation of X by Y. In the third section, we analyze extensively the impact of the Markov assumption on our causality measures. More generally, we document the possible overestimation of causality measures due to the discrepancy between a true value and a pseudo-true value. Besides the theoretical definition of the measures, we describe their sample counterparts and their connection with a Cox procedure for testing non-nested hypotheses. The asymmetry of the Cox procedure leads us to explicitly disentangle two testing strategies, depending on whether the focus of interest is the choice of a practical forecasting formula or really an empirical assessment of some theoretical statement on non-causality. For sake of illustration we provide in section four some numerical evidence of the overestimation of Granger causality due to a confusion with non-markovianity. Finally, section five provides some concluding remarks and sketches some paths for extensions. The main mathematical proofs are gathered in the Appendix.

1.2 Information theoretic causality measures

1.2.1 General framework

We consider a multivariate discrete time process that starts at time $t = -\varpi$ and is denoted by $\{Z_t; t \geq -\varpi\}$. Each component Z_t is partitioned into two subvectors, X_t and Y_t, whose ranges are respectively $\Re(X)$ and $\Re(Y)$, some subsets of Euclidean spaces; the range of Z_t is $\Re(X) \times \Re(Y)$. Vectors $(X'_{t+h}, X'_{t+h-1}, ..., X'_t)', (Y'_{t+h}, Y'_{t+h-1}, ..., Y'_t)'$ and $(Z'_{t+h}, Z'_{t+h-1}, ..., Z'_t)'$ are denoted by X_t^{t+h}, Y_t^{t+h} and Z_t^{t+h} respectively. The probability distribution of the process $\{Z_t; t \geq -\varpi\}$ is defined by:

(i) The probability distribution of the initial vector $Z_{-\varpi}^0$ which is assumed to have a probability density function $f_0\left(z_{-\varpi}^0\right) = f_0\left(x_{-\varpi}^0, y_{-\varpi}^0\right)$ with respect to a product measure $\otimes_{i=0}^{\varpi}\mu(dz_{-i})$ where $\mu(dz)$ is itself a product measure $\mu_x(dx) \times \mu_y(dy)$ on $\Re(X) \times \Re(Y)$;

(ii) The conditional probability distributions of Z_t given $Z_{-\varpi}^{t-1}$, for any $t \geq 1$, which are assumed to have a p.d.f. $f_{0t}\left(z_t \middle| z_{-\varpi}^{t-1}\right)$ with respect to μ (which, for notational simplicity, is assumed to be the same for all t).

The joint p.d.f. of $Z_{-\varpi}^t$ with respect to $\otimes_{i=-\varpi}^t \mu(dz_i)$ is:

$$f_{0t}\left(z_{-\varpi}^t\right) = f_0\left(z_{-\varpi}^0\right) \prod_{i=1}^{t} f_{0i}\left(z_i \middle| z_{-\varpi}^{i-1}\right)$$

Note that the notation $f_{0t}\left(z_t \middle| z_{-\varpi}^{t-1}\right)$ allows the process $\{Z_t; t \geq -\varpi\}$ to be non-stationary. For instance, even in the case of a Markov process of order $p \leq \varpi + 1$, that is, the case when

$$f_{0t}\left(z_t \middle| z_{-\varpi}^{t-1}\right) = f_{0t}\left(z_t \middle| z_{-p+1}^{t-1}\right),$$

the transition density $f_{0t}(z_t \mid z_{-p+1}^{t-1})$ may depend on t. In order to accommodate non-stationarity, our definitions below of non-causality depend on the range of observations, namely $t = 1, ..., T$, on top of possibly some conditioning initial values $t = 0, -1, ..., -\varpi$. In other words, every time we write "Y does not cause X" in some sense, the reader should understand that "Y does not cause X between dates $(-\varpi)$ and T."

1.2.2 Kullback measure of Sims causality

Following Chamberlain's (1982) extension of Sims' (1972) concept of non-causality, we first define:

Definition 1.1 (Sims non-causality)**.** *Y does not cause X in the Sims sense ($Y.NCS.X$) if, for all $t = 1, ..., T$:*

$$f_{0t}\left(y_t \middle| x_{-\varpi}^{T}, y_{-\varpi}^{t-1}\right) = f_{0t}\left(y_t \middle| x_{-\varpi}^{t}, y_{-\varpi}^{t-1}\right)$$

Following GMR, we define a Kullback measure of causality from Y to X, in the Sims sense, by considering the set of all possible p.d.f.s $f_t\left(z_{-\varpi}^{t}\right)$, $t = 1, 2, ..., T$, for which, possibly by contrast with the Data Generating Process (DGP) $f_{0t}\left(z_{-\varpi}^{t}\right)$, $t = 1, 2, ..., T$, there is no causality in the Sims sense from Y to X:

$$HS[Y \nrightarrow X] = \{f_T(\cdot)\,;\; Y.NCS.X\}$$

with, by definition,

$$f_T\left(z_{-\varpi}^{T}\right) = f\left(z_{-\varpi}^{0}\right) \prod_{t=1}^{T} f_t\left(z_t \middle| z_{-\varpi}^{t-1}\right)$$

Then, we can define a discrepancy between the DGP and the non-causality hy-

pothesis in the Sims sense by solving the program:

$$\begin{cases} \text{Min } \dfrac{1}{T}\mathbb{E}_0 \log\left(\dfrac{f_{0T}\left(z_{-\varpi}^{T}\right)}{f_T\left(z_{-\varpi}^{T}\right)}\right) \\ \text{s.t.} \quad f_T(\cdot) \in HS[Y \nrightarrow X] \end{cases} \tag{1.1}$$

We define our Kullback measure of Sims causality from Y to X, denoted by $CS_{Y\to X}(T)$, as the value of the minimization program (1.1). We can show that:

Proposition 1.1. *The value of the program* (1.1) *is:*

$$CS_{Y\to X}(T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_0 \log\left(\frac{f_{0t}\left(y_t\middle|x_{-\varpi}^{T}, y_{-\varpi}^{t-1}\right)}{f_{0t}\left(y_t\middle|x_{-\varpi}^{t}, y_{-\varpi}^{t-1}\right)}\right)$$

The notation $CS_{Y\to X}(T)$ stresses that the Sims causality measure depends on the range $\{1, 2, ..., T\}$ of observations, while the initial date $(-\varpi)$ is considered throughout as given.

Then, obviously:

Corollary 1.1.

$$\begin{aligned} CS_{Y\to X}(T) &\geq 0 \\ CS_{Y\to X}(T) &= 0 \Leftrightarrow Y.NCS.X \end{aligned}$$

1.2.3 Initiation of X by Y

The Sims non-causality property allows the econometrician to write down a model in which endogenous variables y_t are explained by predetermined variables, that comprises past endogenous variables, $y_{-\varpi}^{t-1}$, as well as past and present exogenous variables, $x_{-\varpi}^{t}$. However, in order to fully specify the conditional probability distribution of the observed path z_1^T given the initial values $z_{-\varpi}^{0}$, it also takes a specification of the initiation of the observed path x_1^T of exogenous variables by possibly the endogenous

ones $y^0_{-\varpi}$. This is obvious from the following decomposition where both the roles of Sims non-causality and initiation are displayed in bold notations:

$$\frac{f_T\left(z^T_{-\varpi}\right)}{f_0\left(z^0_{-\varpi}\right)} = f_T\left(x^T_1 \middle| x^0_{-\varpi}, \mathbf{y}^0_{-\varpi}\right) \prod_{t=1}^{T} f_t\left(y_t \middle| x^{\mathbf{T}}_{-\varpi}, y^{t-1}_{-\varpi}\right) \tag{1.2}$$

While Sims non-causality is about some simplification of the second part of the RHS of Eq. (1.2), the initiation issue arises from the first part. We then define:

Definition 1.2 (Non-initiation)**.** *Y does not initiate X ($Y.NI.X$) if:*

$$f_T\left(x^T_1 \middle| x^0_{-\varpi}, y^0_{-\varpi}\right) = f_T\left(x^T_1 \middle| x^0_{-\varpi}\right)$$

Similarly to the methodology of causality measurement described above, we will define a Kullback measure of initiation of X by Y by considering the set of all possible p.d.f. $f_t\left(z^t_{-\varpi}\right)$, $t = 1, 2, ..., T$, for which, possibly by contrast with the DGP $f_{0t}\left(z^t_{-\varpi}\right)$, $t = 1, 2, ..., T$, there is no such initiation:

$$HIN[Y \nrightarrow X] = \{f_T(\cdot);\ Y.NI.X\}$$

Then, we can characterize a discrepancy between the DGP and the non-initiation hypothesis by solving the program:

$$\begin{cases} \text{Min } \dfrac{1}{T}\mathbb{E}_0 \log\left(\dfrac{f_{0T}\left(z^T_{-\varpi}\right)}{f_T\left(z^T_{-\varpi}\right)}\right) \\ \text{s.t.} \quad f_T(\cdot) \in HIN[Y \nrightarrow X] \end{cases} \tag{1.3}$$

We define our Kullback measure of initiation of X by Y, denoted by $CIN_{Y\to X}(T)$, as the value of the minimization program (1.3). We can show that:

Proposition 1.2. *The value of the program* (1.3) *is:*

$$CIN_{Y\to X}(T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_0 \log\left(\frac{f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}, y_{-\varpi}^{0}\right)}{f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}\right)}\right)$$

Then, obviously:

Corollary 1.2.

$$\begin{aligned} CIN_{Y\to X}(T) &\geq 0 \\ CIN_{Y\to X}(T) &= 0 \Leftrightarrow Y.NI.X \end{aligned}$$

It is worth noting that $CIN_{Y\to X}(T)$ is a Cesaro mean of a sequence u_t with

$$u_t = \mathbb{E}_0 \log\left(\frac{f_{0t}(x_t\left|x_{-\varpi}^{t-1}, y_{-\varpi}^{0}\right)}{f_{0t}(x_t\left|x_{-\varpi}^{t-1}\right)}\right)$$

Following Chamberlain (1982), we consider the following condition (R):

Condition (R).

$$\lim_{t=\infty} u_t = 0$$

Under regularity conditions, our Condition (R) is equivalent to condition (R) introduced by Chamberlain (1982, p.571). Chamberlain (1982, p.571) notes that "condition (R) requires that the current effect of y's from the distant past vanishes; similar assumptions are routine in the analysis of aggregate time-series data." Even though stationarity is not a maintained assumption for us, Condition (R) is obviously related to the concept of ergodicity. From the aforementioned Cesaro mean argument, we deduce that:

Proposition 1.3. *Under Condition (R):*

$$\lim_{T=\infty} CIN_{Y\to X}(T) = 0$$

It is worth noting that more often than not, at least in a stationary environment, this limit will be a decreasing limit because the sequence u_t itself should be non-increasing. To see that, imagine the case of a process $\{Z_t;\ t \geq -\varpi\}$ that is stationary Gaussian. Then, the question is akin to the behavior of a function:

$$h(n) = \mathbb{E}\log\left(\frac{f\left(Z|X_1, X_2, ..., X_n, Y\right)}{f\left(Z|X_1, X_2, ..., X_n\right)}\right), \quad n = 1, 2, .., N \tag{1.4}$$

when p.d.f.s are computed from the joint Gaussian distribution of a vector $(Z, X_1, X_2, ..., X_N, Y)'$. Obviously, the difference between the numerator and the denominator in (1.4) is fully explained by the informational content of the difference:

$$Y - \mathbb{E}[Y\,|X_1, X_2, ..., X_n]$$

Thus, this difference should obviously be reduced when n increases.

1.2.4 From Sims causality to Granger causality

Following Granger's (1969) approach to causality, we now define:

Definition 1.3 (Granger non-causality)**.** *Y does not cause X in the Granger sense ($Y.NCG.X$) if, for all $t = 1, ..., T$:*

$$f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}, y_{-\varpi}^{t-1}\right) = f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}\right)$$

Following GMR, we define a Kullback measure of causality from Y to X, in the Granger sense, by considering the set of all possible p.d.f.s $f_t\left(z_{-\varpi}^{t}\right), t = 1, 2, ..., T$ for

which, possibly by contrast with the Data Generating Process (DGP) $f_{0t}\left(z_{-\varpi}^{t}\right), t = 1, 2, ..., T$ there is no causality in the Granger sense from Y to X:

$$HG[Y \nrightarrow X] = \{f_T(\cdot);\ Y.NCG.X\}$$

Then, we can define a discrepancy between the DGP and the non-causality hypothesis in the Granger sense by solving the program:

$$\begin{cases} \text{Min}\ \dfrac{1}{T}\mathbb{E}_0 \log\left(\dfrac{f_{0T}\left(z_{-\varpi}^{T}\right)}{f_T\left(z_{-\varpi}^{T}\right)}\right) \\ \text{s.t.} \quad f_T(\cdot) \in HG[Y \nrightarrow X] \end{cases} \tag{1.5}$$

We define our Kullback measure of Granger causality from Y to X, denoted by $CG_{Y\to X}(T)$, as the value of the minimization program (1.5). We can show that:

Proposition 1.4. *The value of the program* (1.5) *is:*

$$CG_{Y\to X}(T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_0 \log\left(\frac{f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}, y_{-\varpi}^{t-1}\right)}{f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}\right)}\right)$$

Then, obviously:

Corollary 1.3.

$$\begin{aligned} CG_{Y\to X}(T) &\geq 0 \\ CG_{Y\to X}(T) &= 0 \Leftrightarrow Y.NCG.X \end{aligned}$$

Moreover, the comparison of propositions 1.1, 1.2, and 1.4 shows that:

Proposition 1.5.

$$CG_{Y\to X}(T) = CS_{Y\to X}(T) + CIN_{Y\to X}(T)$$

Then, since all these measures are non-negative, we deduce:

Corollary 1.4. *Y does not cause X in the Granger sense if and only if the two following conditions are fulfilled:*

(i) Y does not cause X in the Sims sense

(ii) Y does not initiate X

In other words, Proposition 1.5 sheds some new light on an issue already pointed out by Florens and Mouchart (1982, p.587) "Granger non-causality still implies, but is not equivalent to, Sims non-causality. Some care has to be taken (..) in order to handle the initial condition properly." Similarly to Theorem 3 in Chamberlain (1982, p.575), the conjunction of Proposition 1.4 and 1.5 shows that the equivalence is restored for large T if condition (R) holds:

Corollary 1.5. *Under Condition (R):*

$$\lim_{T=\infty} CG_{Y\to X}(T) = \lim_{T=\infty} CS_{Y\to X}(T)$$

1.3 Causality measures and pseudo-true values

1.3.1 General issue

For the purpose of statistical inference, causality measures have mainly been studied, so far, in the context of stationary Markov processes. Hence, we will maintain throughout the assumption that the process $\{Z_t; t \geq -\varpi\}$ is stationary Markov of order p. This assumption is especially convenient to simplify the expression of the Granger causality measure:

$$CG_{Y\to X}(T) = \mathbb{E}_0 \log\left[f_0\left(x_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}\right)\right] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_0 \log\left[f_{0t}\left(x_t \middle| x_{-\varpi}^{t-1}\right)\right] \tag{1.6}$$

where $f_0\left(x_t\middle|x_{t-p}^{t-1}, y_{t-p}^{t-1}\right)$ stands for the time invariant transition p.d.f. of this stationary Markov process. It is worth stressing that by contrast, in spite of stationarity, $f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}\right)$ depends in general on t because the process $\{X_t;\ t \geq -\varpi\}$ is not Markov and thus the number of relevant lagged values in the conditioning information increases with t. In the context of a parametric model, the marginal transition p.d.f. and the expectation of its logarithm may be computed by using the ARMA structure of the process $\{X_t;\ t \geq -\varpi\}$. However, this difficulty seems to have been largely overlooked in the literature. Instead of computing properly the causality measure $CG_{Y\to X}(T)$ in (1.6) as the value of the minimization program (1.5), people would in general prefer to simplify the computation by solving instead:

$$\begin{cases} \text{Min}\ \dfrac{1}{T} E_0 \log\left(\dfrac{f_{0T}\left(z_{-\varpi}^{T}\right)}{f_T\left(z_{-\varpi}^{T}\right)}\right) \\ \text{s.t.} \quad f_T(\cdot) \in HG_c[Y \nrightarrow X] \end{cases} \tag{1.7}$$

where $HG_c[Y \nrightarrow X]$ is a strict subset of $HG[Y \nrightarrow X]$ because some of the properties assumed for the DGP have been imposed to the candidate density function $f_T(\cdot)$. Obviously, this approach will lead in general to overestimate the actual level of causality, when the distance between the DGP and the reduced set $HG_c[Y \nrightarrow X]$ is actually also due to some other constraints. Examples of this issue in the extant literature are described below.

1.3.2 Causality and markovianity

Most of extant empirical applications of causality measures lie within a Markov framework: the process (X, Y) is assumed to be stationary Markov of some order p. Examples of application include Gaussian VAR(p) models (Geweke, 1982), Qualitative Panel data models (Bouissou, Laffont, and Vuong, 1986) and lattices in physical systems (Schreiber, 2000). All these examples fit within the framework described above

if we maintain the assumption that the number $(\varpi+1)$ of initial values is not smaller than p.

As already mentioned, Geweke (1982, p.309) is led to "begin by supposing that for purpose of estimation, all lag lengths in the canonical form have been truncated at p." It is a way to acknowledge that he has solved the minimization program (1.7) with:

$$\begin{aligned} HG_c[Y \nrightarrow X] &= HG\,[Y \nrightarrow X|p] \\ &= \{f_T\,(\cdot)\,; Y.NCG.X \text{ and } Z \sim Mar(p)\} \end{aligned} \tag{1.8}$$

where the notation $Z \sim Mar(p)$ means that the process $\{Z_t;\ t \geq -\varpi\}$ is stationary Markov of order p. We can show more generally that:

Proposition 1.6. *If $Z \sim Mar(p)$, and $p \leq q \leq \varpi+1$, the value $CG_{Y\to X}(q)$ of the program (1.7) with $HG_c[Y \nrightarrow X] = HG\,[Y \nrightarrow X|q]$ is:*

$$CG_{Y\to X}(q) = \mathbb{E}_0 \log\left[f_0\left(x_t\middle|x_{t-p}^{t-1}, y_{t-p}^{t-1}\right)\right] - \mathbb{E}_0 \log\left[f_0\left(x_t\middle|x_{t-q}^{t-1}\right)\right]$$

The proof of Proposition 1.6 is straightforward from (1.6) with the Mar(q) assumption in the definition of $HG_c[Y \nrightarrow X]$. Moreover, note that since by stationarity and Markov property the Granger causality measure does not depend on T anymore, we have simplified the notation.

Obviously:

Corollary 1.6.

$$q \leq q' \leq \varpi+1 \Rightarrow CG_{Y\to X}(q) \geq CG_{Y\to X}(q') \geq CG_{Y\to X}(T)$$

It is worth noting that $CG_{Y\to X}(q)$ overestimates the actual measure $CG_{Y\to X}(T)$ of causality precisely because it adds a term measuring the amount of non-markovianity

at degree q for the process X. More precisely

Corollary 1.7. *If $Z \sim Mar(p)$, and $p \leq q \leq \varpi + 1$:*

$$CG_{Y\to X}(q) - CG_{Y\to X}(T) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_0 \log\left(\frac{f_{0t}\left(x_t \middle| x_{-\varpi}^{t-1}\right)}{f_0\left(x_t \middle| x_{t-q}^{t-1}\right)}\right) = M_X(q)$$

where $M_X(q)$ is the value of the minimization program:

$$\begin{cases} \text{Min } \dfrac{1}{T}\mathbb{E}_0 \log\left(\dfrac{f_{0T}\left(z_{-\varpi}^{T}\right)}{f_T\left(z_{-\varpi}^{T}\right)}\right) \\ s.t. \quad X \sim Mar(q) \end{cases} \tag{1.9}$$

Recall that $Z \sim Mar(p)$ is the maintained assumption. The reason why it is the markovianity of the process X that is now at stake is the following. Obviously X is Markov of order p when Y does not cause X in the Granger sense (see Florens et al., 1993 for necessary and sufficient conditions). However, when looking for the p.d.f.s with non-causality the closest to the DGP, there is no reason to limit ourselves to stochastic processes that are themselves Markov of order p. After all, when Y does cause X, one may sometimes realize that the accuracy loss (in forecasting a future value of X) due to omission of the past values of Y in the conditioning information is not so important when one allows herself to keep a large number of lagged values of X in the conditioning information. In this case, a large value of $CG_{Y\to X}(p)$ may rather be seen as a signal that X is far from being a Markov process (of the small order q considered) rather than evidence of a high degree of causality from Y to X. Actually, Corollary 1.7 points out the right additive decomposition to assess the trade off between causality and non-markovianity:

$$CG_{Y\to X}(q) = CG_{Y\to X}(T) + M_X(q) \tag{1.10}$$

If we think about Kullback-Leibler discrepancy (KL herefafter) as a squared norm

in some Hilbert space (see for instance the formula of KL in the case of Gaussian distributions, Kullback (1968, p.189) or Gouriéroux and Monfort (1995, p.15), see also the example in the next subsection), equality (1.10) looks like an application of Pythagoras' theorem. The non-causality property $Y.NCG.X$ and the Markov property $X \sim Mar(q)$ (when $Z = (X', Y')' \sim Mar(p),\ p \leq q$) look like defined by two "orthogonal directions", the squared distances to them being additive.

Over-estimation of the causality measure $CG_{Y \to X}(T)$ due to the Markov measure $M_X(q)$ is of course maximum for the choice $q = p$, which is typically the common practice in applications (due to the fact that X is Markov of order p when Y does not cause X). By contrast, since the theory tells us that we should rather consider a model with an infinite number of lags, in practice one would have a large number q. A possible way to devise a proper asymptotic theory for this practice would be to consider a number q_T of lags going to infinity with T, albeit slower than T (see, e.g., Berk, 1974).[1] Another possible approach would be to resort to ARMA models (see Boudjellaba, Dufour, and Roy, 1992 and Dufour, Pelletier, and Renault, 2006 for related discussions). The example of Gaussian transition densities will allow us to assess even further the possible over-estimation of causality due to the use of pseudo-true values.

1.3.3 The case of a parametric model for a Markov process

In this section, we discuss more explicitly the assessment of a causality measure like $CG_{Y \to X}(p)$ as characterized in Proposition 1.6 above:

$$CG_{Y \to X}(p) = \mathbb{E}_0 \log \left[f_0 \left(x_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1} \right) \right] - \mathbb{E}_0 \log \left[f_0 \left(x_t \middle| x_{t-p}^{t-1} \right) \right]$$

[1] We are grateful to a referee for this suggestion.

under the maintained assumption that the process (X, Y) is Markov of order p with transition densities defined by a parametric model:

$$f_0\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}\right) = f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}; \theta^0\right)$$

where θ^0 stands for the true unknown value of the parameter vector $\theta \in \Theta \subset \mathbb{R}^p$.

1.3.3.1 The case of a Gaussian process

It is the case of the Gaussian VAR(p) as studied by Geweke (1982). For the sake of notational simplicity, we will assume throughout that the process X is of dimension 1. Then, the conditional probability distribution of X_t given $(X_{t-p}^{t-1'}, Y_{t-p}^{t-1'})'$ (resp. given X_{t-p}^{t-1}) is the normal distribution with mean $m_t = \mathbb{E}[X_t \mid (X_{t-p}^{t-1'}, Y_{t-p}^{t-1'})']$ (resp. $m_t^* = \mathbb{E}(X_t \mid X_{t-p}^{t-1})$) and variance σ^2 (resp. σ^{*2}) where:

m_t (resp. m_t^*) is the affine regression of X_t on $(X_{t-p}^{t-1'}, Y_{t-p}^{t-1'})'$ (resp. on X_{t-p}^{t-1})

and:

$$\sigma^{*2} = \sigma^2 + Var\left[m_t - m_t^*\right] \tag{1.11}$$

Then, by virtue of the formula for KL in the case of Gaussian distributions:

$$CG_{Y \to X}(p) = \log\left(\frac{\sigma^*}{\sigma}\right) + \frac{1}{2}\left(\frac{\sigma^2}{\sigma^{*2}} - 1\right) + \frac{Var\left[m_t - m_t^*\right]}{2\sigma^{*2}} = \log\left(\frac{\sigma^*}{\sigma}\right) \tag{1.12}$$

where the last equality is obtained thanks to (1.11). As noted by GMR, this provides a parametric interpretation of the KL causality measure in terms of pseudo-true value.

More precisely, we can write:

$$f_0\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}\right) = f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}; \theta^0\right)$$

where θ^0 stands for the true unknown value of the parameter vector θ whose components include the coefficients of the affine function $\mathbb{E}[(X_t, Y_t) \mid (X_{t-p}^{t-1\prime}, Y_{t-p}^{t-1\prime})']$ as well as the parameters to describe the residual variance (including the parameter σ^2 for the residual variance of X_t). Thus, instead of solving (1.7) over a set of probability distributions defined by (1.8), one could have contemplated rather solving the minimization program (1.7) with:

$$\begin{aligned} HG_c[Y \nrightarrow X] = \{ & f_T(\cdot);\ Y.NCG.X \text{ and} \\ & \exists \theta \in \Theta,\ f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}\right) = f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \theta\right)\} \end{aligned} \tag{1.13}$$

It turns out that (1.13) does not add any constraint with respect to (1.8). As it is obvious from (1.11) and (1.12) we actually have:

$$CG_{Y \to X}(p) = \mathbb{E}_0 \log\left[f\left(x_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \theta^0\right)\right] - \mathbb{E}_0 \log\left[f\left(x_t \middle| x_{t-p}^{t-1};\ \nu(\theta^0)\right)\right] \tag{1.14}$$

where $\nu(\theta^0)$ defines the KL pseudo-true value of θ, that is the value of θ that makes the transition density $f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \theta\right)$ the closest possible (in the KL sense) to the DGP, subject to restrictions defined by (1.13). From (1.11) and (1.12) respectively, we see that $f\left(x_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \nu(\theta^0)\right)$ is the normal distribution with mean:

$$m_t^* = \mathbb{E}\left(X_t \middle| X_{t-p}^{t-1};\ \theta^0\right)$$

and variance:

$$\sigma^{*2} = (\sigma^0)^2 + Var\left[\mathbb{E}\left(X_t \middle| (X_{t-p}^{t-1'}, Y_{t-p}^{t-1'})';\ \theta^0\right) - \mathbb{E}\left(X_t \middle| X_{t-p}^{t-1};\ \theta^0\right)\right]$$

Formula (1.14) explains why Geweke (1982) noticed that the sample counterpart of the Kullback causality measure is nothing but (up to scaling) the Likelihood Ratio test (LR hereafter) statistic for the null of non-causality. A sample counterpart of the pseudo-true value is the constrained Maximum Likelihood Estimator (MLE) of θ. Of course, due to inequalities of Corollary 1.6, it may be misleading to test for non-causality by just looking at the sample counterpart of $CG_{Y\to X}(p)$. A reinterpretation of this problem in terms of testing strategy is that, as usual, the likelihood ratio test does not take properly into account the need to compare models of similar dimensions. The Akaike Information Criterion (AIC) approach put forward in this context by Polasek (1994, 2002) does not fix this issue. We would have a valid comparison between models of similar dimensions if we would have compared a specification Markov of order p for Z (with causality from Y to X) with a specification Markov of order $p\left[1+\frac{\dim(Y)}{\dim(X)}\right]$ for X in isolation.

1.3.3.2 The general case

GMR go one step further by proposing more generally to compute causality measures within a given parametric model by using the pseudo-true value. More precisely, they would generally consider:

$$HG_c^*[Y\to X] = \{f_T(\cdot);\ Y.NCG.X \text{ and}$$
$$\exists\theta\in\Theta,\ f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}\right) = f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \theta\right)\}$$

instead of

$$HG_c[Y \nrightarrow X] = \{f_T(\cdot)\,;\; Y.NCG.X \text{ and } Z \sim Mar(p)\}$$

As a result, they come up with an alternative causality measure (see their formula for $\tilde{D}$ on page 384) generalizing the idea of (1.14) above:

$$CG^*_{Y\to X}(p) = \mathbb{E}_0 \log\left[f\left(x_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\; \theta^0\right)\right] - \mathbb{E}_0 \log\left[f\left(x_t \middle| x_{t-p}^{t-1};\; \nu(\theta^0)\right)\right]$$

However, it is worth realizing that the coincidence between the two definitions of causality measures ($CG_{Y\to X}(p) = CG^*_{Y\to X}(p)$) that we found in the former subsection does not generalize beyond the simple Gaussian setting. Obviously, $HG^*_c[Y \nrightarrow X]$ is a strict subset of $HG_c[Y \nrightarrow X]$ and we should find more often than not that:

$$CG^*_{Y\to X}(p) > CG_{Y\to X}(p) \tag{1.15}$$

As already mentioned, GMR show in their Theorem 8 that, under the non-causality hypothesis, the sample counterparts of the two causality measures do not differ by more than a $o_P(1/T)$ term, so that after re-scaling by a factor $(2T)$ to compute a likelihood ratio test statistic, one is led to two asymptotically equivalent test statistics. But once more, this equivalence is valid only under the null hypothesis of non-causality, that is precisely the case when causality measurement is irrelevant. By contrast , when it is relevant, by tightening her hands within a given parametric model, the econometrician will in general overestimate even more the actual amount of causality, by comparison with only maintaining the Markov (of same order p) assumption for the zero-causality proxy of the DGP. To see that, a simple generalization of the example of the former section featuring now conditional heteroskedasticity is sufficiently compelling. Let us assume now that both X and Y are univariate, $p = 2$,

and the conditional probability distribution of X_t given $(X_{t-2}^{t-1'}, Y_{t-2}^{t-1'})'$ is the normal distribution with mean m_t and variance σ_t^2 defined by:

$$\begin{aligned} m_t &= c + aX_{t-1} + bY_{t-1} \\ \sigma_t^2 &= \omega + \alpha \left(X_{t-1} - c - aX_{t-2} - bY_{t-2}\right)^2 \end{aligned} \tag{1.16}$$

Obviously the conditional distribution of X_t given X_{t-2}^{t-1} does not belong in general to the parametric model delineated by conditional normality with mean and variance in the parametric specification (1.16). Therefore:

$$\mathbb{E}_0 \log \left[f\left(x_t \middle| x_{t-2}^{t-1};\ \nu(\theta^0)\right)\right] < \mathbb{E}_0 \log \left[f_0\left(x_t \middle| x_{t-2}^{t-1}\right)\right]$$

which in turns implies (1.15). The econometrician who uses an estimator of $CG^*_{Y\to X}(p)$ to assess the amount of causality from Y to X will overestimate this amount for two reasons. First, when X is forecasted from its own past, a Markov model of order $p = 2$ is not appropriate. A higher order would give a more reliable assessment. Second, even within the set of processes that are Markov of order 2, there is no reason to believe that X in isolation is properly described by a $AR(1) - ARCH(1)$ model conformable to (1.16) (with $b = 0$). It is well known that the GARCH model is not robust to marginalization (see Nijman and Sentana, 1996).

1.3.4 Testing non-nested hypotheses

The main lesson of this section is that, when wondering whether there is a significant amount of causality from Y to X, one should not tight her hands by assuming that a model that sets the focus on the dynamics of X in isolation should be determined by the model at stake for the joint dynamics of X and Y. The two models should be non-nested. Of course, this comes with an important consequence when it comes

to testing. Since, as shown above, the sample counterpart of KL causality measures leads naturally to likelihood ratio tests, we should resort to the Cox procedure for testing non-nested hypotheses (see, e.g., Cox, 1962; Pesaran, 1974; and Gouriéroux and Monfort, 1994). When causality from Y to X is not precluded, we have in mind a parametric model where the process (X, Y) is Markov of order p with transition densities defined by:

$$f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1}\right) = f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \theta\right),\ \theta \in \Theta$$

The corresponding maximum (conditional) log-likelihood function evaluated under this hypothesis, let us call it $H\left(X|Y\right)$, is:

$$L_T\left\{\hat{\theta}_T;\ H\left(X|Y\right)\right\} = \sum_{t=1}^{T} \log\left[f\left(x_t, y_t \middle| x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \hat{\theta}_T\right)\right]$$

where $\hat{\theta}_T$ stands for the (conditional) MLE in this model.

By contrast, when we want to maintain the null hypothesis of (Granger) non-causality from Y to X, we contemplate an asymmetric model in which the process X is Markov of order $q > p$:

$$f\left(x_t, y_t \middle| x_{-\varpi}^{t-1}, y_{-\varpi}^{t-1}\right) = f\left(x_t \middle| x_{t-q}^{t-1};\ \lambda\right) f\left(y_t \middle| x_{t-p}^{t}, y_{t-p}^{t-1};\ \lambda\right),\ \lambda \in \Lambda \tag{1.17}$$

Note that we have assumed in (1.17) that Y_t is independent of Z_{t-r} for $r > p$. Let us call $H(X \mid X)$ this model. It is just an example of a possible way to specify the dynamics of (X, Y) in an asymmetric fashion. For instance, the extant literature on asymmetric VAR (see, e.g., Keating, 2000) would rather assume:

$$f\left(x_t, y_t \middle| x_{-\varpi}^{t-1}, y_{-\varpi}^{t-1}\right) = f\left(x_t \middle| x_{t-q}^{t-1};\ \lambda\right) f\left(y_t \middle| x_{t-q}^{t}, y_{t-p}^{t-1};\ \lambda\right),\ \lambda \in \Lambda$$

Even though we will stick to the specification (1.17) for sake of expositional simplicity, it is important to keep in mind that different specifications are always possible for the factor $f\left(y_t \middle| x_{-\varpi}^t, y_{-\varpi}^{t-1}\right)$ which is not the focus of our interest regarding a KL measure of Granger causality from Y to X. The corresponding maximum (conditional) log-likelihood function evaluated under hypothesis $H(X \mid X)$ is:

$$L_T\left\{\hat{\lambda}_T;\ H\left(X|X\right)\right\} = \sum_{t=1}^{T}\left\{\log\left[f\left(x_t\middle|x_{t-q}^{t-1};\ \hat{\lambda}_T\right)\right] + \log\left[f\left(y_t\middle|x_{t-p}^{t}, y_{t-p}^{t-1};\ \hat{\lambda}_T\right)\right]\right\}$$

The usual LR test statistic is defined by:

$$\xi_T^{LR} = 2\left[L_T\left\{\hat{\theta}_T;\ H(X \mid Y)\right\} - L_T\left\{\hat{\lambda}_T;\ H(X \mid X)\right\}\right]$$

Under the null hypothesis $H(X \mid Y)$, the normalized version of ξ_T^{LR} tends to:

$$\operatorname*{plim}_{T=\infty}\frac{1}{2T}\xi_T^{LR} = \mathbb{E}\left[\log\left(f\left(x_t\middle|x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \theta^0\right)\right)\right] - \mathbb{E}\left[\log\left(f\left(x_t\middle|x_{t-q}^{t-1};\ \tilde{\lambda}(\theta^0)\right)\right)\right]$$

where $\tilde{\lambda}(\theta^0)$ is the pseudo-true value of λ under the maintained hypothesis $H(X \mid Y)$. In other words, with an obvious extension of the notations introduced in the previous subsection we have:

$$\operatorname*{plim}_{T=\infty}\frac{1}{2T}\xi_T^{LR} = CG_{Y\to X}^{*}(q) \tag{1.18}$$

Recall that in the particular case of a Gaussian process, we have $CG_{Y\to X}^{*}(q) = CG_{Y\to X}(q)$. More generally, the message of the result (1.18) is the following regarding the value added by the introduction in this paper of a new causality measure $CG_{Y\to X}^{*}(q)$ with $q > p$. A Cox test based on the above LR test statistic (after proper centering and rescaling) will be interpreted as follows. If the Cox test does not succeed to reject the null while q is large in front of p, it means that the VAR(p) model for

(X, Y) (or more generally a Markov(p) model) does not do a bad job in forecasting X with a parsimonious model of $2p$ regressors instead of a large number q of regressors taken from the own lagged values of X. In contrast, in the case where $q = p$, the non-causality hypothesis becomes nested within the VAR(p) model and failing to reject means that the causality measure is not sufficiently large for a compelling evidence of causality from Y to X.

However, it must be acknowledged that there is no such thing as a free lunch and both strategies, either taking $q = p$ as Geweke (1982) and GMR or taking q much larger than p as advocated in this paper have their own shortcomings. On the one hand, as discussed above (see also the next section for quantitative evidence), taking $q = p$ may lead the researcher to believe that there is a large amount of causality from Y to X, while it may be rather the non-markovanity of X that is at stake. On the other hand, taking $q > p$ implies that we are testing for non-nested hypotheses, which comes with a cost, as well explained by Dastoor (1981, p.115) "we can only accept or reject the hypothesis that is under test, without any implication whatsoever for the alternative."

The bottom line is that the Cox test based on (1.18) can only accept or reject the null hypothesis that the forecasting performance of a VAR(p) model for (X, Y) (or more generally a Markov(p) model) is not dominated by the one based on a large number q of regressors taken from the own lagged values of X. This is exactly the aforementioned Hoel's (1947) point of view. We will reject the null when causality from Y to X is not sufficiently strong to improve the forecast accuracy by comparison with the alternative "univariate" forecasting strategy based only on lagged values of X. In other words, the approach is more "model selection" than "hypothesis testing" according to the classification in Pesaran and Weeks (2001, p.288): "Model selection is more appropriate when the objective is decision making", namely forecasting X in this case. If by contrast we have in mind a hypothesis testing approach that is "better

suited to inferential problems where the empirical validity of a theoretical prediction is the primary objective", like the absence of feedback from income to money in the Sims example, we should rather consider the non-nested testing problem the other way round, that is, under the null hypothesis $H(X \mid X)$ defined by (1.17) against the alternative of the VAR(p) model for (X, Y). In this case:

$$\begin{aligned}\operatorname*{plim}_{T=\infty} \frac{1}{2T}\xi_T^{LR} &= \mathbb{E}\left[\log\left(f\left(x_t\middle|x_{t-p}^{t-1}, y_{t-p}^{t-1};\ \tilde{\theta}(\lambda^0)\right)\right)\right] - \mathbb{E}\left[\log\left(f\left(x_t\middle|x_{t-q}^{t-1};\ \lambda^0\right)\right)\right] \\ &= -CG_{Y\to X}^{**}(q)\end{aligned}$$

where it must be understood that the new causality measure $CG_{Y\to X}^{**}(q)$ is non-negative (it is not the opposite of $CG_{Y\to X}^{*}(q)$) because expectations are now computed under the maintained hypothesis $H(X \mid X)$ with a true value λ^0 and a pseudo-true value $\tilde{\theta}(\lambda^0)$.

1.4 Quantitative Assessment

Consider a Gaussian process (X, Y) with a VAR(1) representation; we take X and Y univariate. We focus on the Granger causality measure from Y to X. The DGP of (X, Y) is given by

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = A \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \quad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \ \Sigma = Var\left[\begin{pmatrix} u_t \\ v_t \end{pmatrix}\right] = \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} \tag{1.19}$$

We can consider a matrix A that can be diagonalized and rewrite it as follows:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = M \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} M^{-1} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix} \tag{1.20}$$

where M is the square (2×2) matrix whose i^{th} column is the m_i eigenvector of A, and λ_i, $i = 1, 2$, is the corresponding i^{th} eigenvalue of A.

This reparametrization is useful for two purposes. First, to easily guarantee a causal VAR process when choosing parameter values, that is, $|\lambda_i| < 1$ (i.e. inside the unit circle), for $i = 1, 2$. Second, to understand how the results change when the persistence of past information (i.e. X_{t-1} and Y_{t-1}) in the DGP changes. More precisely, with

$$\begin{bmatrix} X_t^* \\ Y_t^* \end{bmatrix} = M^{-1} \begin{bmatrix} X_t \\ Y_t \end{bmatrix}$$

we have

$$\begin{bmatrix} X_t^* \\ Y_t^* \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} X_{t-1}^* \\ Y_{t-1}^* \end{bmatrix} + M^{-1} \begin{bmatrix} u_t \\ v_t \end{bmatrix}$$

with

$$\begin{aligned} X_t^* &= \frac{m_{22}}{Det(M)} X_t - \frac{m_{12}}{Det(M)} Y_t \\ Y_t^* &= \frac{-m_{21}}{Det(M)} X_t + \frac{m_{11}}{Det(M)} Y_t \end{aligned}$$

and similar expressions apply for X_{t-1}^* and Y_{t-1}^*. Therefore, to interpret λ_1 as "mostly about" the persistence of X and λ_2 as "mostly about" the persistence of Y, we conveniently fix M as follows

$$M = \begin{bmatrix} 0.9329 & -0.2492 \\ 0.3601 & 0.9684 \end{bmatrix}, \qquad Det(M) = 0.99322$$

Indeed, with this choice of M we obtain that

$$X_t^* = \beta_1 X_{t-1} + \beta_2 Y_{t-1}, \quad \frac{|\beta_1|}{|\beta_2|} > 3.8$$
$$Y_t^* = \beta_3 X_{t-1} + \beta_4 Y_{t-1}, \quad \frac{|\beta_4|}{|\beta_3|} > 2.5$$

such that λ_1 (respectively λ_2), autoregressive coefficient of X_t^* (respectively Y_t^*), is mostly informative about the persistence of X (respectively Y).

Our goal is to study $CG_{Y\to X}$ as computed by a researcher who "wrongly" sticks to GMR and Schreiber (2000) using q lags, with $q \geq p$, for the marginal of X_t instead of the correct measure which takes $\varpi + 1$. Hence, the marginal autoregression for X_t is

$$X_t = \sum_{j=1}^{q} \alpha_j X_{t-j} + \epsilon_t, \qquad \mathbb{E}[\epsilon_t] = 0, \quad V[\epsilon_t] = \sigma_\epsilon^2(q) \tag{1.21}$$

Remark 1.1. *Notice that here the underlying model for the joint process (X, Y) is the following Asymmetric VAR model, where the equation for Y_t remains unchanged when compared to the DGP,*

$$X_t = \sum_{j=1}^{q} \alpha_j X_{t-j} + \epsilon_t, \qquad \mathbb{E}[\epsilon_t] = 0, \quad V[\epsilon_t] = \sigma_\epsilon^2(q)$$
$$Y_t = \gamma\, X_{t-1} + \delta\, Y_{t-1} + \eta_t, \qquad \mathbb{E}[\eta_t] = 0, \quad V[\eta_t] = \sigma_\eta^2$$

or in matrix notation

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = A_1 \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \cdots + A_q \begin{bmatrix} X_{t-q} \\ Y_{t-q} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix},$$

where

$$A_1 = \begin{bmatrix} \alpha_1 & 0 \\ \gamma & \delta \end{bmatrix}, \quad A_j = \begin{bmatrix} \alpha_j & 0 \\ 0 & 0 \end{bmatrix} \text{ for } j = 2, ..., q, \quad \Sigma_q = \begin{bmatrix} \sigma_\epsilon^2 & \rho_q \sigma_\epsilon \sigma_\eta \\ \rho_q \sigma_\epsilon \sigma_\eta & \sigma_\eta^2 \end{bmatrix}$$

This leads to the following Kullback measure of Granger causality

$$CG_{Y \to X}(q) = \frac{1}{2} \log \frac{\sigma_\epsilon^2(q)}{\sigma_u^2} \tag{1.22}$$

where we write $\sigma_\epsilon^2(q)$ instead of σ_ϵ^2 to make explicit the dependence of the result on the choice of q. Our interest is on a population measure of Granger causality as a function of λ_1, λ_2, σ_u, σ_v, and ρ for different values of q. This can be derived in closed form solution as shown in the Appendix. All the numerical evidence is coded in Mathematica version 10.0.1.0.

Figure 1.1 depicts how $CG_{Y \to X}(q)$ changes as q increases, for a fixed configuration of the parameters of the model. Notice that $CG_{Y \to X}(q)$ is decreasing in q, which is consistent with our theoretical finding that if we do not include enough lags of X we end up overestimating the amount of causality. Indeed, the distance between two points is our measure of degree of non-markovianity. In addition, we see that the bigger jump is attained going from the order of the VAR in the DGP $q = p = 1$, to the next order $q = p+1 = 2$.[2] Finally, it is worth noting that after a finite number of lags ($q = 4$ in this figure) the measure stabilizes. The limit value is about 0.127, which we can interpret as the true value of $CG_{Y \to X}$. This shows that, for all practical purposes, we do not need to include an infinite amount of lags to perform a regression; a finite number of lags, large enough, would be sufficient not to overestimate the amount of causality.

[2]Given that we need to choose the value of five parameters to display this graph, it is difficult to show all possible scenarios. However, it is important to remark that the overall observed pattern is consistent across different configurations of parameters, beyond the one chosen for this particular figure.

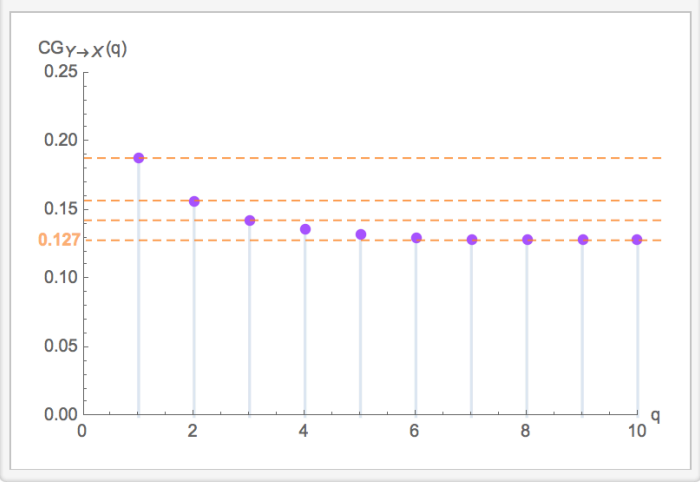


Figure 1.1: Comparison of $CG_{Y\to X}(q)$ across q for fixed values of the parameters: $\lambda_1 = -0.6$, $\lambda_2 = 0.9$, $\rho = 0.5$, $\sigma_u = 1$, $\sigma_v = 1$.

There can be cases in which the amount of overestimation is severe. This is a significant concern as it might lead to wrong conclusions about the amount of causality present in the problem at hand. For instance, in Figure 1.2, if we choose $q = p = 1$ we are lead to think that the amount of causality is around 0.20, while the true measure (i.e. for $q > 3$) is close to zero. The magnitude of the causality measure across values of q is displayed in Table 1.1. Notice that the true amount of causality cannot be exactly zero as in that case, from Corollary 1.6, we know that $CG_{Y\to X}(1) = CG_{Y\to X}(2) = \cdots = CG_{Y\to X} = 0$.

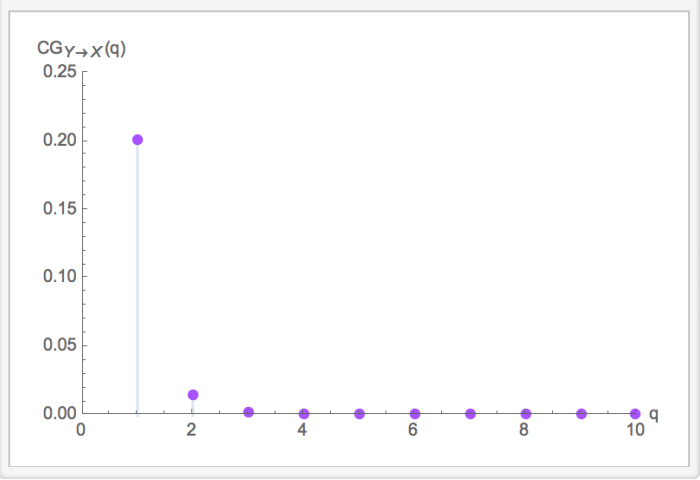


Figure 1.2: Comparison of $CG_{Y\to X}(q)$ for $q = 1, 2, ..., 10$: $\lambda_1 = -0.8$, $\lambda_2 = 0.8$ $\rho = -0.99$, $\sigma_u = 1$, $\sigma_v = 1$

Table 1.1: Value of $CG_{Y\to X}(q)$ for $q = 1, 2, ..., 10$

q	1	2	3	4	5	6	7	8	9	10
$CG_{Y\to X}(q)$	0.2011	0.0146	0.0026	0.0016	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015

The intuition in this particular case is as follows. The persistence of X (i.e. λ_1) and the persistence of Y (i.e. λ_2) are of the same order of magnitude, although of opposite signs since if $\lambda_1 = \lambda_2$ there would be no causality, i.e., $CG_{Y\to X} = 0$. Since the correlation between u_t and v_t is close to -1, after one period most of the information contained in the past of Y is already contained in the past of X.[3] As a result, we see a huge drop in the amount of causality going from $q = 1$ to $q = 2$. This is consistent with the idea that because X is highly persistent, and given that Y does not carry much additional information after one period, the degree of non-markovianity between $q = 1$ and $q = 2$ is of a considerable size.

To further the analysis on how persistence of the past of X impacts $CG_{Y\to X}(q)$ across values of q, in Figure 1.3 we compare $CG_{Y\to X}(1)$, $CG_{Y\to X}(2)$, and $CG_{Y\to X}(3)$ across values of λ_1 for given values of the reminding parameters. The overall effect of λ_1 on $CG_{Y\to X}(q)$ is comprised of two components. On the one hand, the amount of causality increases with the absolute value of the difference between λ_1 and λ_2, i.e., $|\lambda_1 - \lambda_2|$. The larger this difference, all else equal, the larger the value of $CG_{Y\to X}(q)$. In particular, when this difference is zero, i.e. $\lambda_1 = \lambda_2$, there is no causality.[4] As a result, the amount of causality needs not to be close to zero when λ_1 is low. This is particularly the case here since the persistence of the past of Y (i.e. λ_2) is also small.[5]. On the other hand, when there is causality, the amount of overestimation increases with $|\lambda_1 - \lambda_2|$. The difference across $CG_{Y\to X}(q)$ becomes particularly noticeable when

[3] This can be easily seen by thinking in terms of impulse response functions.

[4] This is clear from Eq. (1.20). For instance, in Figure 1.3, $CG_{Y\to X}(q) = 0$ at $\lambda_1 = 0.2$ since $\lambda_2 = 0.2$.

[5] For completeness, supplementary examples where the past of Y is highly persistent are provided in the Additional Figures and Tables section of the appendix. In addition, Table 1.2 in the appendix shows how $CG_{Y\to X}(2)$ changes as a function of σ_u/σ_v for different configurations of parameters that give $CG_{Y\to X}(1) = 0.10$.

going from $q = 1$ to $q = 2$.

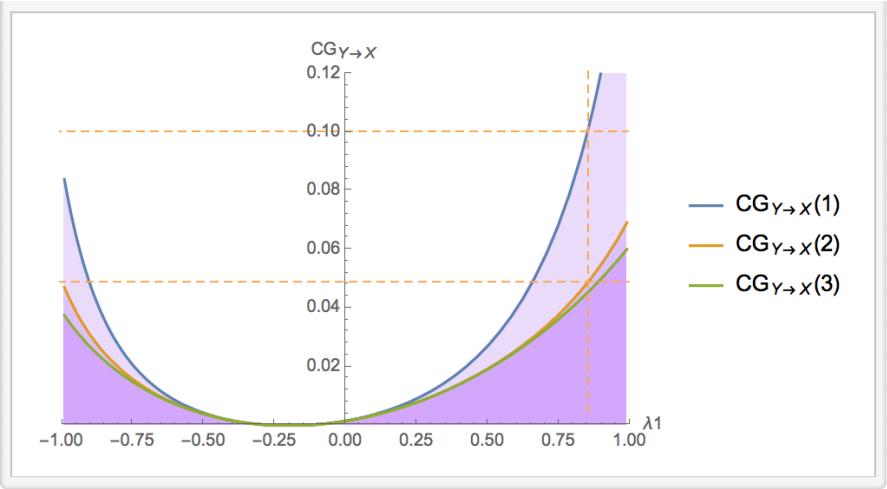


Figure 1.3: Comparison of $CG_{Y\to X}(1)$, $CG_{Y\to X}(2)$ and $CG_{Y\to X}(3)$ for different values of λ_1, and fixed values of the other parameters: $\lambda_2 = -0.2$, $\rho = -0.8$, $\sigma_u = 1$, $\sigma_v = 2$.

1.5 Concluding remarks

A common practice in econometrics amounts to assume that, when considering a sufficiently large number of state variables, the joint dynamics can be properly represented as a Markov process. Prominent examples of this practice are the extensive use of VAR processes in macro-econometrics as well as the use of diffusion processes in finance. This paper starts from the observation that, when a joint process (X, Y) is Markov, the statement that Y does not cause X entails two different phenomena: On the one hand, X will be in isolation a Markov process of the same order as (X, Y), and, on the other hand, information about lagged values of Y does not improve the performance of forecasts of X by comparison with what can be done using only past information about X. We argue in this paper that, in the opposite situation where Y does cause X, the two above properties (and their negations) should be disentangled when it comes to measurement of the strength of causality. On the one hand, how far is X to be a Markov process of the same order as (X, Y)? On the other hand, how much can we improve the accuracy of the forecasts of X by using past information

about Y?

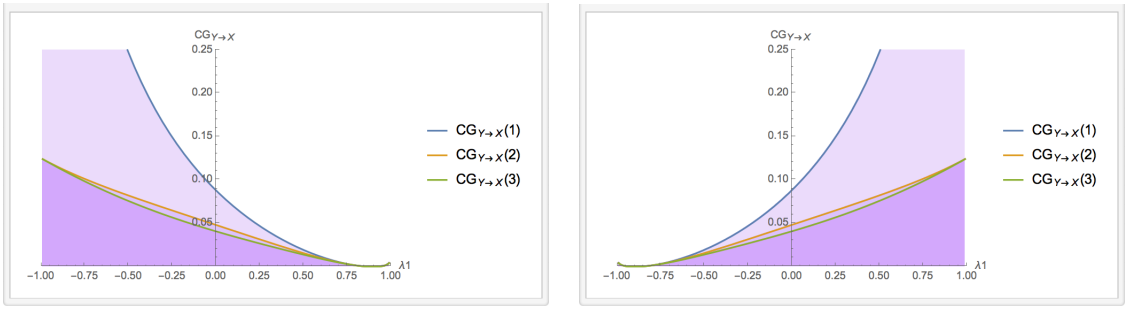
We note that the extant literature on causality measurement (Geweke, 1982; GMR; Schreiber, 2000) has failed to disentangle these two components and, as a result, may lead to the misleading belief that Y powerfully causes X, whereas we would actually be able to forecast X accurately only from its own past, at the cost of a larger number of lags than by using also the past of Y. Our understanding of the reason why the extant literature has failed to disentangle the two is that when testing for non-causality, the main focus is on size control and, thus, the behavior of the test statistic is essentially studied under the null hypothesis of non-causality when the two above statements are true together. We advocate in this paper an alternative strategy when testing for (non)-causality. One would actually consider two non-nested hypotheses: One focused on using Y to help forecasting X, and one focused on using a large number of lags to forecast X from its own past.

The price to pay for testing non-nested hypotheses is that the treatment is not symmetric: "we can only accept or reject the null hypothesis that is under test, without any implication whatsoever for the alternative" (Dastoor, 1981, p.115). We acknowledge that, as stressed more generally by Pesaran and Weeks (2001), the two ways to perform the Cox test for non-nested hypotheses correspond to two different points of view: Either defining the null hypothesis as the Markov property for (X, Y) puts the emphasis on the usefulness of a multivariate model (including Y) to forecast X, or, conversely, one would test the null that a large number of its own lags is sufficient to describe the dynamics of the X process, which is more a statement about "economic exogeneity" of X with respect to Y. In the former case, failing to reject the null advocates for the multivariate forecast. The issue of interest is more about non-nested comparison via predictive ability, as currently fashionable in the forecasting literature (for a recent illustration, see Noureldin, Shephard, and Sheppard, 2014). In the latter case, the issue that is addressed is in line with the research agenda put

forward for instance by Sims (1972) about exogeneity of money. In other words, the first approach corresponds to a measure of forecasting performance while the second one is more about genuine theoretical causality.

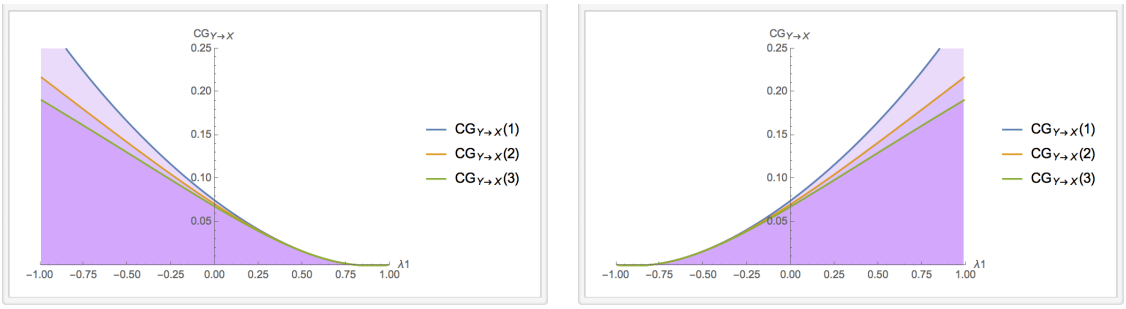
A natural extension of this paper would be to revisit these causality measures between Y and X in a conditional setting given additional state variables Z. Then, one should distinguish direct and indirect causality relationships, paving the way for different intensities of causality relationships at different horizons (see Dufour and Renault, 1998). Then, one should revisit the causality measures at different horizons (see also Dufour and Taamouti, 2010) in order to accommodate two kinds of time lags: time lag for Markov property, and time lag for impact of the cause on the effect.

1.A Additional figures and tables



(a) $\lambda_2 = 0.9$ **(b)** $\lambda_2 = -0.9$

Figure 1.4: Comparison of $CG_{Y\to X}(1)$, $CG_{Y\to X}(2)$ and $CG_{Y\to X}(3)$ for different values of λ_1, and fixed values of the other parameters: $\rho = -0.8$, $\sigma_u = 1$, $\sigma_v = 2$.



(a) $\lambda_2 = 0.9$ **(b)** $\lambda_2 = -0.9$

Figure 1.5: Comparison of $CG_{Y\to X}(1)$, $CG_{Y\to X}(2)$ and $CG_{Y\to X}(3)$ for different values of λ_1, and fixed values of the other parameters: $\rho = 0.5$, $\sigma_u = 1$, $\sigma_v = 1$.

Table 1.2: Analysis of $CG_{Y\to X}(2)$ for $CG_{Y\to X}(1) = 0.10$

5th Pctl	95th Pctl	Median	σ_u/σ_v
0.05697	0.09999	0.0992	0.1
0.06780	0.09994	0.0967	0.25
0.03823	0.09991	0.0913	0.5
0.02877	0.09929	0.0779	1
0.03406	0.08588	0.0664	2
0.04004	0.07411	0.0622	4
0.00000	0.06823	0.0596	10

Note: For each value of σ_u/σ_v, $CG_{Y\to X}(2)$ is computed for different configurations of λ_1,λ_2, and ρ, that give $CG_{Y\to X}(1) = 0.10$. The statistics (5th percentile, 95th percentile and median) are then computed based on these values of $CG_{Y\to X}(2)$.

1.B Mathematical appendix

1.B.1 Proof of Proposition 1.1

Consider the following decomposition

$$
\begin{aligned}
&\frac{1}{T}\underset{f_T\in HS}{\text{Min}}\,\mathbb{E}_0\log\left(\frac{f_{0T}\left(z^T_{-\varpi}\right)}{f_T\left(z^T_{-\varpi}\right)}\right) \qquad (1.23)\\
&=\frac{1}{T}\underset{f_T\in HS}{\text{Min}}\,\mathbb{E}_0\log\left(\frac{f_0(z^0_{-\varpi})f_{0T}\left(x^T_1\middle|z^0_{-\varpi}\right)f_{0T}\left(y^T_1\middle|x^T_{-\varpi},\ y^0_{-\varpi}\right)}{f(z^0_{-\varpi})f_T\left(x^T_1|z^0_{-\varpi}\right)f_T\left(y^T_1|x^T_{-\varpi},\ y^0_{-\varpi}\right)}\right)\\
&=\frac{1}{T}\underset{f_T\in HS}{\text{Min}}\left\{\mathbb{E}_0\log\left(\frac{f_0\left(z^0_{-\varpi}\right)}{f\left(z^0_{-\varpi}\right)}\right)\right.\\
&\quad+\mathbb{E}_0\log\prod_{t=1}^{T}\left(\frac{f_{0t}\left(x_t\middle|x^{t-1}_{-\varpi},\ y^0_{-\varpi}\right)}{f_t\left(x_t\middle|x^{t-1}_{-\varpi},\ y^0_{-\varpi}\right)}\right)\\
&\quad\left.+\mathbb{E}_0\log\prod_{t=1}^{T}\left(\frac{f_{0t}\left(y_t\middle|x^T_{-\varpi},\ y^{t-1}_{-\varpi}\right)}{f_t\left(y_t\middle|x^T_{-\varpi},\ y^{t-1}_{-\varpi}\right)}\right)\right\}
\end{aligned}
$$

Since expectations are by definition computed under the true DGP, we obviously have for any $f_T(\cdot)\in HS$ that each of the three terms in the above decomposition are non-negative. Therefore, the minimization of (1.23) will obviously lead to set the first

two terms to zero, since it is possible to set the denominator equal to the denominator while staying within HS. Then, to fully characterize the solution of (1.23), we only have to minimize the third term, that is, to maximize:

$$\mathbb{E}_0\left[\log\prod_{t=1}^{T} f_t\left(y_t\middle|x_{-\varpi}^{T},\ y_{-\varpi}^{t-1}\right)\right]=\mathbb{E}_0\left[\log\prod_{t=1}^{T} f_t\left(y_t\middle|x_{-\varpi}^{t},\ y_{-\varpi}^{t-1}\right)\right]$$

since, by the constraint of non-Sims causality that must be fulfilled by $f_T(\cdot)$, we know that:

$$f_t\left(y_t\middle|x_{-\varpi}^{T},\ y_{-\varpi}^{t-1}\right)=f_t\left(y_t\middle|x_{-\varpi}^{t},\ y_{-\varpi}^{t-1}\right)$$

By the Law of Iterated Expectations (LIE), the above maximization is equivalent to maximizing

$$\sum_{t=1}^{T}\mathbb{E}_0\mathbb{E}_0\left[\log f_t\left(y_t\middle|x_{-\varpi}^{t},\ y_{-\varpi}^{t-1}\right)\middle|x_{-\varpi}^{t},\ y_{-\varpi}^{t-1}\right]$$

Finally, by means of the Kullback inequality, it follows that for each t and for a given $(x_{-\varpi}^{t},\ y_{-\varpi}^{t-1})$, the conditional expectation is maximized by choosing $f_t\left(y_t\middle|x_{-\varpi}^{t},\ y_{-\varpi}^{t-1}\right)=f_{0t}\left(y_t\middle|x_{-\varpi}^{t},\ y_{-\varpi}^{t-1}\right)$. Hence, we result follows.

□

1.B.2 Proof of Proposition 1.2

Consider the following decomposition

$$
\begin{aligned}
&\frac{1}{T}\underset{f_T\in HIN}{\text{Min}}\,\mathbb{E}_0\log\left(\frac{f_{0T}\left(z^T_{-\varpi}\right)}{f_T\left(z^T_{-\varpi}\right)}\right) \qquad (1.24)\\
&=\frac{1}{T}\underset{f_T\in HIN}{\text{Min}}\,\mathbb{E}_0\log\left(\frac{f_0(z^0_{-\varpi})f_{0T}\left(x^T_1\middle|z^0_{-\varpi}\right)f_{0T}\left(y^T_1\middle|x^T_{-\varpi},\ y^0_{-\varpi}\right)}{f(z^0_{-\varpi})f_T\left(x^T_1|z^0_{-\varpi}\right)f_T\left(y^T_1|x^T_{-\varpi},\ y^0_{-\varpi}\right)}\right)\\
&=\frac{1}{T}\underset{f_T\in HIN}{\text{Min}}\left\{\mathbb{E}_0\log\left(\frac{f_0\left(z^0_{-\varpi}\right)}{f\left(z^0_{-\varpi}\right)}\right)\right.\\
&\quad+\mathbb{E}_0\log\prod_{t=1}^{T}\left(\frac{f_{0t}\left(x_t\middle|x^{t-1}_{-\varpi},\ y^0_{-\varpi}\right)}{f_t\left(x_t\middle|x^{t-1}_{-\varpi},\ y^0_{-\varpi}\right)}\right)\\
&\quad\left.+\mathbb{E}_0\log\prod_{t=1}^{T}\left(\frac{f_{0t}\left(y_t\middle|x^{T}_{-\varpi},\ y^{t-1}_{-\varpi}\right)}{f_t\left(y_t\middle|x^{T}_{-\varpi},\ y^{t-1}_{-\varpi}\right)}\right)\right\}
\end{aligned}
$$

Since expectations are by definition computed under the true DGP, we obviously have for any $f_T(\cdot)\in HIN$ that each of the three terms in the above decomposition are non-negative. Therefore, the minimization of (1.24) will obviously lead to set the first and third terms to zero, since it is possible to set the denominator equal to the denominator while staying within HIN. Then, to fully characterize the solution of (1.23), we only have to minimize the second term, that is, to maximize:

$$
\mathbb{E}_0\left[\log\prod_{t=1}^{T}f_t\left(x_t\middle|x^{t-1}_{-\varpi},\ y^0_{-\varpi}\right)\right]=\mathbb{E}_0\left[\log\prod_{t=1}^{T}f_t\left(x_t\middle|x^{t-1}_{-\varpi}\right)\right]
$$

since, by the non-Initiation constraint that must be fulfilled by $f_T(\cdot)$, we know that:

$$
f_t\left(x_t\middle|x^{t-1}_{-\varpi},\ y^0_{-\varpi}\right)=f_t\left(x_t\middle|x^{t-1}_{-\varpi}\right)
$$

By LIE, the above maximization is equivalent to maximizing

$$\sum_{t=1}^{T} \mathbb{E}_0 \mathbb{E}_0 \left[\log f_t \left(x_t \middle| x_{-\varpi}^{t-1} \right) \middle| x_{-\varpi}^{t-1} \right]$$

Finally, by means of the Kullback inequality, it follows that for each t and for a given $x_{-\varpi}^{t-1}$, the conditional expectation is maximized by choosing $f_t \left(x_t \middle| x_{-\varpi}^{t-1} \right) = f_{0t} \left(x_t \middle| x_{-\varpi}^{t-1} \right)$. Hence, the result follows.

□

1.B.3 Proof of Proposition 1.4

Consider the following decomposition

$$\begin{aligned} \frac{1}{T} \underset{f_T \in HG}{\text{Min}} \mathbb{E}_0 \log \left(\frac{f_{0T} \left(z_{-\varpi}^{T} \right)}{f_T \left(z_{-\varpi}^{T} \right)} \right) &= \frac{1}{T} \underset{f_T \in HG}{\text{Min}} \mathbb{E}_0 \log \left(\frac{f_0(z_{-\varpi}^{0}) \prod_{t=1}^{T} f_{0t} \left(z_t \middle| z_{-\varpi}^{t-1} \right)}{f(z_{-\varpi}^{0}) \prod_{t=1}^{T} f_t \left(z_t \middle| z_{-\varpi}^{t-1} \right)} \right) \\ &= \frac{1}{T} \underset{f_T \in HG}{\text{Min}} \left\{ \mathbb{E}_0 \log \left(\frac{f_0 \left(z_{-\varpi}^{0} \right)}{f \left(z_{-\varpi}^{0} \right)} \right) \right. \\ &\quad + \mathbb{E}_0 \log \prod_{t=1}^{T} \left(\frac{f_{0t} \left(x_t \middle| x_{-\varpi}^{t-1},\ y_{-\varpi}^{t-1} \right)}{f_t \left(x_t \middle| x_{-\varpi}^{t-1},\ y_{-\varpi}^{t-1} \right)} \right) \\ &\quad \left. + \mathbb{E}_0 \log \prod_{t=1}^{T} \left(\frac{f_{0t} \left(y_t \middle| x_{-\varpi}^{t},\ y_{-\varpi}^{t-1} \right)}{f_t \left(y_t \middle| x_{-\varpi}^{t},\ y_{-\varpi}^{t-1} \right)} \right) \right\} \end{aligned} \tag{1.25}$$

Since expectations are by definition computed under the true DGP, we obviously have for any $f_T(\cdot) \in HG$ that each of the three terms in the above decomposition are non-negative. Therefore, the minimization of (1.25) will obviously lead to set the first and last terms to zero, since it is possible to set the denominator equal to the denominator while staying within HG. Then, to fully characterize the solution of (1.25), we only have to minimize the third term, that is, to maximize:

$$\mathbb{E}_0 \left[\log \prod_{t=1}^{T} f_t \left(x_t \middle| x_{-\varpi}^{t-1},\ y_{-\varpi}^{t-1} \right) \right] = \mathbb{E}_0 \left[\log \prod_{t=1}^{T} f_t \left(x_t \middle| x_{-\varpi}^{t-1} \right) \right]$$

since, by the constraint of non-Sims causality that must be fulfilled by $f_T(\cdot)$, we know that:

$$f_t\left(x_t\middle|x_{-\varpi}^{t-1},\ y_{-\varpi}^{t-1}\right) = f_t\left(x_t\middle|x_{-\varpi}^{t-1}\right)$$

By LIE, the above maximization is equivalent to maximizing

$$\sum_{t=1}^{T}\mathbb{E}_0\mathbb{E}_0\left[\log f_t\left(x_t\middle|x_{-\varpi}^{t-1}\right)\middle|x_{-\varpi}^{t-1}\right]$$

Finally, by means of the Kullback inequality, it follows that for each t and for a given $x_{-\varpi}^{t-1}$, the conditional expectation is maximized by choosing $f_t\left(x_t\middle|x_{-\varpi}^{t-1}\right) = f_{0t}\left(x_t\middle|x_{-\varpi}^{t-1}\right)$. Hence, the result follows.

□

1.B.4 Proof of Corollaries 1.1, 1.2 and 1.3

The proof of these corollaries is obvious from the following lemma (See GMR for a proof of this lemma which corresponds to their Lemma 1).

Lemma 1.1. *Let X, Y, Z be three random vectors whose joint p.d.f., with respect to a given σ-finite measure, is denoted by $f(X, Y, Z)$. We have:*

(i)

$$\mathbb{E}\log\left(\frac{f(X\,|Y,Z)}{f(X\,|Y)}\right) \geq 0$$

(ii) This expectation is equal to zero if and only if, almost surely:

$$f(X\,|Y,Z) = f(X\,|Y)$$

1.B.5 Proof of Proposition 1.5

We have:

$$CS_{Y\to X} + CIN_{Y\to X} = \frac{1}{T}\mathbb{E}_0 \log\left(\prod_{t=1}^{T} \frac{f_{0t}(y_t \mid x_{-\varpi}^{T}, y_{-\varpi}^{t-1})}{f_{0t}(y_t \mid x_{-\varpi}^{t}, y_{-\varpi}^{t-1})} . \frac{f_{0t}(x_t \mid x_{-\varpi}^{t-1}, y_{-\varpi}^{0})}{f_{0t}(x_t \mid x_{-\varpi}^{t-1})}\right) \quad (1.26)$$

The numerator of the above product can be rewritten as:

$$\begin{aligned} & f_0(y_1^T \mid x_{-\varpi}^T, y_{-\varpi}^0) f_0(x_1^T \mid x_{-\varpi}^0, y_{-\varpi}^0) \\ = \; & f_0(x_1^T, y_1^T \mid x_{-\varpi}^0, y_{-\varpi}^0) \\ = \; & \prod_{t=1}^{T} f_{0t}(x_t, y_t \mid x_{-\varpi}^{t-1}, y_{-\varpi}^{t-1}) \\ = \; & \prod_{t=1}^{T} f_{0t}(x_t \mid x_{-\varpi}^{t-1}, y_{-\varpi}^{t-1}) f_{0t}(y_t \mid x_{-\varpi}^{t}, y_{-\varpi}^{t-1}) \end{aligned}$$

Plugging this result in (1.26) and noting that one factor cancels out with the denominator, we end up with:

$$CS_{Y\to X} + CIN_{Y\to X} = \frac{1}{T}\mathbb{E}_0 \log\left(\prod_{t=1}^{T} \frac{f_{0t}(x_t \mid x_{-\varpi}^{t-1}, y_{-\varpi}^{t-1})}{f_{0t}(x_t \mid x_{-\varpi}^{t-1})}\right) = CG_{Y\to X}$$

□

1.B.6 Closed form solution for $CG_{Y\to X}(q)$ in equation (1.22)

We first derive a closed form expression for $\sigma_\epsilon^2(q)$ in terms of population moments of X and Y. Let $Z_q = [X_{t-1}, ..., X_{t-q}]'$ then[6]

$$\sigma_\epsilon^2(q) = Var(X_t) - Cov(X_t, Z_q) Var(Z_q)^{-1} Cov(Z_q, X_t) \quad (1.27)$$

[6]Let $\beta := [\alpha_1, ..., \alpha_q]'$ be a $q\times 1$ vector. The equation for $\sigma_\epsilon^2(q)$ is derived using that $X_t = \beta' Z_q + \epsilon_t$, $\beta' = Cov(X_t, Z_q)Var(Z_q)^{-1}$, and $Var(\beta' Z_q) = Cov(X_t, Z_q)Var(Z_q)^{-1}Cov(Z_q, X_t)$.

Notice that the value of $\sigma_\epsilon^2(q)$ depends on q through Z_q. Hence we have

$$CG_{Y\to X}(q) = \frac{1}{2}\ \log \frac{\gamma_{XX0} - \begin{bmatrix}\gamma_{XX1} & \dots & \gamma_{XXq}\end{bmatrix} \Omega_q^{-1} \begin{bmatrix}\gamma_{XX1} \\ \vdots \\ \gamma_{XXq}\end{bmatrix}}{\sigma_u^2} \tag{1.28}$$

where Ω_q is the symmetric matrix of size q whose generic element is:

$$(\Omega_q)_{ij} = \gamma_{XX|i-j|}$$

To compute γ_{XXh}, $h = 0, 1, 2, \cdots, q$, we use the following recursion formulas:

$$\Gamma(h) = \begin{bmatrix}\gamma_{XXh} & \gamma_{XYh} \\ \gamma_{YXh} & \gamma_{YYh}\end{bmatrix} = A^h\Gamma(0)$$
$$\Gamma(0) = A\Gamma(0)A' + \Sigma$$

solved as

$$Vec\,\Gamma(0) = (Id_4 - A \otimes A)^{-1}\ Vec(\Sigma)$$

where Id_4 is the identity matrix of order 4, and $Vec(C)$ stands for the vectorization of some matrix C.

Chapter 2

Structural VAR and Financial Networks: A Minimum Distance Approach to Spatial Modeling

2.1 Introduction

The global financial crisis that originated in the US in 2008 gave rise to a large amount of literature on financial networks, especially in recent years (see, e.g., Billio, Getmansky, Lo, and Pelizzon (2012), Yang and Zhou (2013), Diebold and Yilmaz (2014)). Researchers realized that, in the current global economy, financial institutions are tightly connected to each other, which ultimately leads to some risk amplification. Indeed, financial connectedness plays a crucial role as a spillover mechanism when facing a financial crisis because it can jeopardize the stability of the financial system as a whole. In many contexts, however, financial connectedness or the network structure is unknown due to data limitations and the sensitive nature of financial information. Consequently, the question of how to model a financial network has gained the attention of many researchers.

There are two models commonly used to study network effects: spatial models and Structural Vector Autoregressive (SVAR) models. Spatial models have been studied for decades, in particular since Anselin (1988), and have a growing number of applications. These models have the advantage that they allow for the disentangling of the overall network influence parameter from the network matrix itself. The former is a scalar that provides a macro (or global) measure on how important the network effect is as a whole (i.e., how the network as a whole influences any individual of the system), while the latter focuses on micro (or individual) network effects and provides information on who is connected with whom and the strength of each individual connection. However, the use of spatial models in the financial econometrics literature is quite recent (see e.g., Cohen-Cole, Kirilenko, and Patacchini (2013), Borovkova and Lopuhaa (2012)). One of the main challenges to using spatial models is that the network must be known a priori. Indeed, the network (i.e., spatial weight matrix) relies mainly on the chosen definition for neighbor or distance between geographical units. However, no natural geographic definition of a network exists when it comes to financial settings. As a result, in the absence of data on the network ties, the network matrix is typically ad-hoc when using a spatial model.

Unlike spatial models, SVAR models provide an estimate of the network matrix. These models have been widely used both within Economics, e.g., in Macroeconomics, and outside Economics, e.g., in the Neurosciences, to estimate contemporaneous or network effects among a system of variables. SVAR models are a natural choice as they allow for the modeling of not only lagged dynamics—for instance, volatility tends to be strongly serially correlated—but also of the so called network effect, that is, how individuals in the system affect each other contemporaneously.[1] As an example, if we have data on volatility of the main stock return indexes of countries A, B, and

[1]The reader should understand "individuals" here as the statistical entities being studied (e.g., people, countries, states, banks, etc.). The network for these individuals is built based on some characteristic (i.e., variable) of interest (e.g., volatility of a stock return).

C, we are interested in measuring how much the volatility observed in B (and/or C) at time t contributes to the volatility observed in A at time t after controlling for lagged dynamics. These models are flexible in that the matrix of contemporaneous effects is left unrestricted (beyond the restrictions needed for identification purposes). However, unlike spatial models, SVAR models do not disentangle the overall network influence parameter from the network matrix itself. Having an assessment of the overall network influence is particularly insightful to understanding whether a network has a higher (or lower) impact in different periods of time. In the context of the global financial crisis of 2008, was the network influence stronger during the crisis, when compared to the pre-crisis period, as a result of banks (or countries) becoming highly interconnected with each other?

In this paper, I show that a time series version of a spatial autoregressive model (T-SAR hereafter) is a restricted SVAR model.[2] This allows me to estimate the network while, at the same time, disentangling the overall network influence parameter from the network matrix. Moreover, the elements of a row of the network matrix sum to one (i.e., it is row-standardized), a common choice among spatial models as it eases interpretation. For instance, in our A, B, C example, when the network matrix is row-standardized how much B and C affect A's volatility at time t is a weighted average of B's volatility and C's volatility at time t. Based on the restrictions imposed by the T-SAR on the SVAR, I propose a two-step Minimum Distance Estimation (MDE) approach to estimate the closest spatial model to the SVAR model; this is the main theoretical contribution of the paper. In the MDE approach the constrained estimator is obtained from a quadratic form based on the unconstrained estimator. First, I estimate the network matrix in the unconstrained SVAR. Then, in a second step, I minimize the standardized distance between this estimate and the network matrix that is subject to the constraints implied by the T-SAR model.

[2] I say a "time series" version since spatial models are commonly used for cross-sectional data.

In addition, I construct a test to assess the restrictions on the SVAR imposed by spatial modeling. The test statistic corresponds to a Wald-type test, thus it can be interpreted as a measure of the distance between the unrestricted (SVAR) and restricted (T-SAR) models. The MDE approach has the advantage, over constrained Maximum Likelihood Estimation (MLE), that it delivers a closed form solution for both the constrained estimator and the test statistic. The solution is a function of the matrix of constraints and the unconstrained estimator. This is useful to better understand the relationship between the constrained and unconstrained parameters. Furthermore, the MDE estimators are asymptotically equivalent to constrained MLE, so there is no efficiency cost from using the MDE approach (see, e.g., Gourieroux and Monfort (1989a)).

To estimate the T-SAR model from the SVAR, we first need to choose a SVAR identification strategy. In this paper, I explore machine learning methods based on the PC-algorithm as a data-driven strategy for identification of the SVAR.[3] In view of the limited use of these methods with financial data, I conduct a simulation study to derive guidelines for its implementation.[4] Once we obtain the network estimates, results can be interpreted through various micro and macro connectivity measures, built from the network estimates, and through graphical analysis. In particular, I propose to study the network globally via cohesive-blocks analysis, macro network effect (through the scalar coefficient of the spatial model), and the network impact by order of neighbors (i.e., direct and indirect neighbors). To the best of my knowledge, these three measures are new to the financial networks literature.[5]

Finally, I present an application of the methodology to financial integration among

[3]The acronym "PC" in PC-algorithm refers to that named after the authors **P**eter Spirtes and **C**lark Glymour (Spirtes, Glymour, and Scheines, 2000).

[4]For general simulation studies see, e.g., Spirtes et al. (2000) and Kalisch and Bühlmann (2007).

[5]I also propose the use of the Fruchterman-Reingold (FR hereafter) layout instead of the commonly used circle layout to graph the network. The FR layout, unlike the circle layout, reveals the underlying community structure by assigning node locations such that similar nodes are closer together. For instance, it is useful to visually detect communities (or blocks), bridges, and central nodes.

countries using daily realized return volatility data from June 2003 to March 2015. I apply the MDE methodology to the main stock indexes of 15 countries plus the European Union leading index. Even though the problems caused by the financial crisis of 2008 were due to connectivity among financial institutions, I focus on countries instead because these problems were ultimately dealt with on a country basis and had country level financial implications. The empirical findings show that the overall network influence was stronger during the 2008 crisis period, which is in line with other papers' findings that the network connections became more dense during that period (see e.g., Billio, Getmansky, Lo, and Pelizzon (2012)). Looking at specific time periods, the analysis shows that even though the overall network influence decreased post crisis, it remained above the pre-crisis level during the January 2013 to March 2015 period. This might be attributable to financial problems in the Euro-area, in particular those regarding Greece's debt crisis. Furthermore, a cohesive-blocks study shows that while there were distinct blocks within the network in the pre-crisis period, in the crisis period, however, all countries formed one cohesive block (with the exception of Australia and Mexico). This supports the idea that the crisis was indeed a global one that put many countries in the same "basket."

This paper relates to several papers in both the financial econometrics and spatial models literature. In the financial econometrics literature, two prominent examples are the work of Billio, Getmansky, Lo, and Pelizzon (2012) and Diebold and Yilmaz (2014). Billio et al. (2012) estimate the network in a VAR framework by means of a pairwise-Granger causality approach. It has the advantage that, by focusing on reduced form estimates, it avoids the SVAR identification problem. However, their approach does not measure the strength of the connections since it is based on counting the number of significant Granger causality connections. My constrained SVAR approach encompasses these Granger causality methods, since it identifies in particular the zero coefficients in lag matrices. Diebold and Yilmaz (2014) estimate

the network in a VAR framework by means of generalized variance decomposition (GVD). Even though, the GVD approach has the advantage of treating each variable symmetrically avoiding the order dependency issue of a Cholesky-type decomposition, the structural shocks are not necessarily orthogonal to each other. Also, their paper is different in that all individuals in the network are connected to each other to some degree. In the spatial models literature, two recent examples are Manresa (2015) and Lam and Souza (2015). The former proposes a methodology to estimate spillover effects (i.e., the network) in a panel-data framework, but it looks instead at how characteristics of others affect an individual in the network. The latter proposes a method to estimate the network matrix by means of an Adaptive LASSO estimator; however, it does not disentangle the overall network influence parameter from the network matrix.

This paper is organized as follows. Section 2.2 introduces the theoretical framework for SVAR models and discusses identification issues. Section 2.3 presents the Minimum Distance Estimator of the closest spatial model to the SVAR. Also, this section presents the test to assess how restrictive the constraints imposed by the spatial model on the SVAR model are. Section 2.4 discusses how to implement the MDE methodology using a data-driven identification strategy for the SVAR, based on the PC-algorithm. In addition, this section conducts a simulation study to assess the performance of the PC-algorithm to uncover the causal structure of the network. Section 2.5 addresses how to interpret and analyze the network matrix through different connectivity measures, both at a local and global level. Section 2.6 illustrates the methodology through an application to financial integration among countries by using data on daily log-realized return volatility of each country's main stock index. Finally, Section 2.7 presents concluding remarks. Proofs are gathered in the Mathematical Appendix 2.A; additional graphs and tables are gathered in Appendix 2.B.

2.2 A primer on SVARs

2.2.1 Preliminaries

In this paper I consider a K-dimensional vector of variables $Y_t = (y_{1,t},\ y_{2,t},\ ...,\ y_{K,t})'$ observed at time $t = 1, 2, \cdots, T$, and the number of initial periods needed are defined according to the autoregressive order, p. Without loss of generality, it is assumed that the vector process $\{Y_t\}$ is in deviation from its mean. Hence, $\{Y_t\}$ has zero mean and the deterministic terms are dropped from all the models presented in this paper.

The following matrix notations and operations are used throughout this paper. I_K is the identity matrix of order K. A row-standardized matrix W is a square matrix of real numbers, with each row summing to 1. ι is a vector of ones. O is a matrix of zeros, not necessarily square. The vectorization of a matrix C is denoted by vec(C). If A and B are conformable matrices, define $A \times B$ as the matrix multiplication. Define $A \otimes B$ as the Kronecker product between matrices A and B. If A and B have the same dimensions, define $A \odot B$ as the Hadamard product (or element-wise product) between matrices A and B. Let A be an $m \times m$ matrix, then denote $Diag(A)$ the $m \times 1$ vector containing the diagonal elements of A.

2.2.2 The structural and reduced form VAR model

A Structural Vector Autoregressive (SVAR) model of order p for the dynamics of $\{Y_t\}$ is given by

$$AY_t = B_1 Y_{t-1} + \cdots + B_p Y_{t-p} + \varepsilon_t, \tag{2.1}$$

where A is a $K \times K$ matrix with ones on its main diagonal, B_i, $i = 1, ..., p$, is a $K \times K$ coefficient matrix, and ε_t is a K-dimensional white noise term, that is, $E(\varepsilon_t) = 0$, $E(\varepsilon_t \varepsilon_t') = \Sigma_\varepsilon$ and $E(\varepsilon_t \varepsilon_s) = 0$ for $s \neq t$, with Σ_ε a diagonal and positive definite

covariance matrix. In the literature, this is known as the A-model, in which Σ_ε is always assumed to be a diagonal matrix but the unit main diagonal assumption on A is not required.[6] Although the unit main diagonal assumption on A is not part of the definition of an A-model, it is generally a maintained assumption because it is without loss of generality and it is useful for interpreting the model. This allows us to write the model's k-th equation with $y_{k,t}$ as the left-hand variable. That is, we can re-write the model as

$$Y_t = GY_t + B_1 Y_{t-1} + \cdots + B_p Y_{t-p} + \varepsilon_t \tag{2.2}$$

where $G = I_K - A$ is a $K \times K$ coefficient matrix with zero elements on its main diagonal. Then, G can be interpreted as a network matrix or matrix of contemporaneous effects. In other words, each variable contemporaneously affects other variables of the system but not itself (i.e., no self-loops). The usefulness of this assumption will become clear in Section 2.3 when I discuss the relationship between SVAR models and spatial models.

The parameters of the SVAR model (2.1) are identified from the reduced form model, known as the Vector Autoregressive (VAR) model. The reduced form representation implied by the structural model in (2.1) is given by:

$$Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + u_t \tag{2.3}$$

where $A_i = A^{-1}B_i,\ i = 1, ..., p$, $u_t = A^{-1}\varepsilon_t$, and u_t is a K-dimensional white noise, with $u_t \sim (0, \Sigma_u)$. The parameters of the reduced form model are $(A_1, \cdots, A_p, \Sigma_u)$, where the covariance matrix Σ_u is a symmetric and positive definite matrix. The reduced form VAR in equation (2.3) is causal provided that the roots of the charac-

[6] See Lütkepohl (2007) for further details.

teristic equation $\det(I_K - A_1 z - \cdots - A_p z^p) = 0$ lie outside the complex unit circle.[7] Only causal VAR models are considered in this paper.

2.2.3 Identification of SVAR models

2.2.3.1 The identification problem

Recovering the SVAR parameters from the reduced form VAR coefficients requires further identifying restrictions. The identification problem is discussed in what follows. The parameters of interest are typically comprised of A, the contemporaneous effect parameters, and $B_i, \;\; i = 1, ..., p,$, the lagged effect parameters. Ideally, the structural parameters would be identified from the reduced form parameters. However, there is an infinite set of different values of $B_i, \;\; i = 1, ..., p$, and A which all imply the exact same probability distribution for the observed data, making it impossible to infer from the data alone what the true values for $B_i, \;\; i = 1, ..., p$, and A are.

In other words, without additional restrictions, the SVAR parameters are not identified from the reduced form VAR parameters. This is a well known result in the SVAR literature, a simple proof of this result can be found in Gottschalk (2001, p.3) and further discussion can be found in Rubio-Ramírez, Waggoner, and Zha (2010, p.669) and Kilian (2013, p.519). This is the *identification problem:* without additional assumptions, i.e., identifying restrictions, no conclusions regarding the structural parameters of the "true" model can be drawn from the data. Without them, different structural models give rise to the same reduced form.

[7] For the case of $p = 1$, this condition is often stated as "the VAR is causal provided the eigenvalues of A_1, λ, have modulus less than 1." This is an equivalent statement since the eigenvalues of A_1 satisfy the equation $\det(I_K \lambda - A_1) = 0$ and are equal to the inverses of the roots of the characteristic equation $\det(I_K - A_1 z) = 0$. When $p > 1$ the eigenvalues' condition is about an augmented matrix based on $A_1, ..., A_p$, instead of on the eigenvalues of these matrices directly.

2.2.3.2 Identification restrictions

The matrix of coefficients A and Σ_ε can be recovered from Σ_u if additional restrictions are imposed. For simplicity of exposition, some of these restrictions discussed in what follows have already been imposed in model (2.1). For identification of the SVAR, there are two types of restrictions which are standard in the literature, namely, the orthogonality restriction and restrictions on the matrix A.[8]

The orthogonality restriction consists of assuming that the variance covariance matrix of the structural innovations, Σ_ε, is diagonal. Given Σ_u, we want to write:

$$\Sigma_u = A^{-1}\Sigma_\varepsilon A'^{-1}$$

Since Σ_u is symmetric it has $K(K+1)/2$ free parameters. Given that Σ_ε is assumed diagonal, it has K parameters. Thus, the dimension of the space of parameters for A is:

$$\frac{K(K+1)}{2} - K = \frac{K(K-1)}{2}$$

Out of the K^2 parameters of A, we only have a space of dimension $K(K-1)/2$. Hence we need

$$K^2 - \frac{K(K-1)}{2} = \frac{K(K+1)}{2}$$

restrictions for local identification.

Even though it is without loss of generality that we can normalize the main diagonal elements of A to one, this reduces the number of restrictions needed by K. Therefore, we need $K(K-1)/2$ restrictions after adopting this normalization. The

[8]Identification restrictions and conditions presented in this section are extracted from Lütkepohl (2007). Sign and inequality restrictions are beyond the scope of this paper.

simplest approach in the literature to comply with the remaining number of restrictions is to use exclusion restrictions on the matrix A, i.e., to set some elements of A to zero.

Remark 2.1. *Given ε_t, the k^{th} element of $\varepsilon_t = Au_t$ can be written as:*

$$\begin{aligned}\varepsilon_{kt} &= \sum_{h=1}^{K} a_{kh}u_{ht} \\ &= \sum_{h\neq k} a_{kh}u_{ht} + \left(\frac{a_{kk}}{\lambda_k}\right)(u_{kt}\lambda_k).\end{aligned}$$

We can choose, w.l.o.g., $\lambda_k = a_{kk}$. Therefore, the normalized diagonal elements of A are of the form $\tilde{a}_{kk} = a_{kk}/\lambda_k = 1\ \forall k$. This is w.l.o.g. as it amounts to a rescaling of u_{kt}.

Lütkepohl (2007) characterizes what it takes to give relevant constraints. Given the orthogonality restriction, local and global identification can be formally stated based on the set of restrictions on A as follows. Formally, the restrictions on A can be written as $C_A Vec(A) = c_A$, where C_A is a $K(K+1)/2 \times K^2$ selection matrix and c_A is a suitable $K(K+1)/2 \times 1$ fixed vector. Therefore, the restrictions have to be such that the system of equations

$$\Sigma_u = A^{-1}\Sigma_\varepsilon (A^{-1})' \quad and \quad C_A Vec(A) = c_A \tag{2.4}$$

has a unique solution, at least locally. The proposition below, corresponding to Proposition 9.1 of Lütkepohl (2007), presents a necessary and sufficient rank condition for local uniqueness of the solution.[9]

Proposition 2.1. *(Local identification of the A-model) Let Σ_ε be a $K \times K$ positive definite diagonal matrix. Then, for a given symmetric, positive definite $K \times K$ matrix*

[9]See Lütkepohl (2007, p.360) for a proof of this proposition.

Σ_u, *an* $r \times K^2$ *matrix* C_A *and a fixed* $r \times 1$ *vector* c_A, *the system of equations in* (2.4) *has a locally unique solution for* A *and the diagonal elements of* Σ_ε *if and only if*

$$rk \begin{bmatrix} -2D_K^+(\Sigma_u \otimes A^{-1}) & D_K^+(A^{-1} \otimes A^{-1})D_K \\ C_A & 0 \\ 0 & C_\sigma \end{bmatrix} = K^2 + \frac{1}{2}K(K+1)$$

where D_K *is a* $K^2 \times \frac{1}{2}K(K+1)$ *duplication matrix,* $D_K^+ = (D_K'D_K)^{-1}D_K'$, *and* C_σ *is a* $\frac{1}{2}K(K-1) \times \frac{1}{2}K(K+1)$ *selection matrix which selects the elements of* $vech(\Sigma_\varepsilon)$ *below the main diagonal.*

Although the above proposition provides necessary and sufficient conditions for local identification of the so called A-model, global identification is not guaranteed.[10] In the case of the A-model, if in addition the diagonal elements of A are restricted to 1, then the solution to system (2.4) is also globally unique (Lütkepohl, 2007, p.361).

Remark 2.2. *(Normalization of* Σ_ε*) The covariance matrix of the structural innovations is sometimes normalized such that* $\Sigma_\varepsilon = I_K$. *Nonetheless, this normalization is not convenient for our purposes. The reason is that if* $\Sigma_\varepsilon = I_K$, *since* $\varepsilon_t = Au_t$, *the matrix* A *has to be normalized so that* $A\Sigma_u A' = \Sigma_\varepsilon = I_K$ *is obtained. However, we will see in the next section, that in the analogy with spatial models we need* A *to have main diagonal set to unity. This is another normalization of the system. In other words, if we set* $\Sigma_\varepsilon = I_K$ *we can no longer have the diagonal elements of* A *equal to unity without loss of generality.*

Summing up, in this paper, I follow the orthogonality restriction in taking Σ_ε a diagonal matrix. Based on the above discussion, I do not normalize $\Sigma_\varepsilon = I_K$; instead, I normalize the system by setting the main diagonal elements of A to unity. Regarding

[10] For a thorough discussion on identification, and, in particular, on global identification see Rubio-Ramírez, Waggoner, and Zha (2010). See also Kilian (2013) for an additional discussion and examples in Macroeconomics.

the $K(K-1)/2$ identification restrictions, I choose in this paper to always impose them as exclusion restrictions. A specific set of at $K(K-1)/2$ coefficients of A are constrained to be zero.

These zero restrictions are often founded on some economic theory relevant to the problem at hand. However, this has led to an extensive discussion in the SVAR literature of the subjectivity of these restrictions (see, e.g., Kilian (2013, p.520)). In this paper, I do not take a stand on this issue and simply require the order condition for identification to be satisfied. In the Implementation section, however, I explore a data-driven approach to derive the zero restrictions as a means to avoid the aforementioned subjectivity issue. This is one possible solution, but the reader is left free to choose an alternative strategy.

2.3 A Minimum Distance approach to spatial modeling

2.3.1 SVAR and spatial modeling

The previous section introduced one of the main models discussed in this paper, the SVAR model in (2.2). This section presents the second model used throughout this paper, a "time series" version of a spatial model. The "time series" version of the spatial model considered here corresponds to a time series version of the Spatial Autoregressive model (SAR), hereafter T-SAR model.[11] The T-SAR model over time $t = 1, ..., T$ is given by:

[11] The results could be extended to alternative spatial models like the Mixed regressive SAR model (MSAR) that allows for exogenous covariates. In this case we would have to consider an extended SVAR model as well.

$$Y_t = \rho W Y_t + \Gamma_1 Y_{t-1} + \cdots + \Gamma_p Y_{t-p} + \varepsilon_t \tag{2.5}$$

where ρ is a scalar coefficient, W is a $K \times K$ row-standardized matrix of spatial weights with zero elements on its main diagonal, Γ_i, $i = 1, ..., p$, is a $K \times K$ matrix of coefficients, and ε_t is a K-dimensional noise term with $\varepsilon_t \sim (0, \Sigma_\varepsilon)$ and Σ_ε diagonal. In this model, the network matrix is given by W as its k^{th} row captures the contemporaneous effect of $y_{t,j}$, $\forall j \neq k$, on $y_{t,k}$. The parameter ρ captures how strong this network dependence is.

The next proposition makes use of the following terminology. We refer to the rows of G that have all their elements equal to zero as "zero rows" and those that have both zero and non-zero elements as "mixed rows." In the A-model considered in this paper, the main diagonal of G is always zero. Hence, all rows of G will always have at least one zero element. Other exact zeros in G correspond to the zero restrictions needed for identification. It can be shown that the T-SAR model is a particular case of the SVAR model. This is stated in the proposition below.

Proposition 2.2. *Assume the SVAR is identified. Let l be the number of zero rows of G, with $0 \leq l \leq K-1$. Let P be a $K \times K$ permutation matrix that reorders the rows of G such that the first l rows correspond to the l zero rows of G. The T-SAR model (2.5) is a constrained SVAR model (2.2) with $(K-1-l)$ independent linear restrictions on G given by:*

$$RPG\iota_K = 0 \tag{2.6}$$

where

$$\underset{(K-1)\times K}{R} = \left[\begin{array}{c|c} I_l & O_{l\times(K-l)} \\ \hline O_{(K-l-1)\times l} & \mathcal{R}_{(K-l-1)\times(K-l)} \end{array}\right] \tag{2.7}$$

with

$$\mathcal{R} = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ & & & \cdots\cdots\cdots & & & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \tag{2.8}$$

The proof of this proposition is presented in Appendix 2.A. Notice that if all the rows of G have at least one non-zero element, i.e., $l = 0$, then there is no need for a permutation matrix. Moreover, the restriction matrix in Proposition 2.2 simplifies to $R = \mathcal{R}$.

The intuition of the proof of Proposition 2.2 is discussed in what follows. First, notice that both models take the following form:

$$Y_t = C_0 Y_t + C_1 Y_{t-1} + \cdots + C_p Y_{t-p} + \varepsilon_t$$

where C_i, $i = 0, 1, ..., p$, is a $K \times K$ matrix of coefficients, C_0 has zero elements on its main diagonal, and ε_t is a serially uncorrelated error term with $\varepsilon_t \sim (0, \Sigma_\varepsilon)$.

Based on this observation, and the assumption that the SVAR models is identified, it follows that since Γ_i and B_i are unrestricted for $i = 1, ..., p$, the difference involves the matrices ρW and G. For simplicity of exposition, consider first the case in which all the rows of G have at least one non-zero element. Since W is a row-standardized matrix, each of its rows sums to one, and each row of ρW sums to ρ. Therefore, since

the restrictions on G amount to constraining the rows of G to add up to the same quantity, the result follows. Consider next the case where some rows of G have all their elements equal to zero, i.e., $l > 0$, which becomes a trickier case. The key is to isolate all the zero rows from the rest and impose the restrictions on the submatrix of G that only includes its mixed rows. The former task is accomplished by using the permutation matrix P, while the latter task is accomplished by using the extended matrix of restrictions R instead of $\mathcal{R}$. Intuitively, given that some rows of G have all their elements equal to zero, the sum of these elements is zero as well while the restrictions have to be imposed only on the coefficients that need to be estimated. Finally, notice that in the extreme case where $l = K - 1$ there is only one row with some non-zero elements; therefore showing the result from Proposition 2.2 becomes trivial.

2.3.2 A Minimum Distance estimator

As stressed in the introduction, data on financial networks is either confidential or non-existent, leading researchers to adopt ad-hoc network matrices when it comes to spatial modeling. This motivates the methodology in this section. I propose a Minimum Distance Estimation (MDE) procedure to estimate ρ and W from the estimate of the SVAR network matrix G. In addition, I propose in the next section a test for the restrictions that the spatial model imposes on the SVAR model.

Consider again the SVAR(p) model in (2.1), and its reduced form from (2.3). We want to estimate this model subject to the constraints in (2.6) for observations Y_t, $t = 1, ..., T$, and given initial values $Y_0, ..., Y_{1-p}$. The standard approach would be to estimate it by Maximum Likelihood (ML). This involves first estimating the reduced form VAR to obtain a consistent estimator of Σ_u, and then computing the concentrated log-likelihood function to maximize it subject to the constraints. More

specifically, the constrained log-likelihood to be maximized is

$$\begin{cases} \mathcal{L}(A,B) = \dfrac{-TK}{2}\ln(2\pi) + \dfrac{T}{2}\ln|A|^2 - \dfrac{T}{2}\ln|B|^2 - \dfrac{T}{2}trace\left\{\left(AB'^{-1}B^{-1}A'\right)\hat{\Sigma}_u\right\} \\ s.t. \quad RPA\iota_{_K} = 0 \end{cases} \tag{2.9}$$

where $\hat{\Sigma}_u = T^{-1}\sum_{t=1}^{T}\hat{u}_t\hat{u}_t'$, $\hat{u}_t = Y_t - \sum_{i=1}^{p}\hat{A}_i Y_{t-i}$, the $\hat{A}_i$ are the estimates of the reduced form coefficient matrices A_i, $i = 1, ..., p$, and $\varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$ with $\Sigma_\varepsilon = BB'$ and B diagonal matrix. Notice that since $A = I_K - G$ and $RPI_K\iota_{_K} = 0$ always, it is equivalent to state the constraint in terms of A instead of G.

However, this procedure leads to a non-linear system of equations in terms of A and B to be maximized subject to the constraints. This approach has the disadvantage that it requires numerical methods, and, therefore, it has no closed form solution. A closed form solution is useful to better understand the dependence of the results on the restrictions imposed on the SVAR model, and the relationship between the constrained and unconstrained parameters. It also has the clear advantage that the solution is exact. Therefore, I propose an alternative approach based on minimum distance estimation.

The MDE approach is simpler in that the constrained estimator is obtained from a quadratic form based on the unconstrained estimator. It allows us to replace the initial objective function given by the constrained MLE problem, which can be rather complex, by a quadratic function (in the parameters) which is usually easier to optimize. The constrained estimator appears then as an estimator in two stages: in the first stage, we determine the unconstrained estimator; then, in the second stage, we solve the MDE optimization problem to obtain the constrained estimator. The main advantage of this method is that it delivers closed form solutions for the constrained estimator as well as for the test statistic. A closed form solution is useful to better understand the relationship between the constrained and unconstrained parameters. In addition, it has the nice property that it provides estimators asymptotically equiv-

alent to constrained Maximum Likelihood, so there is no efficiency cost from using the MDE approach. An extensive discussion and a proof of this last result can be found in Gourieroux and Monfort (1989a,b, p.79 and p.383 respectively)

Before stating the MDE problem, a few issues need to be addressed. First, notice that since $A = I_K - G$, it is equivalent to apply minimum distance estimation on G instead of A. The relationship between the T-SAR and SVAR models is rather based on G. Hence, the MDE problem will be based on G as well. Second, the coefficients of the G matrix are not asymptotically degenerate. However, we have some linear restrictions for identification purposes that introduce some degenerate properties. These linear restrictions are zero restrictions. To delete the singularities it is necessary and sufficient to delete the zero coefficients from G. Third, restrictions in Proposition 2.2 have to be redefined accordingly. This will be done through a matrix H. A detailed procedure to adapt the MDE setup to account for these issues is discussed in what follows.

Let G be the $K \times K$ matrix of parameters to estimate, and let $\hat{G}$ and $\hat{G}^c$ be the unconstrained and constrained SVAR estimator of G respectively. Denote $\text{vec}(\hat{G}_{nz})$ the vector where I have erased all the zeros of $\text{vec}(\hat{G})$, with nz the number of non-zero elements in G. I delete the zeros that are identically equal to zero due to the restrictions. Then, $\text{vec}(\hat{G})$ and $\hat{V} = \widehat{Var}[\text{vec}(\hat{G})]$ have to be replaced in the MDE problem by $\text{vec}(\hat{G}_{nz})$ and $\hat{V}_{nz} = \widehat{Var}[\text{vec}(\hat{G}_{nz})]$ respectively. Therefore, the MDE program is given by:

$$\begin{cases} \underset{\text{vec}(G_{nz})}{\text{Min}} & \Big(\text{vec}(\hat{G}_{nz}) - \text{vec}(G_{nz})\Big)' \hat{V}_{nz}^{-1} \Big(\text{vec}(\hat{G}_{nz}) - \text{vec}(G_{nz})\Big) \\ s.t. & H\,\text{vec}(G_{nz}) = 0 \end{cases} \tag{2.10}$$

where $\hat{V}_{nz}$ is an estimator of the asymptotic variance-covariance matrix $V_{nz} = \text{AVar}[\text{vec}(\hat{G}_{nz})]$, with AVar standing for the asymptotic variance. It only remains to

define the matrix of constraints H accordingly.

In our case, G can have either of two types of rows: "zero rows" and "mixed rows" as defined in Section 2.3.1. Let nzr denote the number of mixed rows, where the notation "nzr" stands for "not zero rows." Notice that the number of non-zero elements of G is not equal to the number of mixed rows times the number of columns of G, i.e., $nz \neq nzr * K$. The reason is that there are also zero elements coming from mixed rows. The following example illustrates these concepts:

Example 2.1.

$$G = \begin{bmatrix} 0 & 0 & 0 \\ g_{21} & 0 & g_{23} \\ g_{31} & g_{32} & 0 \end{bmatrix} \begin{array}{l} \longrightarrow \textit{zero row} \\ \left.\begin{array}{l} \\ \end{array}\right\} \textit{mixed rows} \end{array}$$

where $K = 3$, $nz = 4$, $nzr = 2$.

The matrix H of constraints imposed on $\text{vec}(G_{nz})$ can be constructed according to algorithm 2.2 below. Mathematically, the resulting H is given in the following theorem:

Theorem 2.1. *(Matrix of constraints H) Restrictions (2.6) that characterize the T-SAR as a constrained SVAR are equivalent to:*

$$H \,\text{vec}(G_{nz}) = 0$$

with

$$\underset{(nzr-1)\times nz}{H} = \left\{ P_s \left[(\iota_K' \otimes \mathcal{R}_{nzr})' \odot \left(\text{vec}(G_{nzr}^*) \iota_{(nzr-1)}' \right) \right] \right\}' \tag{2.11}$$

where $\text{vec}(G_{nz})$ *is* $\text{vec}(G)$ *with non-zero elements erased,* G_{nzr} *is* G *with zero rows erased,* G_{nzr}^* *is* G_{nzr} *with its non-zero elements replaced by ones, and* P_s *is a* $nz \times$

$(nzr * K)$ *selection matrix that selects non zero rows of a given matrix or vector. If the* i^{th} *element of* $\text{vec}(G^*_{nzr})$ *is different from zero, then the* i^{th} *column of* P_s *corresponds to the* i^{th} *canonical vector; otherwise the* i^{th} *column of* P_s *is a column of zeros.*

Algorithm 2.2 (Algorithm to construct H)**.**

1. *Let* G_{nzr} *be equal to* G *but with zero rows deleted, and let* $\mathcal{R}_{nzr}$ *be built as* $\mathcal{R}$ *from equation* (2.8) *but with dimensions* $(nzr-1)\times nzr$*. Notice that* G_{nzr} *will still contain the zeros coming from the mixed rows. Compute* $H_0 = \iota'_K \otimes \mathcal{R}_{nzr}$*.*[a]

2. *Replace non-zero elements of* G_{nzr} *by ones, and denote this matrix by* G^*_{nzr}*. The matrix* G^*_{nzr} *simply tracks where zero and non-zero elements of* G_{nzr} *are located.*

3. *Multiply each column of* H'_0 *by* $\text{vec}(G^*_{nzr})$ *element wise (i.e., Hadamard product). Call this object* $\mathcal{H}$*. This will pin down rows in* H'_0 *that should be deleted as they correspond to zero elements in* $\text{vec}(G^*_{nzr})$*.*

4. *Delete the zero rows in* $\mathcal{H}$*.*

5. *Finally,* $H = \mathcal{H}'$ *and has dimensions* $(nzr-1)\times nz$*.*

[a] For computational ease, when coding this algorithm G_{nzr} can be replaced by $\hat{G}_{nzr}$.

The solution of the MDE program is presented in the next proposition. Let $\text{vec}(\hat{G}^c_{nz})$ be the minimum distance estimator of $\text{vec}(G_{nz})$ subject to the constraint $H\,\text{vec}(G_{nz}) = 0$, and let λ be the Lagrange multiplier of the MDE problem. Then, $\text{vec}(\hat{G}^c_{nz})$ can be derived from the Lagrangian by minimizing with respect to $\text{vec}(G_{nz})$

and λ, and then solving for $\text{vec}(G_{nz})$. The solution to this minimization program is given in the following proposition:

Proposition 2.3. *(Minimum Distance Estimator) The Minimum Distance Estimator* $\text{vec}(\hat{G}^c_{nz})$ *of the minimization problem* (2.10)*, is given by*

$$\text{vec}(\hat{G}^c_{nz}) = \text{vec}(\hat{G}_{nz}) - \hat{V}_{nz}\, H' \left\{ H\, \hat{V}_{nz}\, H' \right\}^{-1} H\, \text{vec}(\hat{G}_{nz}) \tag{2.12}$$

where H *is defined as in equation* (2.11)*, and* $\hat{V}_{nz} = \widehat{Var}[\text{vec}(\hat{G}_{nz})]$*. The estimator of the asymptotic variance-covariance matrix of* $\text{vec}(\hat{G}^c_{nz})$ *is given by*

$$\hat{V}^c_{nz} = \hat{V}_{nz} - \hat{V}_{nz} H' \left\{ H\, \hat{V}_{nz}\, H' \right\}^{-1} H\, \hat{V}_{nz} \tag{2.13}$$

The proof is given in Appendix 2.A.

Finally, it is straightforward to obtain the estimate of the spatial weights or network matrix $\hat{W}$ from $\hat{G}^c$. Indeed, it only requires to recall that $\hat{G}^c = \hat{\rho}\hat{W}$ and that the sum of the elements of any mixed row of $\hat{G}^c$ is equal to $\hat{\rho}$. Furthermore, the standard error of $\hat{\rho}$ can be computed from the variance-covariance matrix of $\hat{G}^c$.

Corollary 2.1. *($\hat{W}$ and $\hat{\rho}$) Let* $\mathcal{MP}$ *be the set of mixed rows of* $\hat{G}^c$*, and denote* $[\hat{G}^c]_j$ *the* j^{th} *row of* $\hat{G}^c$*. Then, for any* $j \in \mathcal{MP}$

$$\hat{\rho} = [\hat{G}^c]_j\, \iota_K, \quad \widehat{SE}_{\hat{\rho}} = \sqrt{\widehat{Var}\left([\hat{G}^c]_j\, \iota_K \right)} = \sqrt{\sum_{i=1}^{K} \widehat{Var}\left(g^c_{ji} \right) + \sum_{i \neq i'} \widehat{Cov}\left(\hat{g}^c_{ji}, \hat{g}^c_{ji'} \right)}$$

and

$$\hat{W} = \hat{\rho}^{-1} \hat{G}^c$$

The proof is omitted since Corollary 2.1 holds by construction of $\hat{G}^c$.

2.3.3 Test for a spatial model

We want to test whether the restrictions imposed by the T-SAR spatial model hold. For this purpose, we let the null hypothesis correspond to the spatial model (restricted model), and the alternative hypothesis be the SVAR(p) model (unrestricted model). That is, we are interested in testing:

$$H_0: \text{ T-SAR} \quad \text{[restricted model]}$$

$$H_1: \text{ SVAR}(p) \quad \text{[unrestricted model]}$$

The MDE approach provides a natural Wald-type test statistic; this is stated in the next proposition.

Proposition 2.4. *(Minimum Distance Test Statistic) Consider the MDE problem in (2.10) and the solution given in Proposition 2.3. The test statistic to test the constraints* $H \operatorname{vec}(G)_{nz} = 0$ *imposed on* $\operatorname{vec}(G)_{nz}$*, has a closed form representation given by*

$$\mathcal{SD} = T\left(\operatorname{vec}(\hat{G}_{nz}) - \operatorname{vec}(\hat{G}^c_{nz})\right)' V_{nz}^{-1}\left(\operatorname{vec}(\hat{G}_{nz}) - \operatorname{vec}(\hat{G}^c_{nz})\right) \xrightarrow{d} \chi^2_{nzr-1} \tag{2.14}$$

where $V_{nz} = AVar[\operatorname{vec}(\hat{G}_{nz})]$.

The number of restrictions is $nzr-1$, that is the number of mixed rows minus one, since the restrictions impose that all mixed rows of G add up to the same quantity. In addition, notice that V_{nz}^{-1} is a generalized inverse of $AVar[\operatorname{vec}(\hat{G}_{nz}) - \operatorname{vec}(\hat{G}_{nz})^c]$. The proof is given in Appendix 2.A.

Since the test statistic takes the quadratic form $m'AVar(m)^- m$, with $AVar(m)^-$ a generalized inverse of $AVar(m)$, it can be interpreted as a standardized distance measure. As stated in Ullah (1996, p.137), "Many of the currently used econometric tests, such as the likelihood ratio, the score and Wald tests, can in fact be shown to be

in terms of appropriate distance measures." A Wald-type test can be interpreted as a measure of the distance between the unrestricted and restricted value under the null hypothesis. Given that the null hypothesis in this paper corresponds to the spatial model, the test statistic can be interpreted as a measure of the distance to a spatial model. In other words, it provides a measure of how far the spatial model is from the SVAR model, or how restrictive the parametrization imposed by a spatial model is. The larger the value of the $\mathcal{SD}$ statistic, the bigger the distance is and the more restrictive the spatial model is.

2.4 Implementation

2.4.1 Identification via machine learning

Implementation of the MDE procedure is simple given identification of the SVAR. However, the issue of identification of SVAR models is not an easy problem to address. In the time series literature, a popular approach has been the use of recursive systems as they can be identified through a Cholesky decomposition of the reduced form VAR residuals; yet it is well known in this literature that identification of SVAR models through a Cholesky decomposition carries an ordering problem (see e.g., Cooley and Leroy (1985), Stock and Watson (2001), Kilian (2013)). Namely, a different order of the variables in the system leads to a different network pattern.

To circumvent the ordering issue, in this section, I examine the use of data-driven methods that deliver a data driven ordering in a recursive overidentified system. This solution, which was proposed in the SVAR literature to deal with the ad-hoc ordering, consists of inferring the causal structure from the data by means of machine-learning, specifically a graph-theoretic approach. The most popular algorithm in the machine learning literature is the PC-algorithm from the seminal work by Spirtes, Glymour, and Scheines (2000). This machine learning method has been widely used

in macroeconomics, psychology, and biostatistics, but it has not been explored much in the context of financial networks. An example in the financial networks literature is the work of Yang and Zhou (2013) that applies the PC-algorithm for the study of credit risk spillovers.

2.4.1.1 The general issue

Section 2.2.3.2 stressed that the orthogonality restriction plus the restriction on the main diagonal elements of A do not deliver identification of the SVAR model, as $K(K-1)/2$ additional restrictions are still needed. One solution adopted in the SVAR literature to address this identification problem is to recover the structural matrix of coefficients A through a triangular factorization, also known as LDL decomposition, of Σ_u given by $A^{-1}\Sigma_\varepsilon(A')^{-1}$. This decomposition facilitates identification of A and Σ_ε as it requires A to be lower triangular, therefore imposing the required number of restrictions for identification.

However, it is well known that making A lower triangular, as a result of the triangular factorization, is not without consequence. This imposes a recursive ordering of the K variables involved; which means that, depending on the variables ordering, different matrices A are obtained. As an illustration of the foregoing, consider the following example for $K = 3$:

Example 2.2. *Let $Y_t = (y_{1,t}, y_{2,t}, y_{3,t})$. Consider the SVAR model (2.2) with $p = 1$. The triangular factorization of Σ_u implies the following ordering:*

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ a^{21} & 1 & 0 \\ a^{31} & a^{32} & 1 \end{bmatrix} \quad \Rightarrow A = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix}$$

where a^{ij} represents the ij^{th} element of the inverse of A, and I have used that the

inverse of an invertible lower triangular matrix is also lower triangular. Recalling that $G = (I_K - A)$*, the SVAR model has the following causal ordering:*

$$y_{1,t} = b_{1,11}y_{1,t-1} + b_{1,12}y_{2,t-1} + b_{1,13}y_{3,t-1} + \varepsilon_{1,t}$$
$$y_{2,t} = -a_{21}y_{1,t} + b_{1,21}y_{1t-1} + b_{1,22}y_{2t-1} + b_{1,23}y_{3t-1} + \varepsilon_{2,t}$$
$$y_{3,t} = -a_{31}y_{1,t} - a_{32}y_{2,t} + b_{1,31}y_{1,t-1} + b_{1,32}y_{2,t-1} + b_{1,33}y_{3,t-1} + \varepsilon_{3,t}$$

In other words, the implied causal ordering is: $y_{1,t} \to y_{2,t}$*, and* $y_{1,t} \to y_{3,t} \leftarrow y_{2,t}$*. Now consider a different ordering of the variables in* Y_t*, say* $Y_t = (y_{3,t}, y_{2,t}, y_{1,t})$*, where I have interchanged* $y_{1,t}$ *and* $y_{3,t}$*. It is easy to see that the implied causal ordering now is* $y_{3,t} \to y_{2,t}$*, and* $y_{3,t} \to y_{1,t} \leftarrow y_{2,t}$*.*

The lower triangular factorization approach, which is a variant of the classical Cholesky decomposition, corresponds to a just-identified system. That is, each Cholesky ordering corresponds to a just-identified SVAR, and, as formerly shown, they are all observationally equivalent. However, as illustrated in the above example, different orderings translate into different network dynamics. In absence of additional information, this poses the problem that we cannot identify the network, which is precisely what we are interested in, from this decomposition alone.

2.4.1.2 A graph-theoretic approach to identification: The PC-algorithm

A solution that has been proposed in the SVAR literature to deal with the ad-hoc ordering consists of inferring the causal structure from the data by means of a graph-theoretic approach.[12] This idea of a data driven methodology was first promoted by Swanson and Granger (1997), and has been referred to as "Empirical Identification". Furthermore, it has been extensively discussed by Demiralp and Hoover (2003) and later by Moneta (2008), and Hoover, Demiralp, and Perez (2009) among

[12]The reader unfamiliar with graph theory is advised to read the Network Terminology section of Appendix 2.A.

others. Causal relationships, like the ones described by a SVAR, can be represented by a graph where causal variables are connected to their effects through arrows, as shown by Pearl (2000) and Spirtes, Glymour, and Scheines (2000). The graph of the data generating process can alternatively be represented by zero restrictions over the matrix A. Moreover, to reduce the complexity of a graph-theoretic approach the literature focus on the so called "Directed Acyclic Graphs" (DAGs) as these graphs represent recursive orderings and hence are easier to handle.

Graphical causal models are based on conditional independence analysis and are implemented sequentially on the estimated residuals from the reduced form VAR through a search algorithm. Once the contemporaneous causal structure is recovered, the estimation of the lagged autoregressive coefficients allows us to identify the complete SVAR model. Several search algorithms have been developed to recover the contemporaneous causal ordering. The algorithm discussed in this paper is the so-called PC-algorithm, since it is one of the most widely used among search algorithms.

The PC algorithm is a search algorithm that has been adopted in a variety of fields to build DAGs, which are a type of graphical models as stated in definition 2.7. The search procedure is comprised of a few steps that can be easily summarized as follows. The algorithm starts with a complete undirected graph; the maximum number of edges in an undirected graph without a self-loop (i.e., directed cycle) is $K(K-1)/2$ for K nodes. This leaves at least $K(K-1)/2$ zero restrictions on A which are needed to fulfill the order condition for identification. The algorithm then continues by recursively removing edges between vertices based on conditional independence tests of size α. This yields an undirected graph called the skeleton of the DAG as defined in 2.8. The last step consists of (partially) orienting the remaining edges to produce a (partial) DAG as a final output.[13] The exact algorithm and and

[13]The PC-algorithm runs in the worst case in exponential time, as a function of the number of nodes. However, if the true DAG is sparse, which is often a reasonable assumption, this reduces time to a polynomial runtime.

a simple example based on five nodes is provided in Appendix 2.A.

Remark 2.3. *The PC algorithm is a useful tool to uncover the underlying causal ordering. However, many times, the PC algorithm cannot determine the DAG uniquely but only the corresponding equivalence class of the DAG. An equivalence class contains DAGs that have the same conditional independence information. That is, they share the same skeleton and directed edges, but some of the edges remain undirected. Consequently, the resulting output of the PC algorithm is most commonly referred to as a Completed Partially Directed Acyclic graph (CPDAG) (see Spirtes et al. (2000)). In spite of this, the PC algorithm has proven to be useful in that it reveals at least partially the causal ordering.*

The PC algorithm is not assumption free. Indeed, the algorithm is based on two main assumptions that establish a link between causation and partial correlation as stated in Spirtes et al. (2000).[14,15] Let $\mathcal{G} = (V, E)$ be a graph consisting of a set of nodes or vertices $V = \{1, ..., K\}$ and a set of edges $E \subseteq V \times V$, i.e., the edge set is a subset of ordered pairs of distinct nodes. In our SVAR framework, the set of nodes corresponds to the components of a random vector $Y_t \in \mathbb{R}^K$. If there is a directed edge $i \to j$, node i is said to be a *parent* of node j, while node j is said to be a *descendant* of node i. Descendants can also be more than one vertex away. For instance, if there is a directed edge $i \to j \to l$, then l is also a descendant of i. Naturally, i is an ancestor of j if and only if j is a descendant of i.

Let $\mathbb{P}$ be a probability distribution on $\mathbb{R}^K$, and let $V \sim \mathbb{P}$. For each $i \in V$, denote $Parents(i)$ the set of parents of i and $Descendants(i)$ the set of descendants of i. Then, we can state the first assumption:

[14] An additional assumption that is in general (implicitly) assumed is causal sufficiency, which states that there are no omitted variables that cause two of the included variables. In other words, causal sufficiency assumes that there are no latent confounders of any two observed variables.

[15] Consistency of the PC-algorithm, and other search algorithms, has been addressed in particular in Spirtes et al. (2000) and Robins, Scheines, Spirtes, and Wasserman (2003) in the context of causal inference. See also Zhang and Spirtes (2002) and Uhler, Raskutti, Bühlmann, and Yu (2013) for a discussion on uniform consistency.

Assumption A1. *(Causal Markov Condition) The directed acyclic graph $\mathcal{G}$ over V and the probability distribution $\mathbb{P}(V)$ satisfies the Markov condition if and only if for every vertex $i \in V$,*

$$i \perp\!\!\!\perp V \setminus \{Descendants(i) \cup Parents(i)\} \mid Parents(i)$$

where $V \setminus C = \{j \in V \mid j \notin C\}$.

In words, $\mathcal{G}$ satisfies the causal Markov condition if and only if each variable is conditionally independent of its non-descendants given its parent variables. It implies that we can write probabilities of variables by conditioning just on each variable's parents (i.e., we do not have to condition on grandparents, great grandparents, aunts, uncles or children). This assumption is also known as the local Markov property. As an example, assume we have data on volatility of main stock return indexes of countries A, B, C, D, and E. Also assume that $A \to B \to C \to D$, and $A \to E$. In this case, to test independence between A and D it is enough to condition on B (i.e., parent), we do not need to condition on C (i.e., grandparent) nor E (i.e., descendant).

The second assumption requires the use of d-separation concept, which can be formally defined as:

Definition 2.1. *(d-separation) Let $i, j \in V$, $i \neq j$, a set $S \subseteq V$ is said to d-separate (directionally separate) i from j if and only if S blocks every path from node i to j.*

Notice that d-separation implies conditional independence: if S blocks all paths from i to j, then $i \perp\!\!\!\perp j \mid S$. However, the converse is not necessarily true. To reverse this, and conclude that if $i \perp\!\!\!\perp j \mid S$ then it must be that S d-separates i and j, an additional assumption called the "Faithfulness condition" is necessary. This assumption states that a distribution $\mathbb{P}$ is faithful to a DAG $\mathcal{G}$ if no conditional independence relations other than the ones entailed by the Markov property are present. That is,

Assumption A2. *(Faithfulness condition)* $\mathbb{P}$ *is faithful with respect to* $\mathcal{G}$ *if for any* $i, j \in V$*, with* $i \neq j$*, and any set* $S \subseteq V$*,*

$$i \perp\!\!\!\perp j \mid \{r; r \in S\} \iff i \text{ and } j \text{ are d-separated by the set } S$$

In other words, a distribution is faithful to a DAG if all the conditional independence relations for $\mathbb{P}$ can be derived from d-separation. Intuitively, the faithfulness condition states that if we see zero correlation between two variables, the reason we see it is because there is no edge between these variables and not cancellation of structural parameters. Consider again the example from the introduction where we have data on volatility of main stock return indexes of countries A, B, and C. Assume A affects B directly (i.e., $A \to B$), and A affects B indirectly through C (i.e., $A \to C \to B$). Moreover, assume the direct effect increases volatility on B while the indirect effect decreases volatility on B. If the two effects happen to be of the exact same magnitude, given that they are of opposite sign, they would cancel each other out. As a result, we would obtain that A is independent of B contradicting the graph. This is an example of a violation of faithfulness. Notice that this type of violation is unlikely to happen, at least in the financial networks context, since it requires effects to exactly offset one another.

Remark 2.4. *In the multivariate Gaussian case, conditional independence can be inferred from partial correlations. Therefore, the normality assumption allow us to assess conditional independence by conducting tests based on partial correlations instead. Therefore, in the multivariate Gaussian case the Markov and faithfulness assumptions can be equally stated in terms of partial correlations.*

2.4.1.3 Implementing the algorithm

The implementation of the PC Algorithm is overall simple. In this paper, I use the R package *pcalg* of Kalisch, Mächler, Colombo, Maathuis, and Bühlmann (2012). The "original" PC-algorithm is known to be order-dependent, in the sense that the output depends on the order in which the variables are given (Colombo and Maathuis (2014, p.1)). As a result, in this paper I use the modifications to the PC-algorithm proposed by Colombo and Maathuis (2014), which the authors show produce a fully order independent output. This has been incorporated in the pcalg package through the majority rule together with the solve conflict options.[16,17] The input of the PC-Algorithm is the estimate of the covariance matrix of u_t, $\hat{\Sigma}_u$, obtained from the reduced form VAR.[18] Notice that the search algorithm employs a statistic based on the residuals and not the variable Y_t itself. The main reason for this choice of input is to filter Y_t from its VAR dynamics.

As explained in Section 2.4.1.2, the output from imputing $\hat{\Sigma}_u$ into the PC-Algorithm is a CPDAG. Sometimes the algorithm returns a fully directed graph, in which case we refer to the output as a DAG. Many times, however, some edges are not oriented (bidirected edges in the graph) which results in a CPDAG. If the output is indeed a CPDAG, this implies a set of possible DAGs. A possible solution, is to select a DAG from this set based on some criterion and/or prior information. In this paper, I propose to select a DAG based on a Maximum Likelihood criterion, this is explained in the simulations section.

For the remainder of this subsection, assume that we have a DAG either directly

[16] Colombo and Maathuis (2014) introduced a less strict version of the conservative PC-algorithm option for the v-structures called majority rule. In this case, the triple $a - b - c$ is marked as "ambiguous" if and only if b is in exactly 50 percent of such separating sets or no separating set was found. If b is in less than 50 percent of the separating sets it is set as a v-structure, and if in more than 50 percent it is set as a non v-structure (for more details see Colombo and Maathuis (2014)).

[17] Sampling errors (or hidden variables) can lead to conflicting information about edge directions. The solve conflict option introduces a modification to the PC-algorithm to address this problem.

[18] Alternatively, it is also possible to use the residuals directly; the results do not change based on this choice.

from the PC-algorithm or from choosing one from the set of possible DAGs. This DAG shows graphically the causal structure among the K variables of interest, and tells us which entries of G are set to exact zeros. To recover this causal order in a matrix form, we need to compute the adjacency matrix of the graph. This matrix will only have zeros and ones. The matrix G with zero and one entries is equal to the transpose of the adjacency matrix from the PC-algorithm. For instance, if we observe $j \to i$, then the entry (i, j) in G is non-zero while the entry (j, i) is set to zero. Then, once we know the causal structure of the contemporaneous variables, we know which entries of $A = I_K - G$ are non-zero. Based on this information, we can estimate the non-zero entries of the overidentified system through Maximum Likelihood. This will provide an estimate of the weight of each connection discovered by the search algorithm.

2.4.2 Estimation of overidentified SVAR models

The SVAR model in this paper is estimated by Maximum likelihood (ML).[19] Assume we estimated the reduced form model to obtain $\hat{A}_1, ..., \hat{A}_p$ and $\hat{\Sigma}_u$, and we applied the PC-algorithm to pin down the zero and non-zero entries of A. For simplicity of exposition, consider that the PC-algorithm delivers a DAG; that is, there are no undirected edges and the DAG is unique. Based on this result, the matrix A is estimated imposing the zero restrictions implied by the PC-algorithm. Assume next that $\varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$, with $\Sigma_\varepsilon = BB'$ and B a $K \times K$ matrix. Then, we can substitute $\hat{A}_1, ..., \hat{A}_p$, in the log-likelihood function and maximize it with respect to A and B.

The steps to estimate the SVAR can be summarized as follows:

1. Estimate the reduced form coefficients $\hat{A}^p = (\hat{A}_1, ..., \hat{A}_p)$ and the covariance matrix of the residuals $\hat{\Sigma}_u$.

[19] See Lütkepohl (2007, Chapter 9, p.372) for further details.

2. Apply the PC-algorithm to retrieve the causal structure of the variables and, therefore, the zero restrictions in the matrix $A = I_K - G$.

3. Substitute A^p for $\hat{A}^p$ in the log-likelihood, rearrange, and maximize w.r.t. A and B, with $\Sigma_\varepsilon = BB'$. The concentrated log-likelihood to be maximized after rearranging is:

$$\mathcal{L}(A, B) = -\frac{KT}{2}\ln(2\pi) + \frac{T}{2}\ln|A|^2 - \frac{T}{2}\ln|B|^2 - \frac{T}{2}trace\left\{\left(A'B'^{-1}B^{-1}A\right)\hat{\Sigma}_u\right\} \tag{2.15}$$

where $\hat{\Sigma}_u = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_t\hat{u}_t'$, and $\hat{u}_t = Y_t - \sum_{l=1}^{p}\hat{A}_l Y_{t-l}$.

This procedure leads to a non-linear system of equation in terms of A and B to be maximized subject to the identifying restrictions.[20] Maximization of this function is done by means of numerical methods.

2.4.3 Simulations

For the purpose of assessing how well the PC-algorithm pins down the "zeros" (no connection) and "ones" (connections) in the network matrix G, I conduct several simulation exercises. I work with G instead of the restricted network matrix G^c from the MDE procedure, since their zeros and ones structure coincide by construction. This exercise is based on K individuals, i.e., $Y_t = [y_{1,t}, ..., y_{K,t}]$, and the data is simulated from the following SVAR process:

$$AY_t = \sum_{l=1}^{p} B_l Y_{t-l} + \varepsilon_t$$

where Y_t is in deviation from its mean.

[20] We can easily extend this procedure to allow for a constant term. If the model includes a constant term, we need to replace A^p by $\Pi = [A_0\ A^p]'$, and define X_t as $X_t = [1\ Y_{t-1}', ..., Y_{t-p}']'$.

The matrices A, Σ_ε, and $B_1, .., B_p$ are calibrated to match features of financial data since we are interested in assessing the performance of the PC-algorithm in a financial network framework. I use data from the Application section to construct the Data Generating Process (DGP). The data consists of daily realized log-volatility of returns, for K countries' main stock indexes, from June 25, 2003 to March 1, 2007.[21] I apply the methodology to this data and take the resulting matrices $\hat{A}$, and $\hat{B}_1, ..., \hat{B}_p$ as the true DGP for simulating samples from the SVAR process. The lag order (i.e., p) of the SVAR is chosen by means of a Bayesian Information Criterion (BIC) in the reduced form model, with $p_{max} = 10$. Finally, Y_t is simulated using the described specification for $T = 1000$ periods, and $s = 500$ replications. In addition, for each simulation exercise, the Monte Carlo simulation study compares the performance of the PC-algorithm across values of its tuning parameter α, i.e., significance level of the independence test. This is is done to provide guidelines in terms of the choice of α.

The different scenarios considered are: $\alpha = \{0.2, 0.15, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ and $K = \{8, 16\}$. In what follows, I focus on results for $K = 16$ as this is the same K as in the application; results for $K = 8$ are available in Appendix 2.B. For each of these configurations, and each simulation, I compute the following measures to assess performance: Accuracy (ACC), Structural Hamming Distance (SHD), True Positive Rate (TPR) and Specificity (SPC). For a given configuration, the values reported below correspond to an average of each measure across simulations (i.e., mean ACC, mean TPR, etc.). These measures are standard in the DAG and network literature; definitions and interpretations are provided in Appendix 2.A for the reader unfamiliar with them.

The simulation exercise presented here explores the performance of the PC-

[21] This corresponds to the so called "Pre-crises" period in the Application section. This period has been chosen because it is a period of relatively normal activity in the stock market, though other periods could be used instead.

algorithm implemented with the majority rule together with the solve conflict options (MajRSC option hereafter). As previously mentioned, the combination of these options, which was proposed by Colombo and Maathuis (2014), renders the algorithm fully order independent. Figure 2.1 displays the simulation results.[22]

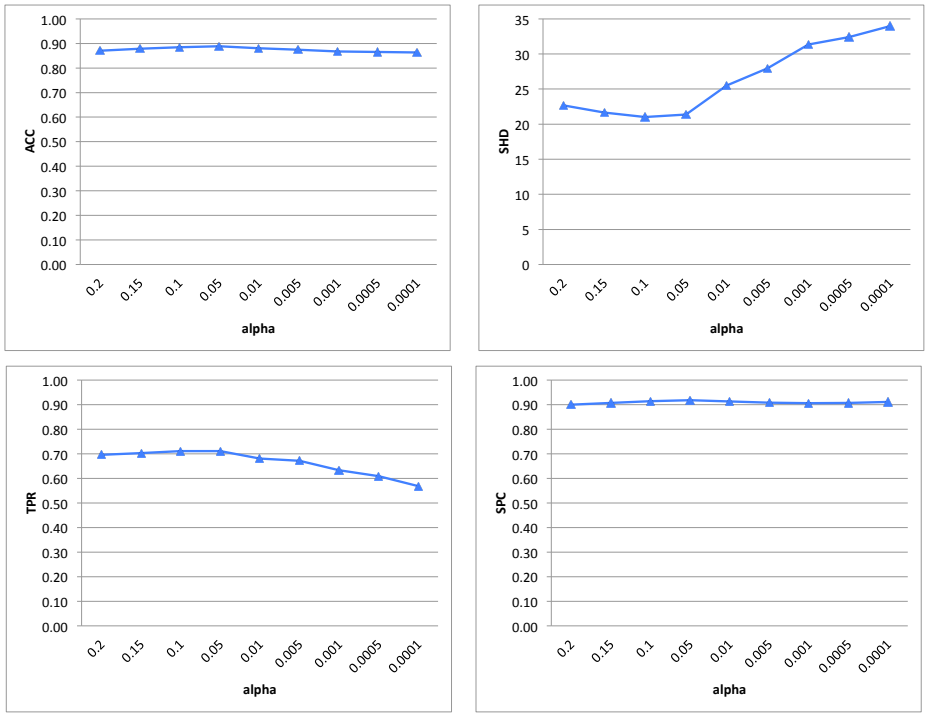


Figure 2.1: Comparison of MajRSC option to implement the PC-algorithm across values of α, $s = 500$. DGP generated using MajRSC option, $\alpha = 5\%$, $K = 16$, and $p = 1$. Mean ACC, TPR, SPC, and SHD reported.

The MajRSC option performs the best across the four measures for a value of α equal to 0.05, although results for 0.1 are very similar. For $\alpha = 0.05$ the value of ACC is quite high (approx. 0.9) meaning that, on average, 90% of the zeros and ones observed in a simulation are also present in the DGP. This suggests that, overall, the algorithm performs well. If we look at its performance for detecting zeros and ones separately we observe some disparity. On average, the value of SPC is still high (approx. 0.9), while the TPR is a bit lower (i.e., 70%). This means that the PC-algorithm is better at correctly picking up zeros (i.e., no connection) between

[22]The DGP was calibrated using the MajRSC option with $\alpha = 5\%$. Other options and values of α are explored in Appendix 2.B. Conclusions are similar.

individuals. Nevertheless, a TPR of 70% is quite good, as it means that on average the algorithm correctly recovers 70% of the links (i.e., connections) present in the DGP. To assess whether it is always the case that the best α coincides with the one chosen to calibrate the DGP, I also constructed DGPs with $\alpha = \{0.10, 0.0001\}$ and repeated the simulation exercise. The results are presented in Appendix 2.B, and they also suggest the use of α equal to 0.05 setting aside the former concern.[23] Other simulation exercises are presented in Appendix 2.B for the interested reader.

Finally, I address what to do in cases where the PC-algorithm returns some edges that are not oriented. As explained in Section 2.4.1.3, many times, some edges are not oriented (bidirected edges in the graph) which results in a CPDAG and gives rise to a set of possible DAGs. The simplest option is to leave these edges as bidirected as long as the order condition for identification is still satisfied. An alternative, is to direct these edges based on some additional criterion and assess the performance of this choice through simulations. In this paper, I explore the use of a maximum likelihood criterion. First I generate all possible DAGs from the partially directed graph produced by the PC-algorithm. Notice that some of the possible combinations of edges orientation may deliver a graph that contains cycles. To restrict the possibilities to acyclical graphs I check for cycles by applying the following lemma (see, e.g., Hearon (1972) or Saaty and Busacker (1965) for a proof of this lemma):

Lemma 2.1. *Let $\mathcal{A}$ be an adjacency matrix for a directed graph $\mathcal{G} = \{V, E\}$, such that $\mathcal{A}_{ij} = 1$ if $V_i \to V_j \in E$, and $\mathcal{A}_{ij} = 0$ otherwise. $\mathcal{G}$ has no directed cycles if and only if $\mathcal{A}$ is nilpotent. That is, $\mathcal{G}$ is acyclic if and only if $\mathcal{A}^K = 0$.*

Once I single out the acyclic graphs, I estimate the model for each possible DAG and pick the one that maximizes the log-likelihood.

Of course, there is no free lunch: this approach is computationally intensive espe-

[23] Notice that the results in the Appendix also compare MajRSC with another option to implement the algorithm.

cially when the number of undirected edges goes beyond eight, which may occasionally be the case. Also, sometimes, there could be no acyclical graph and one has to choose a graph with a cycle, but this problem may as well arise in any other approach. To address the former issue, I apply the following rule: if the number of bidirected edges is above eight, then, I first orient the edges in excess of eight using a t-statistic ranking criterion. This criterion consists of orienting each of these bidirected edges in the direction of the more significant connection. In other words, it consists of deleting the edge that is less significant in the pair of edges that conforms the bidirected edge. For each pair of edges that conforms a bidirected edge I compute the difference of t-statistics. Then, I rank these differences from highest to lowest. Finally, assume that the number of bidirected edges in excess of eight is q, I orient the pairs with the q highest differences using the t-statistic ranking criterion. The remaining bidirected edges are oriented with the ML criterion.

The results of this approach are shown in Table 2.1. In this table, I compare the percentage of times a coefficient in the network is correctly recovered in a set of 100 simulations; this allows to assess improvements in terms of prediction. The top and bottom panels present this performance measure for before and after orienting bidirected edges with the ML criterion respectively. Non-zero coefficients are highlighted in blue. For instance, in the top panel 72% of the time the PC-algorithm correctly recovers the coefficient for (Y_1, Y_2), while after redirecting bidirected edges this percentage increases to 82%. Our goal of orienting the bidirected edges is to increase the probability of correctly detecting a connection when there is one. From the comparison of the two panels in Table 2.1 it is clear that the ML criterion helps to achieve that goal as all the values in blue are higher in the bottom panel. Therefore, I apply this criterion in the application section.

Table 2.1: Percentage of times each coefficient in the network is correctly recovered in $s = 100$ simulations. Top panel corresponds to the network matrix produced by the PC-algorithm. In the bottom panel bidirected edges had been oriented using ML criterion. DGP and simulations use MajRSC option, $\alpha = 5\%$, and $K = 16$. $p = 1$ in the DGP.

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}
Y_1	100	72	56	66	44	47	55	100	100	100	99	100	72	100	100	100
Y_2	88	100	100	40	85	100	51	93	100	100	72	100	100	100	100	100
Y_3	77	100	100	63	74	66	100	81	100	100	100	100	100	100	100	100
Y_4	49	45	33	100	45	52	33	80	100	99	100	100	100	100	100	100
Y_5	72	51	37	37	100	62	61	100	100	100	100	100	100	100	100	100
Y_6	89	100	36	31	32	100	35	99	100	100	99	100	100	100	99	100
Y_7	74	56	100	57	33	68	100	100	100	100	99	100	100	100	100	100
Y_8	99	95	82	81	100	98	99	100	78	100	100	88	100	87	100	100
Y_9	100	99	100	99	100	100	98	78	100	38	100	100	67	99	100	35
Y_{10}	99	100	100	100	100	100	100	100	86	100	99	99	69	100	99	98
Y_{11}	100	91	100	100	100	100	99	100	100	99	100	100	100	65	70	99
Y_{12}	100	100	100	100	100	100	100	45	99	98	100	100	21	61	100	98
Y_{13}	74	100	100	100	100	100	100	100	53	29	100	51	100	99	99	91
Y_{14}	100	100	100	100	100	99	100	64	100	100	92	30	99	100	54	99
Y_{15}	100	100	100	100	100	99	100	100	100	99	95	100	100	88	100	97
Y_{16}	100	100	100	100	100	100	100	100	81	98	99	98	29	98	98	100

(a) Network from PC-algorithm

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}
Y_1	100	82	71	75	62	42	67	100	100	100	99	100	65	100	100	100
Y_2	77	100	100	26	64	100	36	95	100	100	87	100	100	100	100	100
Y_3	66	100	100	42	55	49	100	88	100	100	100	100	100	100	100	100
Y_4	36	50	53	100	57	38	51	85	100	99	100	100	100	100	100	100
Y_5	55	69	66	21	100	43	34	100	100	100	100	100	100	100	100	100
Y_6	92	100	58	48	53	100	56	99	100	100	99	100	100	100	99	100
Y_7	59	73	100	37	55	46	100	100	100	100	100	100	100	100	100	100
Y_8	99	95	82	79	100	98	99	100	84	100	100	85	100	91	100	100
Y_9	100	99	100	99	100	100	98	78	100	69	100	100	80	99	100	63
Y_{10}	99	100	100	100	100	100	100	100	66	100	100	100	38	100	100	98
Y_{11}	100	81	100	100	100	100	100	100	100	99	100	100	100	81	89	99
Y_{12}	100	100	100	100	100	100	100	97	99	98	100	100	49	49	100	99
Y_{13}	84	100	100	100	100	100	100	100	47	52	100	39	100	99	99	62
Y_{14}	100	100	100	100	100	99	100	63	100	100	81	56	100	100	71	99
Y_{15}	100	100	100	100	100	99	100	100	100	99	87	100	100	69	100	98
Y_{16}	100	100	100	100	100	100	100	100	57	100	100	100	65	99	99	100

(b) Network with no bidirected edges: ML approach

2.5 Measures of connectivity

In this section, I discuss several measures of connectivity to analyze the network W from the previous section. Since most of these measures are from the network literature, they rely heavily on network or graph theory concepts. All the terminology used in what follows is properly defined in the Network Terminology section of Appendix 2.A. The reader unfamiliar with graph theory is advised to read that section of the appendix before going through the current section.

2.5.1 Network preliminaries

Mathematically, a network of K individuals can be represented by a $K \times K$ adjacency matrix $\mathcal{A}$. In its most simple form, the adjacency matrix is basically a matrix of zeros (no connections) and ones (connections) such that entry ij equals 1, i.e., $[\mathcal{A}]_{ij} = 1$, if node i and node j are connected, otherwise $[\mathcal{A}]_{ij} = 0$. The adjacency matrix needs not be limited to have entries comprised of zeros and ones ("zero-one" type hereafter). Many times, weights are added to each connection to account for the strength of the connection, instead of considering each as equally important. Furthermore, in a general network context, the adjacency matrix need not be symmetric. This is particularly the case of directed networks where $[\mathcal{A}]_{ij} = 1$ reads as "i influences j", while $[\mathcal{A}]_{ji} = 1$ reads as "j influences i." Notice that, by definition, the adjacency matrix corresponds to the transpose of the network matrix depicted by W in the restricted SVAR (or by G in the unrestricted SVAR) from the previous section.

In this paper, I work with weighted adjacency matrices, unless otherwise stated, since the entries of W (or G) are not assumed to be zeros and ones.[24] Moreover, the adjacency matrix corresponding to W (or G) is not restricted to be symmetric. As a result, this adjacency matrix provides not only the strength of connections but also their direction, and whenever necessary either or both dimensions may be ignored for the sake of analysis.

2.5.2 The measures

In what follows, I summarize a subset of measures that potentially reveal different features of the network. Several measures of connectedness can be computed from a given network to uncover the relationships that it embodies. Some of the measures from the networks literature have been adopted in the financial network literature for

[24] For some network measures these weights might be ignored. This will be explicitly stated when necessary.

the study of financial phenomena. For instance, Billio et al. (2012) uses degree, closeness, and eigenvector centrality, among other measures for the analysis of systemic risk; while Diebold and Yilmaz (2014) focus on degree (in-degree and out-degree) and diameter for the study of volatility connectedness among banks. It is important to bear in mind, however, that a network is a complex system and, as such, each measure only captures a dimension of it. The ensemble of these measures is what allows us to build a broad picture about the network. In particular, discrepancies among them (i.e., low correlation) are valuable as they may reveal a different aspect of the network.

In the spatial econometrics literature, spatial statistics are classified as either global or local, depending on whether they characterize the network as a whole or a particular component of it. Similarly, in the network literature measures can be classified as either macro, to describe broad characteristics of the network, or micro, to compare nodes individually and understand how each of them relates to the network as a whole (see Jackson (2008, p.37)). In this paper, therefore, I distinguish between global (or macro) and local (or micro) measures and split the analysis accordingly. In addition, in Appendix 2.A I briefly discuss the layout chosen in this paper to graph the network.

Measures of centrality are the most widely used to assess micro aspects of a network. In this paper I consider four measures of centrality, from the network literature, as they complement each other.[25] Let $\mathcal{A}$ be the adjacency matrix corresponding to the network W, i.e., $\mathcal{A} = W'$, we define the following micro measures:

1. *Degree centrality:* This is the simplest measure of centrality. It ranges between 0 and 1 and represents how well a node is connected in terms of number of

[25] See Jackson (2008) for further details.

direct connections (in or out), on average. Formally, for node i we have

$$C_d^{i\leftarrow j} = \frac{\#\{j:\ [\mathcal{A}]_{ji} \neq 0\}}{K-1} \qquad [\text{In-degree}_i]$$
$$C_d^{i\rightarrow j} = \frac{\#\{j:\ [\mathcal{A}]_{ij} \neq 0\}}{K-1} \qquad [\text{Out-degree}_i]$$
$$C_d^{i-j} = C_d^{i\leftarrow j} + C_d^{i\rightarrow j} \qquad [\text{Total-degree}_i]$$

However, degree centrality overlooks several important features of a network. For instance, it does not account for how well located a node is in the network. That is, a node could have few links, but it could lie in a pivotal location in the network.

2. *Closeness centrality:* This measure reflects how close a node is, on average, to any other node. Intuitively, it describes the extent of influence of a node on the network. Being close to every other node can be important in situations where something is transmitted through the network. Formally, the closeness centrality of a vertex i is defined by the inverse of the average length of the shortest paths to/from any other vertex j in the graph:

$$C_{cl}^i = \frac{K-1}{\sum_{j\neq i} l(j,i)}$$

where $l(j,i)$ is the (shortest) distance between i and j given by the shortest path between i and j.[26]

3. *Betweenness centrality*: According to Borgatti (2005, p.60) this measure can be defined as "the share of times that a node j needs a node i (whose centrality is being measured) in order to reach a node k via the shortest path." Formally, let $P_i(kj)$ denote the number of shortest paths from j to k through i, and let $P(kj)$ be the total number of shortest paths between k and j. We can estimate

[26]If there is no (directed) path between vertex j and i then the total number of vertices is used in the formula instead of the path length.

how important i is in terms of connecting k and j by looking at the ratio $P_i(kj)/P(kj)$ (Jackson, 2008, p.39). Then, averaging over all pairs of nodes (excluding node i) gives

$$C_{be}^{i} = \sum_{k \neq j:\ i \notin \{k,j\}} \frac{P_i(kj)/P(kj)}{(K-1)(K-2)/2}$$

Intuitively, high-betweenness vertices (i.e., ratio closer to 1) lie on a large number of non-redundant shortest paths between other vertices; they can thus be thought of as "bridges" or "boundary spanners." In contrast, low-betweenness vertices (i.e., ratio closer to 0) are less critical to other vertices.

Given that we have a weighted directed network, I compute the shortest path between node i and j by summing the inverse weights involved in each path connecting the two nodes and then choosing the path with the smallest value of total inverse weight. In addition, I "normalize" the weights by the average weight in the network for interpretation purposes (see Opsahl, Agneessens, and Skvoretz (2010)). The main advantage of this normalization is that it makes the measure (e.g., closeness) comparable across networks with different weight ranges.

These three first measures are straightforward to compute from the adjacency matrix. However, they focus mostly on "quantity" rather than "quality" of connections. That is, we may be interested in assessing not only how well connected or how close a node is to many other nodes, but also whether it is connected to other "central" nodes or "key players" in the network. Hence the next measure is based on the premise that a node's importance is determined by how important its neighbors are.

4. *Bonacich Power centrality*: This measure generalizes degree centrality by taking into account the prestige of a node's neighbors. Bonacich proposed that a node's centrality (or prestige) is equal to a function of the prestige of those they are connected to. Hence, if a node is connected to very central nodes then it should

have higher centrality than those with the same degree but connected to less central nodes.[27] Bonacich Power centrality is defined as

$$C_{BP}(\alpha, \beta) = \alpha(I_K - \beta\mathcal{A})^{-1}\mathcal{A}\iota$$

where $C_{BP}(\alpha, \beta)$ is a $K \times 1$ vector containing the power centrality measure of every node, β is an attenuation parameter with $|\beta| < 1/\lambda_{\mathcal{A}\iota}$ (the reciprocal of the largest eigenvalue of $\mathcal{A}$), and $\alpha > 0$ is a scaling parameter. Intuitively, β allows one to control how the value of being connected to other nodes decays with distance. In other words, it reflects the radius of power. Small values of β weight local structure, larger values weight global structure. If β is positive, then a node has higher centrality when tied to nodes who are central (e.g., status matters). If β is negative, a node is more powerful only as its neighbors become weaker (e.g., competition dominates). As β approaches zero, we consider only direct connections, hence, we obtain degree centrality.[28]

Even though centrality measures are quite popular, they only focus on local aspects of the network, while, as discussed above, it is also important to analyze the network as a whole. In this paper, I propose to study the network globally via cohesive-blocks analysis, ρ, and the network impact by order of neighbors. While cohesive-blocks analysis comes from the network literature, the two other measures are related to the spatial econometrics literature. They are defined as follows:

1. *Cohesive-blocks analysis*: This analysis is based on the skeleton of the network graph (i.e., the undirected graph). Cohesion is closely related to concepts of strong ties among the members of embedded social groups or closed social circles. Structural cohesion is defined as the minimal number of individuals in a

[27] As each node's power depends on each other node's power simultaneously, the solution to this problem is based on fixed point theory and matrix algebra.

[28] When $\beta \to 1/\lambda_{\mathcal{A}}$, this is equal to the familiar eigenvector centrality score, up to a multiplicative constant.

network that need to be removed to disconnect the group (Moody and White (2003)). Equivalently, it can be defined as the minimum number of independent paths linking each pair of individuals in the network. Intuitively, networks are structurally cohesive if they remain connected even when nodes are removed.

Formally, define vertex connectivity as the minimum number of nodes κ whose deletion from a graph $\mathcal{G}$ disconnects it. Then, cohesive blocking is a method of determining hierarchical subsets of graph vertices based on their structural cohesion (or vertex connectivity).

Definition 2.2. *(κ-cohesive) For a given graph $\mathcal{G} = (V, E)$, $S \subset V$ is said to be maximally κ-cohesive if $\nexists\ S' \supset S$, such that S' is l-cohesive with $l \geq \kappa$.*

Cohesive blocking is a process through which, given a κ-cohesive set of vertices, maximally l-cohesive subsets are recursively identified with $l > \kappa$. Thus a hierarchy of vertex subsets is found, with the entire graph $\mathcal{G}$ at its root (i.e., $\kappa = 0$).[29] As a result, cohesive blocking generates a network hierarchy in which individuals are "nested" at different levels. In this paper, I use the structural cohesion algorithm of Moody and White (2003) that is implemented in the igraph R package (i.e., cohesive.blocks).

2. *ρ-measure*: The parameter ρ from Section 2.3.2 is a natural global measure of the overall network influence. Since W is row-standardized, the parameter ρ can be interpreted as a measure of the overall strength of network dependence. For instance, in the volatility connectedness example, a positive value of ρ will indicate that nodes in the network are expected to have higher volatility values if, on average, their neighbors have high volatility values.

3. *Impact by order of neighbors*: Taking the r-th power of the network matrix W provides the impact of r-th order neighbors. We can expect that the impact

[29] As defined in the igraph R package documentation.

declines, on average, as we move from lower-order neighbors to higher-order neighbors. In particular, the pattern of decay can be of interest to assess the persistence of spillovers. Since the network influence as a whole is governed by ρ, I compute the impact by order of neighbor using powers of ρW. The average total impact for r-order neighbors is given by

$$\overline{Nh}^{r} = K^{-1}\iota_{K}'(\rho W)^{r}\iota_{K}$$

2.6 Application to financial integration

A natural application of the MDE methodology is studying financial integration among countries. Financial integration is a phenomenon in which financial markets in neighboring, regional and/or global economies are closely linked together. Financial integration is a natural application since it relies on the idea that countries affect each other through time at a financial level and, hence, belong to a certain network. When a crisis takes place, the negative shock spreads to multiple countries (in different intensities) due to the fact that they are financially connected. This may affect financial integration patterns among countries as, for example, some links might break while others may strengthen. These ideas are closely related to the concept of financial contagion "contagion will be said to occur when the impact of the systematic risk on individual volatility processes is even stronger during the crisis period" (Dungey and Renault, 2013, p.2).

For this application, I use daily realized return volatility data at 10-min intervals (RV10) of main stock indexes of a set of countries from the Oxford Man Realized Volatility Library (Heber, Lunde, Shephard, and Sheppard, 2009). The sampling interval was chosen to account for the trade-off between minimizing micro-structural bias (from using high-frequency intra-day data at low intervals) and minimizing sam-

pling error (from using larger sampling intervals). The sample consists of 16 world leading indexes from June 2003 through March 2015. The list of indexes is given in Table 2.2. The European Union leading index (EURO STOXX50) is included to control for missing eurozone countries that are not readily available in the Oxford Man data set (e.g., Greece, and Portugal).[30] The application is coded in R.

I study volatility connectedness since, as phrased in Diebold and Yilmaz (2014, p.125), "First, if volatility tracks investors fear... then volatility connectedness is the 'fear connectedness' expressed by market participants as they trade... Second, ... we are particularly interested in crises, and volatility is particularly crisis-sensitive." Connectedness in mean has also been studied in the literature, though to a lesser degree. An example is the paper by Billio, Getmansky, Lo, and Pelizzon (2012) which looks at connectedness in mean return in a Granger Causality sense. However, they acknowledge that there might be higher order effects not captured by connectedness in mean, and they found that causal relationships are even stronger if we also take into account the level of risk financial institutions may face, i.e., their volatility (Billio et al., 2012, p.551). The use of volatility has the additional advantage that log-volatility is approximately Gaussian; therefore it is suitable for the PC-algorithm which invokes normality.[31]

Furthermore, I analyze the network in different periods of time. The underlying reason is to explore whether the 2008 crisis changed the connectivity patterns observed in the network before, during and after the crisis. Accordingly, the analysis is conducted by dividing the data into four time periods called hereafter: pre-crisis period (06/25/2003 - 03/01/2007), crisis period (03/02/2007 - 02/26/2010), post-crisis period (03/01/2010 - 12/28/2012), and euro-crisis period (01/02/2013 - 03/05/2015).

[30]The EURO STOXX50 index covers 50 stocks from 12 eurozone countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal and Spain.

[31]There are other algorithms available, like the VAR-LiNGAM algorithm in Moneta, Entner, Hoyer, and Coad (2013) but it is currently limited to a rather small number ($N = 8$) of individuals in the network.

Table 2.2: List of stock return indexes in the sample by country, K=16.

Index	Country	Code	Index	Country	Code
AEX	The Netherlands	NLD	Nikkei 225	Japan	JPN
DAX	Germany	DEU	KOSPI	South Korea	KOR
IBEX 35	Spain	ESP	HSI	Hong Kong	HKG
FTSE 100	United Kingdom	GBR	FTSTI	Singapore	SGP
CAC 40	France	FRA	DJIA	USA	USA
FTSE MIB	Italy	ITA	MXX	Mexico	MEX
SSMI	Switzerland	CHE	BVSP	Brazil	BRA
EURO STOXX 50	European Union	EURO	AORD	Australia	AUS

Notice that the period after the "crisis period" is split into two for the purpose of separately studying the period where a new set of austerity measures in Greece were being discussed, upon revision of Greece's second bailout package, to deal with the Greek government-debt crisis. The dates delimiting each period have been carefully selected based on several sources: dates chosen in the financial network literature (e.g., Diebold and Yilmaz (2014), Billio et al. (2012)), the detailed survey on crisis dates in Dungey, Milunovich, Thorp, and Yang (2015), and the discussion on the topic in Contessi, De Pace, and Guidolin (2014). In addition, the time periods were chosen based on key dates, for instance: the turning point in the US monetary policy in 2003 (in 06/25/2003 the Fed set the interest rate at 1%), the first time Bear Stearns hedge fund acknowledged its financial problem (03/01/2007), and the time when the real magnitude of Greece's high government-debt was revealed (03/01/2010). Time series and distributional plots of the data for the full period are shown in Appendix 2.B.

2.6.1 The local measures

To study the network I use the local and global measures described in Section 2.5.[32] Hereafter, to simplify exposition, I will refer to members of the network as countries even though EURO STOXX50 corresponds to more than one country. The micro measures are summarized in Tables 2.3 and 2.4 below. The first table shows in-degree, out-degree, and total-degree for each country and each period. Overall, the number of connections, on average, decreased after the crisis, and, in particular, during the euro-crisis period. Intuitively, many links might have been broken due to the crisis. Moreover, even though degree increased during the crisis, after this period it decreased to levels below the pre-crisis period.

If we look at in-degree, we see that the European countries had higher, though diverse, levels prior to the crisis. This may be associated with membership in the European Union. During the crisis period, however, some of the observed disparities evened out possibly because countries in the EU were in one way or another subject to the same financial shocks. In addition, the higher level of in-degree from the UK in the euro-crisis period may reflect its status as a global financial hub. In terms of out-degree, it is also the case that dispersion as well as mean degree decreased after the crisis period. A feature worth noticing is that US out-degree was higher than in-degree during the whole period. Another interesting feature is that the higher amount of out-degree during the crisis belongs to France (0.40) while during the post-crisis period it belongs to Germany (0.27). This may be attributable to their level of involvement during the corresponding periods. Finally, in the euro-crisis period, the highest value of out-degree corresponds to the EURO STOXX50 index which may be due to the effect that Greece's debt crisis events were having at the time on the rest of the European Union.

[32]Notice that I do not look into pairwise values of these measure given that there are sixteen countries, four centrality measures (plus the in, out and total distinction in degree centrality), and four time periods.

Table 2.3: Degree Centrality. The sample is from June 25th, 2003 to March 5th, 2015, K=16. The analysis is divided in four periods: pre-crisis period (06/25/2003 - 03/01/2007), crisis period (03/02/2007 - 02/26/2010), post-crisis period (03/01/2010 - 12/28/2012), and euro-crisis period (01/02/2013 - 03/05/2015).

Index	Pre-crisis			Crisis			Post-crisis			Euro-crisis		
	In	Out	Total	In	Out	Total	In	Out	Total	In	Out	Total
NLD	0.13	0.33	0.47	0.20	0.13	0.33	0.20	0.07	0.27	0.07	0.20	0.27
DEU	0.27	0.13	0.40	0.13	0.07	0.20	0.07	0.27	0.33	0.13	0.07	0.20
ESP	0.27	0.07	0.33	0.20	0.13	0.33	0.00	0.13	0.13	0.13	0.00	0.13
GBR	0.13	0.33	0.47	0.13	0.20	0.33	0.13	0.20	0.33	0.27	0.07	0.33
FRA	0.20	0.20	0.40	0.00	0.40	0.40	0.13	0.20	0.33	0.07	0.13	0.20
ITA	0.00	0.33	0.33	0.27	0.20	0.47	0.20	0.13	0.33	0.13	0.07	0.20
CHE	0.27	0.07	0.33	0.27	0.07	0.33	0.13	0.13	0.27	0.07	0.13	0.20
EURO	0.20	0.20	0.40	0.13	0.27	0.40	0.27	0.07	0.33	0.00	0.27	0.27
JPN	0.07	0.20	0.27	0.27	0.07	0.33	0.07	0.13	0.20	0.13	0.07	0.20
KOR	0.13	0.00	0.13	0.13	0.20	0.33	0.13	0.13	0.27	0.13	0.00	0.13
HKG	0.20	0.13	0.33	0.13	0.07	0.20	0.13	0.00	0.13	0.07	0.13	0.20
SGP	0.07	0.07	0.13	0.13	0.20	0.33	0.13	0.13	0.27	0.07	0.13	0.20
USA	0.07	0.13	0.20	0.13	0.20	0.33	0.07	0.20	0.27	0.00	0.20	0.20
MEX	0.20	0.07	0.27	0.07	0.07	0.13	0.13	0.07	0.20	0.13	0.07	0.20
BRA	0.13	0.00	0.13	0.27	0.07	0.33	0.13	0.07	0.20	0.13	0.07	0.20
AUS	0.07	0.13	0.20	0.00	0.13	0.13	0.07	0.07	0.13	0.07	0.00	0.07

Note: "In + Out" may slightly differ from "Total" in some cases due to rounding.

Even though each country has only a few in or out connections, it can still be central to the network in different ways. Centrality measures capturing features other than degree are closeness, betweenness, and Bonacich power. The results for these measures are given in Table 2.4. As explained in Section 2.5, closeness is a good measure to single out countries that may influence others in the network because they are "close" to most countries. They are "close" in the sense that they are positioned in the network near to many nodes. During the pre-crisis period, the highest value of closeness correspond to Italy (0.46), followed by The Netherlands (0.36). In the case of Italy, the high value of closeness may have been related to the unstable economic situation that Italy was facing well before the crisis due to a prolonged low-growth period starting around 1995 (the last year when the Lira devalued with respect to the Deutsche Mark). In the case of The Netherlands the observed value of closeness may have been related to its strong position in world trade. Not surprisingly, other values that are relatively high correspond to France (0.30), UK (0.22), and Germany

(0.23), as these countries are leading countries within the European Union.

In the crisis period, although values of closeness remained similar on average, dispersion decreased, implying that countries became more equally close to each other. This is consistent with the idea that as the financial instability spread during this time period, countries became more financially connected and the situation was indeed global. In particular, not surprisingly, we see that the US became much closer to the rest of the world (increase from 0.07 to 0.15), and that France became even more influential (i.e., closer). The lowest values are observed for Hong Kong and Japan. In the post-crisis period, the average level of closeness and dispersion returned overall to pre-crisis levels, with the higher values corresponding to EU countries. In the euro-crisis period, however, closeness values decreased even further possibly because the inability of the EU to solve Greece's debt crisis problems continued to create more financial instability. The higher value of closeness corresponds to the EURO STOXX50 (0.27), followed by France (0.20).

Betweenness serves as a complementary measure to closeness. It allows us to pin down countries considered central to the network because they are a means to reaching other countries. In other words, it allows us to identify which countries play the role of a bridge connecting different countries in the network. It is natural to see some zero values for this measure since some countries may not connect others. In the pre-crisis period, the highest value of betweenness corresponds to The Netherlands (0.36) and Germany (0.36) followed by EURO STOXX50 (0.30). Comparing the pre-crisis to the crisis period, we see that many countries with a zero value of betweenness now have positive values. In fact, most countries have a positive value of betweenness in the crisis-period. This is an interesting feature because it may be reflecting spillovers that took place during the crisis.

Table 2.4: Centrality Measures: closeness, betweenness, and Bonacich power centrality. The sample is from June 25th, 2003 to March 5th, 2015, K=16. The analysis is divided in four periods: pre-crisis period (06/25/2003 - 03/01/2007), crisis period (03/02/2007 - 02/26/2010), post-crisis period (03/01/2010 - 12/28/2012), and euro-crisis period (01/02/2013 - 03/05/2015).

Index	Pre-crisis			Crisis		
	Closeness	Betweenness	B. Power b = 0.33	Closeness	Betweenness	B. Power b = 0.33
NLD	0.36	0.36	1.91	0.17	0.06	0.73
DEU	0.23	0.36	0.64	0.12	0.03	0.37
ESP	0.00	0.00	0.42	0.15	0.06	0.53
GBR	0.22	0.00	1.65	0.17	0.13	1.10
FRA	0.30	0.28	0.83	0.43	0.00	2.69
ITA	0.46	0.00	2.33	0.17	0.16	0.97
CHE	0.11	0.00	0.34	0.13	0.11	0.44
EURO	0.20	0.30	0.88	0.22	0.16	1.24
JPN	0.15	0.24	0.81	0.08	0.15	0.54
KOR	0.06	0.00	0.00	0.13	0.20	0.93
HKG	0.15	0.29	0.88	0.08	0.00	0.31
SGP	0.15	0.21	0.42	0.13	0.21	0.78
USA	0.07	0.09	0.29	0.15	0.25	1.17
MEX	0.07	0.01	0.12	0.14	0.24	0.52
BRA	0.06	0.00	0.00	0.13	0.25	0.30
AUS	0.04	0.00	0.58	0.13	0.00	0.54

Index	Post-crisis			Euro-crisis		
	Closeness	Betweenness	B. Power b = 0.33	Closeness	Betweenness	B. Power b = 0.33
NLD	0.26	0.15	0.59	0.17	0.05	1.54
DEU	0.27	0.10	2.07	0.10	0.03	0.63
ESP	0.30	0.00	1.02	0.06	0.00	0.00
GBR	0.32	0.30	1.49	0.09	0.14	0.54
FRA	0.33	0.55	1.45	0.20	0.01	1.32
ITA	0.27	0.41	1.34	0.07	0.01	0.34
CHE	0.30	0.18	1.23	0.10	0.05	0.87
EURO	0.15	0.32	0.63	0.27	0.00	2.21
JPN	0.08	0.15	0.69	0.07	0.00	0.34
KOR	0.07	0.03	0.52	0.06	0.00	0.00
HKG	0.06	0.00	0.00	0.07	0.05	0.80
SGP	0.07	0.23	0.46	0.08	0.13	1.07
USA	0.11	0.27	1.01	0.10	0.00	1.64
MEX	0.07	0.20	0.31	0.08	0.15	0.70
BRA	0.06	0.28	0.39	0.09	0.14	0.58
AUS	0.06	0.16	0.29	0.06	0.00	0.00

Note: In Bonacich power centrality the scaling parameter α is set to 1.

These spillovers continued during the post-crisis period, but did not extend to the euro-crisis period. In particular, in the crisis period we see the role of the US as a connecting bridge; it has the highest value of betweenness (0.25) together with Brazil who is a leader in Latin America and is more internationally exposed than other Latin American countries. In the post-crisis period, the highest values of betweenness correspond to France (0.55), Italy (0.41), EURO STOXX50 (0.32), and UK (0.30). In particular, it may be reflecting the central role of France in the EU and the role of the UK as a financial center. Also, the value for the US remained relatively high (0.27) during this period.

In the euro-crisis period, however, only the UK preserves a high value of betweenness relative to other EU countries (0.14) possibly due to the UK's continued position as a financial hub. After the post-crisis period, low values of betweenness are observed in most countries. A plausible interpretation is that many financial relations were broken, helping, to some extent, to stop negative spillover effects. Other high values, in relative terms, correspond to Mexico (0.15), Brazil (0.14) and Singapore (0.13).

Regarding Bonacich prestige or power centrality, a few features are worth noticing. Bonacich power centrality allows us to detect "powerful" countries in the sense that they are connected to countries who are powerful (i.e., very central) themselves. Recall that this measure depends on the tuning parameter β. In Table 2.4, I present Bonacich power for β (b in the table) equal to 0.33 to weight more global structure.[33] Looking at the broad picture, it is interesting to see that pre-crisis UK had more "prestige" than the US (1.65 vs 0.29). However, in the crisis period this difference evened out (1.10 vs 1.17). In addition, in the crisis period we see the rising role of France who became a very powerful player (highest value of Bonacich power, 2.69). An interesting feature, not already captured by other measures, is the role of Ger-

[33]I also tried a lower value of β, $b = 0.25$, that weights more local structure. Results are fairly similar with either choice of b.

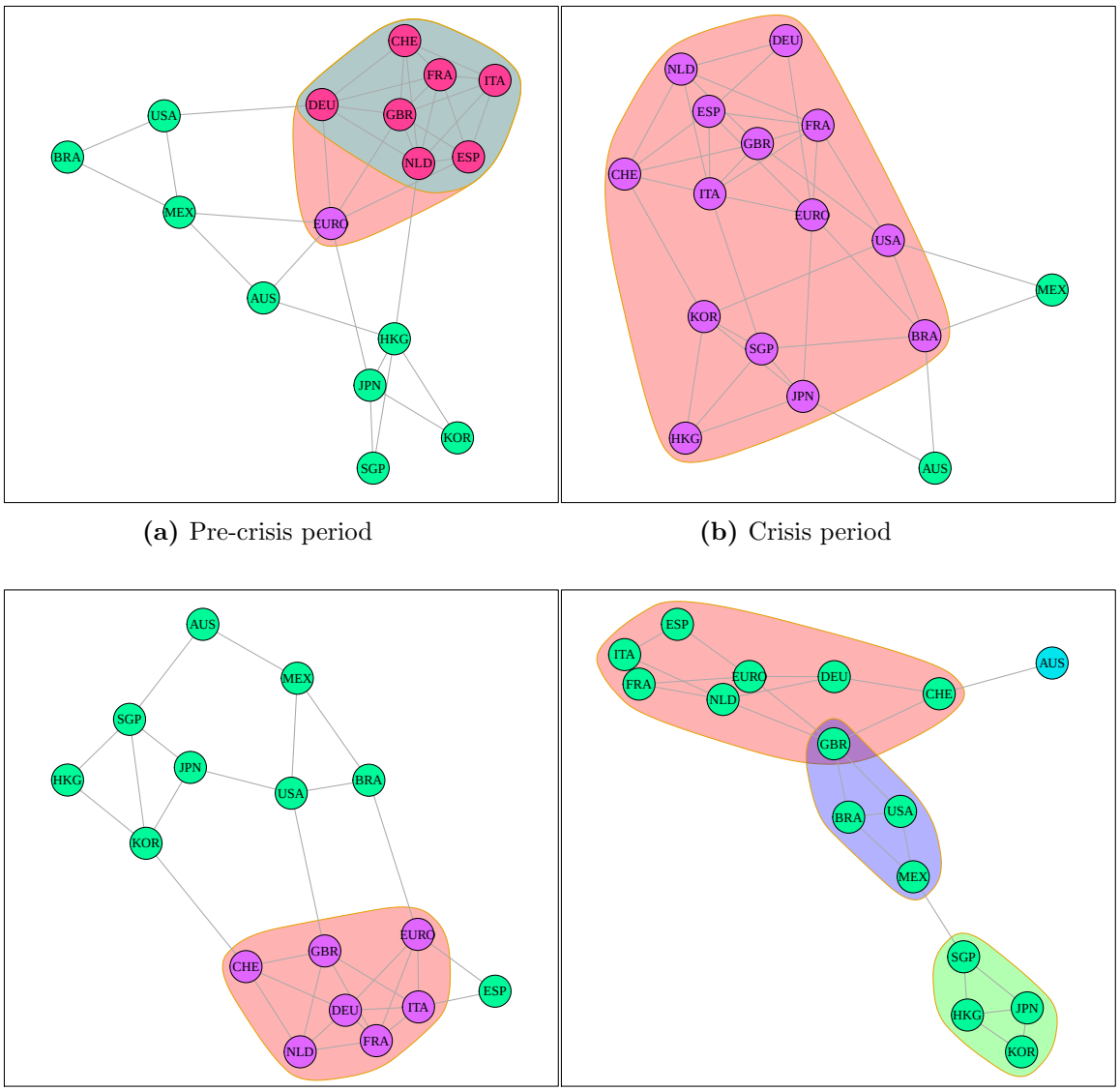
many in the post-crisis period. Bonacich power shows that in the post-crisis period Germany became the most powerful player (2.07) followed by the UK (1.49) and France (1.45). However, in the euro-crisis period Germany's position is no longer that prestigious. We see that EURO STOXX50 now has the highest value (2.21), followed by France (1.32) and the US (1.64).

2.6.2 The global measures

A graphical representation of the network in each period, using the Fruchterman-Reingold layout, is given in Appendix 2.B. These graphs provide evidence of distinctive changes in connectivity patterns across the periods considered. For instance, the network in the euro-crisis period compared to the crisis period has tighter clusters. Notice that nodes positions are not fixed across graphs exactly because fixing them would prevent us from finding changes in connectivity patterns; this is one advantage over circle layouts. To uncover these changes in connectivity patterns we can use a cohesive-blocks analysis.

The cohesive-blocks analysis in Figure 2.2 shows very distinctive patterns across periods. Not only is it consistent with the features previously discussed in the centrality measures analysis, but also it allows us to better understand what these measures were capturing. In the pre-crisis period we see that only the EU countries belong to a marked cohesive-block. However, the events of the crisis may have brought countries closer financially as the result of important spillovers. Indeed, with the exception of Mexico and Australia, all countries became part of one big cohesive-block.

In the post crisis period, we see a picture similar to that of the pre-crisis period. This is likely the case as the spillovers that took place during the crisis decreased in this period. Many countries entered a recovery phase, while some others were still dealing with the financial and economic instability brought by the financial crisis. One clear difference, though, is the case of Spain. The financial crisis hit harder in the



(a) Pre-crisis period **(b)** Crisis period

(c) Post-crisis period **(d)** Euro-crisis period

Figure 2.2: Cohesive-blocks analysis by period. The sample is from June 25th, 2003 to March 5th, 2015, K=16. The analysis is divided in four periods: pre-crisis period (06/25/2003 - 03/01/2007), crisis period (03/02/2007 - 02/26/2010), post-crisis period (03/01/2010 - 12/28/2012), and euro-crisis period (01/02/2013 - 03/05/2015). Node's colors correspond to different degree of cohesiveness: light blue $\kappa = 1$, green $\kappa = 2$, purple $\kappa = 3$, dark pink $\kappa = 4$. Cohesive-blocks are marked by shaded area.

EU among its weaker members. Namely, Ireland, Portugal, Italy, Spain, and Greece. This is likely why we see Spain left out of the EU cluster. One may have expected to see the same pattern for Italy, however, this may not be the case as Italy was already in a critical situation before the crisis even started; the crisis only worsened Italy's

situation. In contrast, Spain was doing well before the crisis and was brought into a deep financial crisis only after the events of 2008.

In the euro-crisis period, however, the pattern is quite different from the previous pictures. There are three very distinctive clusters plus Australia: the EU countries, the Asian countries, and countries in the American continent. A likely explanation is that Greece's debt crisis situation isolated the EU from the rest of the world, turning the UK into a central link to the non-EU world. Furthermore, this may suggest that, after the crisis, countries tried to reduce their future vulnerability to spillovers effects. Since Greece's situation became a deeper problem after the second bailout was unsuccessful (i.e., end of post-crisis period), it is reasonable not to see the euro-crisis patterns in the post-crisis period.

The ρ-measure in each period of time (i.e., $\hat{\rho}$) is given in Table 2.5 below, with standard errors in parentheses. During the crisis period this measure was the highest (0.82), meaning that the network as a whole mattered the most during this time period. Furthermore, after the crisis period even though the strength of the network connectedness decreased (from 0.75 to 0.71), it did not return to pre-crisis levels (0.67). This may be attributable to the financial problems in the Euro area, in particular, those regarding Greece's debt crisis situation.

Table 2.5: ρ-measure across time periods. The sample is from June 25th, 2003 to March 5th, 2015, K=16.

Period	rho-measure	
Pre-crisis	0.6712	(0.0072)
Crisis	0.8187	(0.0072)
Post-crisis	0.7455	(0.0068)
Euro-crisis	0.7075	(0.0076)

Note: std. errors in parentheses.

In addition to period specific analysis of the ρ-measure, I conduct a rolling-sample

exercise to shed some light on the dynamics of connectedness embedded in this measure. I estimate the model using a rolling-sample window of 500 days; the windows are rolled through the sample one day at a time. This gives a total of 2471 samples and, therefore, 2471 conditional values of $\hat{\rho}$. The advantage of using this technique is basically to look at any changing property, the ρ-measure in our case, of a series over time. We obtain an estimate of the property over time instead of one single constant measure for specific periods (i.e., pre-crisis, crisis, post-crisis, and euro-crisis).

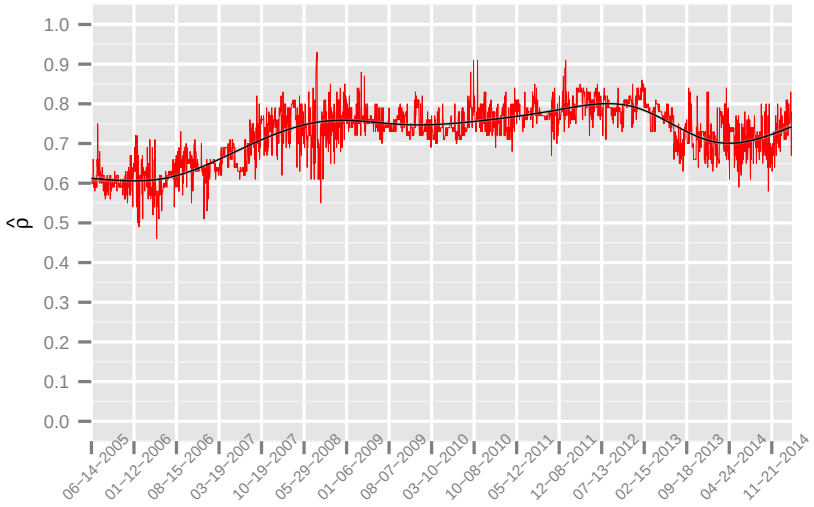


Figure 2.3: Rolling-sample plot of $\hat{\rho}$. The rolling estimation window is 500 days, and windows are rolled by one day at a time. Dates reported correspond to the ending date of the rolling window. The solid black line corresponds to a smoothed conditional mean.

Figure 2.3 shows that until the beginning of 2006 the overall network influence was around 0.6. Afterwards, we see that there is a persistent increase in the overall strength of connectivity that achieves its maximum around August/September 2008; this coincides with the peak moment of the 2008 crisis. Following that peak we see the network influence stays relatively high though stable (around 0.75) for a considerable period of time until the end of 2009, and then it slowly starts to increase during 2010. This last observation coincides with the timing of the sovereign debt crisis that erupted in Greece, Ireland, and Portugal in 2010. The overall network influence then continues to slowly increase until July of 2012, which coincides with

the timing of when the European Stability Mechanism (ESM) was planned to be ratified.[34] Subsequently, the overall strength of connectivity decreases; this pattern persists during most of 2013. After that, $\hat{\rho}$ starts to increase again. This last increase may be associated with the unstable financial and economic situation taking place in Greece and the associated instability brought to the EU. In particular, it is important to notice that at the end of the full period the levels of $\hat{\rho}$ did not return to pre-crisis levels.

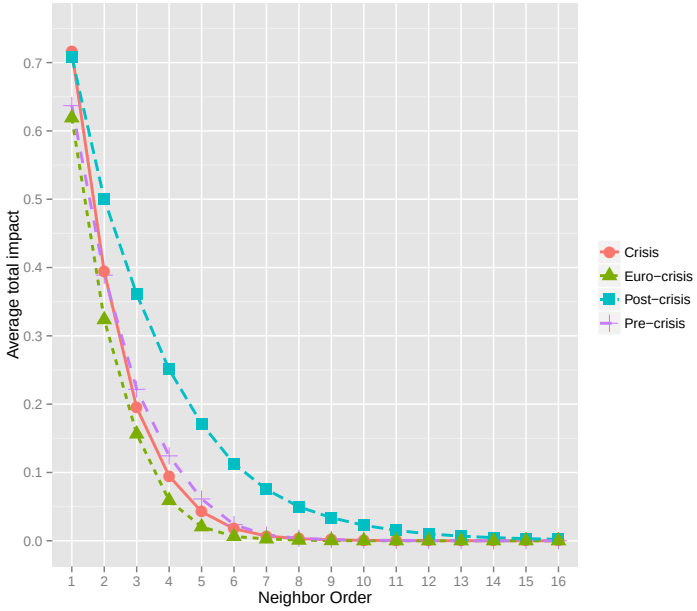


Figure 2.4: Average total network impact by order of neighbor across periods. The sample is from June 25th, 2003 to March 5th, 2015, K=16. The analysis is divided in four periods: pre-crisis period (06/25/2003 - 03/01/2007), crisis period (03/02/2007 - 02/26/2010), post-crisis period (03/01/2010 - 12/28/2012), and euro-crisis period (01/02/2013 - 03/05/2015).

Finally, I look at the average impact of the network by order of neighbor for each time period. The underlying idea of looking at this measure is to see whether countries get affected mostly through their direct connections or whether the connections of their connections also play an important role. For instance, in an analogy with a

[34]The European Stability Mechanism (ESM) was established by a treaty among the eurozone countries as a permanent rescue funding program to succeed the temporary European Financial Stability Facility (EFSF) and the European Financial Stabilization Mechanism (EFSM) in July 2012. However, it had to be postponed and only entered into force in September 2012.

network of friends: first order neighbors are direct friends, second order neighbors correspond to friends of direct friends, while 3rd order neighbors are friends of friends of direct friends, and so on so forth. In Figure 2.4 we see that, overall, except for the post-crisis period, up to 6th order neighbors have some indirect impact. As expected, the impact is decreasing with the neighbor order, starting at around 0.7 in the crisis and post-crisis periods and at 0.63 in the pre-crisis and euro-crisis periods. It is interesting to see that in the post-crisis period the rate of decay is lower than in other periods and that some influence is found up to 11th order neighbor. This finding may be the result of strong spillover effects happening during this period, which is in agreement with observations made when analyzing the betweenness measure. Surprisingly, we do not see a differential rate of decay for the crisis period when compared to the pre-crisis period. However, the impact of direct connections is higher in the crisis period (0.70) which is consistent with our previous findings. As expected, in the euro-crisis period the effect of higher order neighbors decays at the fastest rate and also displays the lowest levels.

2.7 Conclusion

A common practice in the spatial econometrics literature is to assume that the spatial weight (or network) matrix is row-standardized and pre-multiplied by a scalar parameter that captures the overall network influence. This parametrization of the network has shown to be very useful for interpretation purposes; hence its popularity among users of spatial-type models. When modeling networks across geographical space, the network is known and defined based on some notion of geographical distance. All efforts center on estimating the overall network influence parameter. In contrast, in settings unrelated to geographical ties, the use of spatial models is generally unfeasible due to lack of data on network ties. Some papers have addressed this issue by

proposing a methodology to estimate the network from the data. Two recent examples are Manresa (2015) and Lam and Souza (2015), although in the former spillovers take place through a variable X (i.e., individual's characteristics) instead of Y (i.e., the output variable) and in the latter there is no disentangling of the overall network influence parameter from the network matrix (i.e., $W^* = \rho\, W$ is estimated instead).

In this paper I addressed these shortcomings by developing a two step procedure based on a minimum distance approach which allows for the estimation of both the scalar parameter ρ and the row-standardized network matrix W. This was done in a SVAR context by interpreting the time series spatial model (i.e., T-SAR) as a constrained SVAR. Moreover, I developed a test to assess the constraints imposed by the T-SAR model on the SVAR model. Implementation of the MDE methodology is straightforward given identification of the SVAR, which is an unresolved issue in the time series literature. In this paper I explored one possible identification strategy involving machine learning methods, although the MDE methodology is not limited to this identification approach. However, there is no free lunch. Data-driven methods deliver many times only partial solutions to the ordering problem, and there are no general guidelines on which algorithm to implement and which options to choose when implementing it. This paper, through a simulation exercise, shed some light on these issues when using financial data. Finally, the MDE methodology was illustrated through an application to financial integration among countries based on daily realized volatility data for the period June 2003 through March 2015. Both the cohesive-blocks analysis and the overall network influence parameter moving window exercise supported the existence of an interplay between crises and changes in financial integration patterns.

The financial networks literature currently faces two main challenges. On the one hand, how does one estimate the network links given that data on the network structure is seldom available, and, on the other hand, how does one model the channels

or mechanisms through which the network or spillover effects take place. While addressing the former makes possible the study of network effects, providing answers to the latter is crucial to better understanding the determinants underlying the observed spillover patterns. This paper contributed to the literature with regard to the first challenge. The simplest approach to the second challenge would consist of extending the model to allow for exogenous covariates (e.g., individual's characteristics). A more sophisticated approach would require the design of a structural model for network formation in financial settings. This is left for future research.

Appendix

2.A Mathematical appendix

2.A.1 Proof of Proposition 2.2

First, notice that both models take the following form:

$$Y_t = C_0 Y_t + C_1 Y_{t-1} + \cdots + C_p Y_{t-p} + \varepsilon_t$$

where C_i, $i = 0, 1, ..., p$, is a $K \times K$ matrix of coefficients, C_0 has zero elements on its main diagonal; and ε_t is a serially uncorrelated error term with $\varepsilon_t \sim (0, \Sigma_\varepsilon)$.

In the T-SAR model $C_0 = \rho W$ and $C_i = \Gamma_i$, for $i = 1, ..., p$, while in the SVAR model $C_0 = G$ and $C_i = B_i$, for $i = 1, ..., p$. Under the assumption that the SVAR is identified, and since Γ_i and B_i are unrestricted, it is enough to show that if $RPG\iota_K = 0$ the SVAR can be written as the T-SAR model.

If $l = K - 1$ the proof is trivial. Take $0 \leq l < K - 1$, without loss of generality assume that the first l rows of G have all its elements equal to zero. Then

$$PG\iota_K = G\iota_K = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sum_{j=1}^{K} g_{(l+1)j} \\ \vdots \\ \sum_{j=1}^{K} g_{Kj} \end{bmatrix}$$

Hence,

$$RPG\iota_{_K} = 0 \iff \begin{cases} g_{(l+1),1} + \cdots + g_{(l+1),K} = g_{(l+2),1} + \cdots + g_{(l+2),K} \\ g_{(l+2),1} + \cdots + g_{(l+2),K} = g_{(l+3),1} + \cdots + g_{(l+3),K} \\ \vdots \\ g_{(K-1),1} + \cdots + g_{(K-1),K} = g_{K,1} + \cdots + g_{K,K} \end{cases} \tag{2.16}$$

This imposes that all the rows of constrained G sum to the same quantity. Let that quantity be ρ, then taking common factor ρ

$$C_0 = \rho \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \\ \frac{g_{(l+1),1}}{\rho} & \cdots & \frac{g_{(l+1),K}}{\rho} \\ \frac{g_{(l+2),1}}{\rho} & \cdots & \frac{g_{(l+2),K}}{\rho} \\ \vdots & \vdots & \vdots \\ \frac{g_{K,1}}{\rho} & \cdots & \frac{g_{K,K}}{\rho} \end{bmatrix}$$

and noticing that by construction $\sum_{j=1}^{K} \frac{g_{i,j}}{\rho} = 1$, for $i = (l+1), ..., K$, the result follows.

Finally, it is left to show that there are $(K-1-l)$ independent linear restrictions. It is obvious from (2.16) that the constraints are linear, and that there are $(K-1-l)$ nontrivial restrictions. Assume again, w.l.o.g., that the first l rows of G correspond to the zero rows. Therefore:

$$RPG\iota_{_K} = 0 \iff RG\iota_{_K} = 0 \iff \mathcal{R}G_{[K-l]}\iota_{_K} = 0 \tag{2.17}$$

where $\mathcal{R}$ is the $(K-1-l) \times (K-l)$ matrix defined in equation (2.8), $G_{[K-l]}$ is the $(K-l) \times K$ submatrix of G obtained by selecting only the the $K-l$ mixed rows of

G. By properties of the Kronecker product we can rewrite 2.17 as:

$$(\iota'_K \otimes \mathcal{R}) \operatorname{vec}(G_{[K-l]}) = 0$$

To show independence notice that this system is an homogeneous system of linear equations of the form $\Phi x = 0$, with $\Phi = (\iota'_K \otimes \mathcal{R})$ and $x = \operatorname{vec}(G_{[K-l]})$. This system has $K - 1 - l$ independent linear restrictions if the rank of Φ is $K - 1 - l$. From equation (2.8) it is clear that the rank of $\mathcal{R}$ and, therefore, of $(\iota'_K \otimes \mathcal{R})$ is $K - 1 - l$. The latter holds since by properties of the Kronecker product $rank(C_1 \otimes C_2) = rank(C_1)\, rank(C_2)$. Hence, the result follows. □

2.A.2 Proof of Theorem 2.1

We need to show that the matrix of constraints H applied to $\operatorname{vec}(G_{nz})$ is correctly defined. That is, it constrains the mixed rows of G to add up to the same quantity. We have that:

$$\begin{aligned} H &= \left\{ P_s \left[(\iota'_K \otimes \mathcal{R}_{nzr})' \odot \left(\operatorname{vec}(G^*_{nzr}) \iota'_{(nzr-1)} \right) \right] \right\}' \\ &= \left[(\iota'_K \otimes \mathcal{R}_{nzr}) \odot \left(\operatorname{vec}(G^*_{nzr}) \iota'_{(nzr-1)} \right)' \right] P'_s \end{aligned}$$

where $\mathcal{R}_{nzr}$ has dimensions $(nzr - 1) \times nzr$ and is defined as in (2.8), and P_s is a $nz \times (nzr\ ^*\ K)$ selection matrix. Using that $\operatorname{vec}(G_{nz}) = P_s \operatorname{vec}(G_{nzr})$ we have:

$$\begin{aligned} H \operatorname{vec}(G_{nz}) &= \left[(\iota'_K \otimes \mathcal{R}_{nzr}) \odot (\operatorname{vec}(G^*_{nzr}) \iota'_{(nzr-1)})' \right] P'_s P_s \operatorname{vec}(G_{nzr}) \\ &= \left[(\iota'_K \otimes \mathcal{R}_{nzr}) \odot (\operatorname{vec}(G^*_{nzr}) \iota'_{(nzr-1)})' \right] \operatorname{vec}(G_{nzr}) \end{aligned}$$

where I used that $P_s'P_s$ is equal to a modified $(nzr * K) \times (nzr * K)$ identity matrix such that the i^{th} element of its main diagonal is replaced by a zero if the i^{th} element of $\text{vec}(G_{nzr})$ is zero.

The next step makes use the following property of Hadamard product:

Lemma A1. *(See Horn and Johnson (1994, p.305)) Define the diagonal matrix, D_x, of size n with entries from a vector $x \in \mathrm{R}^n$ by*

$$[D_x]_{ij} = \begin{cases} [x]_i & if \;\; i = j \\ 0 & otherwise \end{cases}$$

Suppose A,B are $m \times n$ matrices. Then the i^{th} diagonal entry of the matrix AD_xB' coincides with the i^{th} entry of the vector $(A \cdot B)x$ for all $i = 1, 2, ..., m$. That is,

$$[AD_xB']_{ii} = [(A \odot B)x]_i \quad for \;\; all \;\; 1 \leq i \leq m$$

By Lemma A1,

$$\begin{aligned} &\left\{ [\iota_K' \otimes \mathcal{R}_{nzr}] \odot \left[\text{vec}(G_{nzr}^*) \; \iota_{(nzr-1)}' \right]' \right\} \text{vec}(G_{nzr}) \\ &= Diag \left\{ [\iota_K' \otimes \mathcal{R}_{nzr}] \;\; D_{\text{vec}(G_{nzr})} \left[\text{vec}(G_{nzr}^*) \; \iota_{(nzr-1)}' \right] \right\} \end{aligned} \tag{2.18}$$

Without loss of generality, assume that the first l rows of G are zero, with $0 \leq l \leq K$. Therefore, $nzr = K - l$ and we have:

$$\text{vec}(G_{nzr}) = \begin{bmatrix} g_{(l+1),1} & \cdots & g_{K,1} & \cdots\cdots & g_{(l+1),K} & \cdots & g_{(K-1),K} & g_{K,K} \end{bmatrix}' \tag{2.19}$$

where the elements corresponding to $g_{i,i}$ are equal to 0 for $i = (l+1), .., K$. Notice that if $l = K - 1$ there exist only one row that is not all zeros, hence the problem of specifying the restriction matrix becomes trivial. As a result, I consider $l < K - 1$

hereafter.

Using (2.19) we can write:

$$D_{\text{vec}(G_{nzr})} \times \left[\text{vec}(G^*_{nzr})\ \iota'_{(nzr-1)}\right] = \begin{bmatrix} g_{(l+1),1}\mathbb{1}_{\{g^*_{(l+1),1}\neq 0\}} & \cdots & g_{(l+1),1}\mathbb{1}_{\{g^*_{(l+1),1}\neq 0\}} \\ \vdots & \vdots & \vdots \\ g_{K,1}\mathbb{1}_{\{g^*_{K,1}\neq 0\}} & \cdots & g_{(K+1),1}\mathbb{1}_{\{g^*_{K,1}\neq 0\}} \\ \cdots\cdots\cdots & \cdot & \cdots\cdots\cdots \\ g_{(l+1),K}\mathbb{1}_{\{g^*_{(l+1),K}\neq 0\}} & \cdots & g_{(l+1),K}\mathbb{1}_{\{g^*_{(l+1),K}\neq 0\}} \\ \vdots & \vdots & \vdots \\ g_{K,K}\mathbb{1}_{\{g^*_{K,K}\neq 0\}} & \cdots & g_{K,K}\mathbb{1}_{\{g^*_{K,K}\neq 0\}} \end{bmatrix} \tag{2.20}$$

Next, notice that $[\iota'_K \otimes \mathcal{R}_{nzr}] = [\mathcal{R}_{nzr}|\cdots|\mathcal{R}_{nzr}]$ is a $(nzr-1)\times(nzr\ ^*\ K)$ matrix. Since all the columns of (2.20) are the same, pre-multiplying it by $[\iota'_K \otimes \mathcal{R}_{nzr}]$ gives a matrix with all its columns equal to:

$$[\mathcal{R}_{nzr}|\cdots|\mathcal{R}_{nzr}] \times \begin{bmatrix} g_{(l+1),1}\mathbb{1}_{\{g^*_{(l+1),1}\neq 0\}} \\ \vdots \\ g_{K,1}\mathbb{1}_{\{g^*_{K,1}\neq 0\}} \\ \cdots\cdots\cdots \\ g_{(l+1),K}\mathbb{1}_{\{g^*_{(l+1),K}\neq 0\}} \\ \vdots \\ g_{K,K}\mathbb{1}_{\{g^*_{K,K}\neq 0\}} \end{bmatrix}$$

$$= \begin{bmatrix} -\left[g_{(l+1),1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{(l+1),K}\mathbb{1}_{\{\cdot\}}\right] + \left[g_{(l+2),1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{(l+2),K}\mathbb{1}_{\{\cdot\}}\right] \\ \vdots \\ -\left[g_{(K-1),1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{(K-1),K}\mathbb{1}_{\{\cdot\}}\right] + \left[g_{K,1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{K,K}\mathbb{1}_{\{\cdot\}}\right] \end{bmatrix}$$

Finally, using that all the columns of $\left\{[\iota'_K \otimes \mathcal{R}_{nzr}]\ D_{\text{vec}(G_{nzr})}\left[\text{vec}(G^*_{nzr})\ \iota'_{(nzr-1)}\right]\right\}$ are

the same, we obtain that:

$$
\begin{aligned}
&Diag\left\{\left[\iota_K' \otimes \mathcal{R}_{nzr}\right]\ D_{\mathrm{vec}(G_{nzr})}\left[\mathrm{vec}(G^*_{nzr})\ \iota'_{(nzr-1)}\right]\right\} \\
&= \begin{bmatrix} -\left[g_{(l+1),1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{(l+1),K}\mathbb{1}_{\{\cdot\}}\right] + \left[g_{(l+2),1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{(l+2),K}\mathbb{1}_{\{\cdot\}}\right] \\ \vdots \\ -\left[g_{(K-1),1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{(K-1),K}\mathbb{1}_{\{\cdot\}}\right] + \left[g_{K,1}\mathbb{1}_{\{\cdot\}} + \cdots + g_{K,K}\mathbb{1}_{\{\cdot\}}\right] \end{bmatrix}
\end{aligned}
$$

hence the result follows.

□

2.A.3 Proof of Proposition 2.3

The Minimum Distance Estimator, $\mathrm{vec}(\hat{G}^c_{nz})$, can be derived from the Lagrangian by minimizing with respect to $\mathrm{vec}(G_{nz})$ and λ, and then solving for $\mathrm{vec}(G_{nz})$. The Lagrangian for the minimization problem in (2.10) is given by

$$
\mathcal{L} = \left(\mathrm{vec}(\hat{G}_{nz}) - \mathrm{vec}(G_{nz})\right)' \hat{V}_{nz}^{-1} \left(\mathrm{vec}(\hat{G}_{nz}) - \mathrm{vec}(G_{nz})\right) - \lambda' H\, \mathrm{vec}(G_{nz}) \tag{2.21}
$$

Differentiation with respect to $\mathrm{vec}(G_{nz})'$ and λ yields the first order conditions

$$
\frac{\partial \mathcal{L}}{\partial\, \mathrm{vec}(\hat{G}^c_{nz})}: \quad -2\hat{V}_{nz}^{-1}\, \mathrm{vec}(\hat{G}_{nz}) + 2\hat{V}_{nz}^{-1}\, \mathrm{vec}(\hat{G}^c_{nz}) - H'\lambda = 0 \tag{2.22}
$$

$$
\frac{\partial \mathcal{L}}{\partial \lambda}: \quad H\, \mathrm{vec}(\hat{G}^c_{nz}) = 0 \tag{2.23}
$$

Premultiplying (2.22) by $H\hat{V}_{nz}$ we have

$$
-2H\, \mathrm{vec}(\hat{G}_{nz}) + 2H\, \mathrm{vec}(\hat{G}^c_{nz}) - H\hat{V}_{nz}H'\lambda = 0
$$

Then solving for λ and using (2.23) gives

$$\lambda = -2(H\hat{V}_{nz}H')^{-1}H\,\text{vec}(\hat{G}^c_{nz})$$

Substituting back into (2.22) gives the solution for $\text{vec}(\hat{G}^c_{nz})$

$$\text{vec}(\hat{G}^c_{nz}) = \text{vec}(\hat{G}_{nz}) - \hat{V}_{nz}\,H'\left\{H\,\hat{V}_{nz}\,H'\right\}^{-1}H\,\text{vec}(\hat{G}_{nz}) \tag{2.24}$$

To derive the estimator of the asymptotic covariance matrix of $\text{vec}(\hat{G}^c_{nz})$, $\hat{V}^c_{nz}$, define $M = I_{nz} - \bar{P}$, with M an idempotent matrix and $\bar{P} = \hat{V}_{nz}\,H'\left\{H\,\hat{V}_{nz}\,H'\right\}^{-1}H$. Rewrite (2.24) as $\text{vec}(\hat{G}^c_{nz}) = M\,\text{vec}(\hat{G}_{nz})$. Then,

$$\begin{aligned}\hat{V}^c_{nz} &= M\hat{V}_{nz}M' = \hat{V}_{nz} - 2\hat{V}_{nz}\bar{P}' + \bar{P}\hat{V}_{nz}\bar{P}' = \hat{V}_{nz} - \hat{V}_{nz}H'\left\{H\,\hat{V}_{nz}\,H'\right\}^{-1}H\,\hat{V}_{nz} \\ &= M\hat{V}_{nz}\end{aligned} \tag{2.25}$$

where I have used the expression for $\bar{P}$. □

2.A.4 Proof of Proposition 2.4

From (2.12) we have that

$$\sqrt{T}\left(\text{vec}(\hat{G}_{nz}) - \text{vec}(\hat{G}^c_{nz})\right) = \hat{V}_{nz}\,H'\left\{H\,\hat{V}_{nz}\,H'\right\}^{-1}\sqrt{T}H\,\text{vec}(\hat{G}_{nz}) \tag{2.26}$$

We know that $\sqrt{T}\left(\text{vec}(\hat{G}_{nz}) - \text{vec}(G_{nz})\right) \xrightarrow{d} \mathcal{N}(0, V_{nz})$, then

$$\sqrt{T}\left(H\,\text{vec}(\hat{G}_{nz}) - 0\right) \xrightarrow{d} \mathcal{N}(0, HV_{nz}H') \tag{2.27}$$

Also, from equation (2.13) we have that

$$V_{nz} - V_{nz}^c = V_{nz} H' \{ H V_{nz} H' \}^{-1} H V_{nz} \tag{2.28}$$

Using the results from (2.27) and (2.28) in (2.26) we have that

$$\sqrt{T} \left(\text{vec}(\hat{G}_{nz}) - \text{vec}(\hat{G}_{nz}^c) \right) \xrightarrow{d} \mathcal{N}(0, V_{nz} - V_{nz}^c)$$

Recalling that a quadratic form of a normal distribution has a χ^2 distribution,

$$T \left(\text{vec}(\hat{G}_{nz}) - \text{vec}(\hat{G}_{nz}^c) \right)' [V_{nz} - V_{nz}^c]^- \left(\text{vec}(\hat{G}_{nz}) - \text{vec}(\hat{G}_{nz}^c) \right) \xrightarrow{d} \chi^2_{(nzr-1)}$$

where $(nzr - 1)$ is the number of restrictions on $\text{vec}\,(G)_{nz}$, and $[V_{nz} - V_{nz}^c]^-$ is a generalized inverse of $[V_{nz} - V_{nz}^c]$.

It is left to show that V_{nz}^{-1} is a generalized inverse of $[V_{nz} - V_{nz}^c]$. This amounts to show that if C^g is a generalized inverse of C then $C\, C^g\, C = C$:

$$[V_{nz} - V_{nz}^c]\ V_{nz}^{-1}\ [V_{nz} - V_{nz}^c] = [V_{nz} - V_{nz}^c] - V_{nz}^c + V_{nz}^c V_{nz}^{-1} V_{nz}^c$$

We need to show that the last two terms cancel out. Using the expression for V_{nz}^c in (2.13) we have:

$$\begin{aligned}
&V_{nz}^c V_{nz}^{-1} V_{nz}^c - V_{nz}^c \\
&= V_{nz}^c \left[V_{nz}^{-1} V_{nz}^c - I_{nz} \right] \\
&= \left[V_{nz} - V_{nz} H' \{ H V_{nz} H' \}^{-1} H V_{nz} \right] \left[I_{nz} - H' \{ H V_{nz} H' \}^{-1} H V_{nz} - I_{nz} \right] \\
&= -V_{nz} H' \{ H V_{nz} H' \}^{-1} H V_{nz} + V_{nz} H' \{ H V_{nz} H' \}^{-1} H V_{nz} H' \{ H V_{nz} H' \}^{-1} H V_{nz} \\
&= 0 \qquad \square
\end{aligned}$$

2.A.5 Steps of the PC-algorithm

I present here the steps of the PC-Algorithm as described in Spirtes et al. (2000, p.117-119).

The PC-Algorithm

1. Form the complete undirected graph $\mathcal{G}$ on the vertex set V.

2. $n = 0$.

   repeat

   repeat

   select an ordered pair of variables i and j that are adjacent in $\mathcal{G}$ such that $Adjacencies(\mathcal{G}, i)\backslash\{j\}$ has cardinality greater than or equal to n, and a subset S of $Adjacencies(\mathcal{G}, i)\backslash\{j\}$ of cardinality n, and if i and j are d-separated given S delete edge $i - j$ from $\mathcal{G}$ and record S in $Sepset(i, j)$ and $Sepset(j, i)$;

   until all ordered pairs of adjacent variables i and j such that $Adjacencies(\mathcal{G}, i)\backslash\{j\}$ has cardinality greater than or equal to n and all subsets S of $Adjacencies(\mathcal{G}, i)\backslash\{j\}$ of cardinality n have been tested for d-separation;

   $n = n + 1$;

   until for each ordered pair of adjacent vertices i, j, $Adjacencies(\mathcal{G}, i)\backslash\{j\}$ is of cardinality less than n.

3. For each triple of vertices i, j, k such that the pair i, j and the pair j, k are each adjacent in $\mathcal{G}$ but the pair i, k are not adjacent in $\mathcal{G}$, orient $i - j - k$ as $i \rightarrow j \leftarrow k$ if and only if j is not in $Sepset(i, k)$.

The PC-Algorithm *(continued)*

4. repeat

   If $i \to j$, j and k are adjacent, i and k are not adjacent, and there is no arrowhead at j, then orient $j - k$ as $j \to k$.

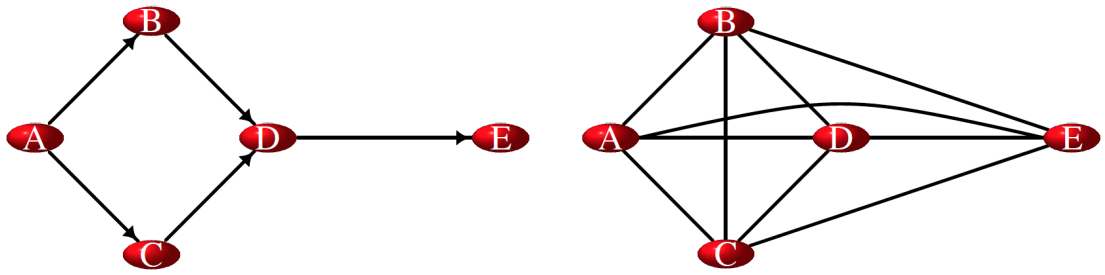
   If there is a directed path from i to j, and an edge between i and j, then orient $i - j$ as $i \to j$.

   until no more edges can be oriented.

Let $Adjacencies(\mathcal{G}, V)$ be the set of vertices adjacent to V in the directed acyclic graph $\mathcal{G}$. In the algorithm, the graph $\mathcal{G}$ is continually updated, so $Adjacencies(\mathcal{G}, V)$ is constantly changing as the algorithm progresses.

The next example, based on five nodes, provides further details on each step.

Example 2.3. *(Based on Spirtes, Glymour, and Scheines (2000, p.118)[35]) Assume we have 5 nodes (or variables) A, B, C, D, E. The true graph (i.e., DGP) is given in Figure 2.5a.*



(a) Unknown True DGP (b) Complete undirected graph

Figure 2.5: PC-Algorithm

[35]This example was adopted from Richardson (2012)

1. ***Complete undirected graph:*** *The starting point is to form a complete undirected graph $\mathcal{G}$ on the vertex set V as shown in Figure 2.5b. The complete undirected graph shows an undirected edge between every variable of the system, i.e., every variable in V is connected with lines having no arrows.*

2. ***Remove edges:*** *Edges between variables are removed sequentially based on zero correlation or partial correlation (conditional correlation). Each edge is subjected to tests that the correlation between its endpoints is zero. For instance, we first test $H_0: \rho_{ij} = 0$, where ρ_{ij} is the population (unconditional) correlation between nodes i and j. If a correlation is judged to be not different from zero, we remove the edge between the two end points of the corresponding edge. Edges surviving these unconditional correlation tests are then subjected to conditional correlation tests, e.g., $H_0: \rho_{ij.k} = 0$, where $\rho_{ij.k}$ is the population correlation between i and j conditional on variable k. If these conditional correlations equal zero we pick up the edge i, j.*[36]

   This stage makes use of Fisher's z statistic, "d-separation" concept, and gives rise to the concept of "sepset" (i.e., separation set) which is used in the next stage for orientation purposes:

   (a) ***Statistic:*** *Fisher's z-transformation is used to test for significance from zero, i.e., $\rho_{ij.S} = 0$:*

$$z(\rho_{ij.\boldsymbol{S}}, T) = \frac{1}{2} \log \left(\frac{|1 + \rho_{ij.\boldsymbol{S}}|}{|1 - \rho_{ij.\boldsymbol{S}}|} \right)$$

   where $\rho_{ij.\boldsymbol{S}}$ is the population correlation between i and j given S. Let $|S|$ equal the number of variables in S. If the variables i, j, and S are normally

[36] The PC-algorithm adds some sophistication to this step in order to increase tractability and reduce the curse of dimensionality. In other words, the algorithm is done such that it avoids testing every single conditional correlation unnecessarily. The details of this sophistication are given in the description of the algorithm.

distributed then[37]

$$\sqrt{T - |\boldsymbol{S}| - 3}\,(\hat{z} - z) \overset{a}{\sim} \mathcal{N}(0, 1)$$

(b) ***d-separation:*** *A non-endpoint vertex V on a path between i and j is said to be a collider if the path takes the form $i \cdots \rightarrow V \leftarrow \cdots j$. A path π is said to d-connect i and j unconditionally if i and j are at the endpoints of π and there are no colliders on π. If there is no path d-connecting i and j then they are said to be d-separated.*

(c) ***Sepset:*** *The conditioning variable(s) on removed edges between two variables is called the sepset of the variables whose edge has been removed. For vanishing zero order conditioning information the sepset is the empty set.*

- *If we remove the edge between i and j through unconditional correlation test, $\rho_{ij} = 0$, then the $Sepset(i, j)$ is $\{\}$, and we denote it as $S_{ij} = \emptyset$.*
- *If we remove this edge by conditioning on k, $\rho_{ij.k} = 0$, then the $Sepset(i, j)$ is k, and we denote it as $S_{ij} = k$.*

In the current example we check dependencies of degree $\mathscr{D} = 0, 1, 2$ since there are no dependencies of degree 3 or more:

- *$\mathscr{D} = 0$ (Zero order independencies): There is no pair of variables d-separated given the empty set, so the initial graph is unchanged.*
- *$\mathscr{D} = 1$ (First order independencies): We have that B and C are d-separated given $\{A\}$, thus we remove the $B - C$ edge and record the sepset of BC as $S_{BC} = \{A\}$. Next, since A and E are d-separated given $\{D\}$ we remove the $A - E$ edge and record $S_{AE} = \{D\}$. Similarly, since B and E are d-separated given $\{D\}$ we remove the $B - E$ edge and write $S_{BE} = \{D\}$.*

[37] See Spirtes et al. (2000, p.128) for further details.

Finally, we have that C and E are d-separated given $\{D\}$, thus, we remove the $C - E$ edge and record $S_{CE} = \{D\}$. In summary we have the following independencies of first order:

$$B \perp\!\!\!\perp C \mid A \quad A \perp\!\!\!\perp E \mid D$$
$$B \perp\!\!\!\perp E \mid D \quad C \perp\!\!\!\perp E \mid D$$

- $\mathscr{D} = 2$ *(Second order independencies): We have only one second order independency. Namely, A and D are d-separated given B, C, thus, we remove edge $A - D$ and the record its sepset as given by $S_{AD} = \{B, C\}$. Formally, $A \perp\!\!\!\perp D \mid \{B, C\}$.*

This completes the elimination stage, and Figure 2.5b reduces to Figure 2.6. This last figure gives the skeleton of the graph.

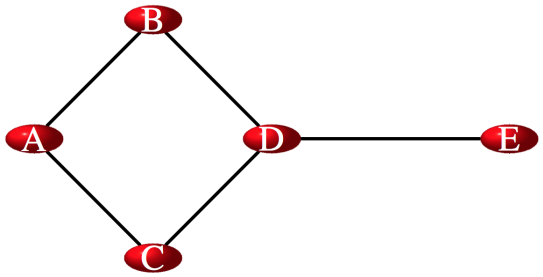


Figure 2.6: Skeleton PC-Algorithm

3. ***Edge Direction:*** *Edges are directed by considering triples, the following rules apply:*

   (a) *If we have $i - k - j$ but i and j are not adjacent, and $k \notin S_{ij}$ then orient as $i \to k \leftarrow j$.*

   (b) *If there is a directed path from i to k, and an edge between k and j, e.g., $i \to k - j$, then direct $k - j$ as $k \to j$.*

(c) *If* $i \to k \to j$ *and there is an edge* $i - j$*, then orient it as* $i \to j$*.*

(d) *If* i *and* k *are not adjacent, but* $i - l - k$ *and* $i \to j \leftarrow k$ *and* $l - j$*, then direct this last edge as* $l \to j$*.*

In our example, since $D \notin S_{BC} = \{A\}$*, we orient* $B \to D \leftarrow C$*. The other triples* (B, A, C)*,* (A, B, D)*,* (A, C, D)*,* (B, D, E) *and* (C, D, E) *do not lead to further orientation, since the middle vertex is in each sepset. Since* (B, D, E) *is such that* $B \to D - E$*, then we orient* $D - E$ *as* $D \to E$*. The resulting graph from the PC-algorithm is shown in Figure 2.7. Notice that the bidirected edges* $A \leftrightarrow B$ *and* $A \leftrightarrow C$ *simply represent edges that were not oriented (i.e., these are undirected edges).*

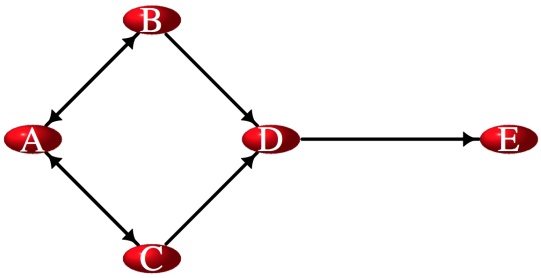


Figure 2.7: CPDAG after PC-Algorithm

2.A.6 Measures of performance for the PC-algorithm

Each element of the network matrix can take two possible values: 0 or 1. This leads to two possible outcomes, positive (predicting that there is a connection, i.e., a value of 1), or negative (predicting that there is no connection, i.e., a value of 0). The elements in the network matrix obtained from each simulation may or may not match those of the network from the DGP. There are four possible cases to be considered:

i. True positive (TP): 1 is identified as 1

ii. False positive (FP): 0 incorrectly identified as 1

iii. True negative (TN): 0 correctly identified as 0

iv. False negative (FN): 1 incorrectly identified as 0

Based on this classification we can defined the following measures:

i. Accuracy: $ACC = (TP + TN)/(P + N)$

ii. True Positive Rate (or Sensitivity): $TPR = TP/P$

iii. Specificity: $SPC = TN/N$

where P and N stand for the number of positives and negatives in the DGP respectively.

While ACC provides an overall assessment of how well the PC-algorithm performs at returning the true network structure (both zeros and ones), TPR and SPC focus on correctly detecting a connection and a no-connection respectively. In a given simulation, ACC provides the proportion of true results (both true connections and true no-connections) among the total number of elements in the network delivered by the simulation. An accuracy of 100% means that the simulated network is exactly the same as the one given by the DGP. In contrast, TPR focus exclusively on whether we are correctly detecting a connection where there should be one. That is, what proportion of the ones in the DGP are also present in a given simulation. As an example, a TPR of 60% means that 60% of the links in the DGP are also present in the network produced by a given simulation. It is important to keep in mind that even if TPR is 100% there could be other links in the simulated network that come from false positives.[38] SPC is the true negative rate (TNR), it gives what proportion of the zeros found in a simulation correspond to true zeros in the DGP. Notice that

[38]To address this concern, we could also look at the Positive Predictive Value (PPV) which measures the proportion of positives found in a given simulation that are in fact true positives. The difference between the TPR and the PPV lies in the denominator of the rate: while the former divides by the number of ones in the DGP, the latter divides by the number of ones in the network delivered by the simulation.

this rate is the complement of the False Positive Rate (FPR), i.e., $FPR = 1 - SPC$. Moreover, from the above formulas, it is clear that ACC is a weighted average of TPR and SPC with weights related to the prevalence rate (i.e., $P/(P+N)$):

$$ACC = TPR \times \text{prevalence} + SPC \times (1 - \text{prevalence})$$

Finally, the Structural Hamming Distance (SHD) is a metric that directly compares the structure of the learned and the original networks. The SHD between two CPDAGs is the number of the following operators required to make the CPDAGs match: add or delete an undirected edge, and add, remove, or reverse the orientation of an edge (see Tsamardinos, Brown, and Aliferis (2006) for further details). It can be interpreted as the minimum number of operations needed to go from the graph delivered by a simulation to the true graph. This measure is usually used to assess the overall quality of fit which requires to focus simultaneously on the TPR and SPC (or FPR). A large SHD suggests a poor fit, while a small SHD suggests a good fit.

2.A.7 Network terminology

In simple terms, a network is a graph which is comprised of an arrangement of nodes, each representing an individual in the network, joined by a set of lines or connecting arrows, that depict the relationships among these individuals. Formally,

Definition 2.3. *(Graphical model) A graphical model, $\mathcal{G} = (V, E)$, is a system of nodes (or vertices) $V = \{1, ..., K\}$ and connecting edges (or lines) $E \subseteq V \times V$, where K is the number of variables or nodes in the system.*

A graph can be classified as either undirected, i.e., when the edges have no orientation, or directed, when the edges are arrows indicating the direction of the relation. Based on these notions, two main types of graphs can be defined; namely, undirected and directed graphs. A special case of directed graphs corresponds to the so called

Directed Acyclic Graph (DAG), which has gained attention in the SVAR literature as the graphical representation of a recursive system is given by a DAG. Hence, in this paper I particularly focus on DAGs. Moreover, I adopt the following convention, from DAGs, regarding edge definitions. An edge $(i,j) \in E$ is called directed if $(i,j) \in E$ but $(j,i) \notin E$ for some $i,j \in V$. If both $(i,j) \in E$ and $(j,i) \in E$, the edge is called undirected.[39] In the former case, the directed edge is denoted as $i \to j$, while in the latter case the undirected edge is denoted as $i - j$.

Formal definitions of both undirected and DAG graphs are given next. An undirected graph, $\mathcal{G} = (V,E)$, consists of the set V of nodes and E of edges, which are unordered pairs of elements of V. Formally,

Definition 2.4. *(Undirected graph) A graph $\mathcal{G} = (V,E)$ is called undirected if $\forall i,j \in V$, $(i,j) \in E$ if and only if $(j,i) \in E$.*

The definition of Directed Acyclic Graph requires the following two concepts:

Definition 2.5. *(Directed cycle) A graph $\mathcal{G} = (V,E)$ has a directed cycle, if $\exists$ a sequence of directed edges $\{(i_1,i_2),(i_2,i_3),\cdots,(i_{K-1},i_K)\}$, such that $(i_k,i_{k+1}) \in E$ for each $k \in \{1,\cdots,K-1\}$, with $i_1 = i_K$.*

A directed graph, $\mathcal{G} = (V,E)$, consists of the set V of nodes and E of edges, which are ordered pairs of elements of V. The convention adopted in this paper is that all edges need to be directed to have a directed graph. If only some edges are directed, then the graph is called partially directed. Formally,

Definition 2.6. *(Directed graph) A graph $\mathcal{G} = (V,E)$ is directed if $\forall i,j \in V$, if $(i,j) \in E$ then $(j,i) \notin E$.*

Now we are condition to define a DAG:

[39] This definition of undirected edges is specific to DAGs since it assumes that a bidirected edge is an edge with no direction. This may not be the convention adopted in other contexts where bidirected edges and undirected edges need to be distinguished.

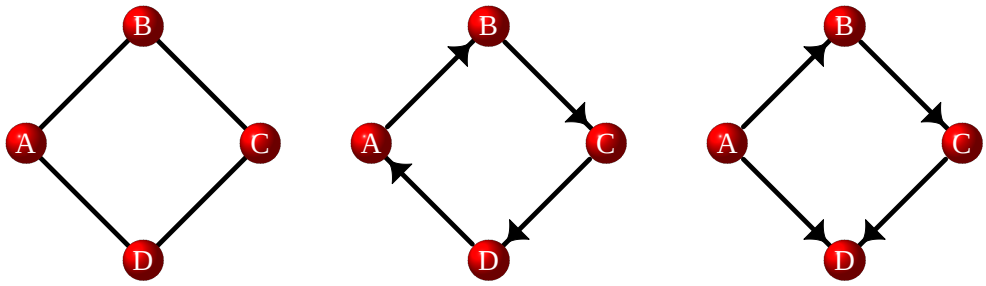
Definition 2.7. *(Directed Acyclic Graph) A graph $\mathcal{G} = (V, E)$ is called a Directed Acyclic Graph (DAG) if all edges are directed and there are no directed cycles.*

A related concept is the notion of skeleton of a DAG,

Definition 2.8. *(Sketelon of a DAG) The graph generated by replacing all directed edges of a DAG with undirected edges is called a skeleton.*

This concept has shown to be useful whenever we want to focus on the structure of the network regardless of the direction of connectivity. For instance, we will look at the skeleton to study the presence of cohesive-blocks within the network.

Figure 2.8 below illustrates the aforementioned graph concepts. Notice that the skeleton of the DAG shown in (c) coincides with the undirected graph given in (a).



(a) Undirected Graph (b) Directed Graph (with a cycle) (c) Directed Acyclic Graph

Figure 2.8: Examples of Graphs

Finally, I define the concept of a path in a network as many of the connectedness measures presented in the next section invoke it.

Definition 2.9. *(Path) A path in a graph $\mathcal{G} = (V, E)$ between nodes i and j is a sequence of edges $i_1i_2, i_2i_3, \cdots, i_{K-1}i_K$ such that $i_ki_{k+1} \in E$ for each $k \in \{1, 2, \cdots, K-1\}$, with $i_1 = i$ and $i_K = j$ and such that each node in the sequence $\{i_1, \cdots, i_K\}$ is distinct.*

2.A.8 Graphical representation of the network

It has become a common practice in the financial networks literature to graph the network using a circle layout. Circle graphs are commonly used to visualize which nodes are most highly connected, their emphasis is on rankings. The nodes are located at equal distances around a circle, and nodes that are highly connected are very easy to quickly locate because of the density of lines. Moreover, it is convenient for comparing graphs since node's positions can be easily fixed. However, this representation has an important shortcoming, it fails at revealing the underlying structure of the network. For instance, a circle layout does not allow one to visually detect communities (or blocks), nor bridges or central nodes.

In this paper, I propose the use of the Fruchterman-Reingold layout as an alternative for analyzing either one network or when the number of graphs to be compared is relatively small. The Fruchterman-Reingold algorithm is a force-based graph layout algorithm. Force-based means that it treats each vertex and edge as if it were a physical object whose position is influenced by forces around it. It is based on the spring algorithm, which introduces attraction forces between connected nodes and repulsion forces between disconnected nodes. All edges are treated as springs, so that the network will oscillate until a minimum force between nodes is reached. The force-directed graph drawing algorithms is a class of algorithms for drawing graphs in an aesthetically pleasing way. Their purpose is to position the nodes of a graph so that all the edges are of more or less equal length and there are as few crossing edges as possible. In simple words, it assigns locations to nodes such that nodes that are "more similar" are closer together. For instance, two nodes are "similar" to the extent that they have similar shortest paths to all other nodes.

2.B Additional tables and graphs

2.B.1 Other simulation exercises

2.B.1.1 $K = 16$

It is well known in the VAR literature that "... standard VARs rarely employ more than six to eight variables" (Bernanke, Boivin, and Eliasz, 2005, p.388). Since the number of parameters in a VAR increases with the square of the number of variables included, as the number of variables increases the estimates become noisier. One possible solution to address the concern of using a relatively large number of variables, i.e., K=16, is to estimate the reduced form VAR by Least Absolute Shrinkage and Selection operator (LASSO-VAR). Sparsity in the reduced form model allows us to delete weak and indirect links among variables that would lead to spurious relations in the reduced form; hence, it delivers VAR estimates that are more reliable (see Davis, Zang, and Zheng (2012)).

The conjecture in this paper is that LASSO-VAR techniques are potentially useful in two dimensions. First, it provides less noisy estimates of the variance-covariance matrix of the residuals, which is the input of the PC-algorithm. In other words, the conjecture is that a better input could help pin down the connections in the network and their directions better. Second, LASSO-VAR provides more precise estimates of the reduced form lagged coefficients, which are used to estimate the structural lagged coefficients. Although there is evidence in the literature that LASSO-VAR provides more precise reduced form lagged coefficients, it is unclear whether it could be beneficial for the purpose of uncovering the network structure via the PC-algorithm.

The first simulation exercise compares the performance of the PC-algorithm for two choices of reduced form estimation methods: LASSO-VAR versus standard VAR estimation. It is clear from Figure 2.9 that LASSO-VAR performance to uncover the network is almost identical to standard VAR, mitigating the second potential benefit.

A plausible explanation for this result is that the benefits of LASSO in the reduced form do not translate well due to the presence of nonlinearities when going from the VAR to the SVAR. Another possibility is that even $K = 16$ may not be big enough to capture the advantages of using LASSO; but this is beyond the scope of this paper. Since LASSO-VAR at least improves in one dimension, I adopt this method in the paper.

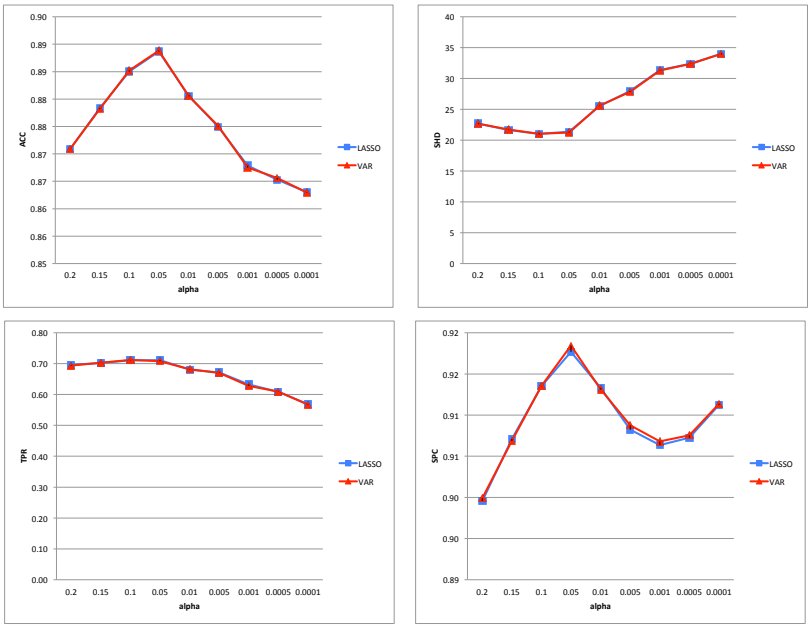


Figure 2.9: Comparison of LASSO-VAR and VAR estimation methods in the reduced form across values of α, $s = 500$. DGP generated using MajRSC option, $\alpha = 5\%$, $K = 16$, and $p = 1$. Mean ACC, TPR, SPC, and SHD reported.

The second simulation exercise compares the MajRSC option to an alternative option for calling the PC-algorithm in R. Sometimes "sampling errors, non faithfulness, or hidden variables can also lead to non-extendable CPDAGs, meaning that there does not exist a DAG that has the same skeleton and v-structures as the graph found by the algorithm. An example of this is an undirected cycle consisting of the edges a-bc-d and d-a. In this case it is impossible to direct the edges without creating a cycle or a new v-structure." (Kalisch et al., 2012). In this situation one can use the option "u2pd = retry" (Retry option hereafter), then, up to 100 combinations of

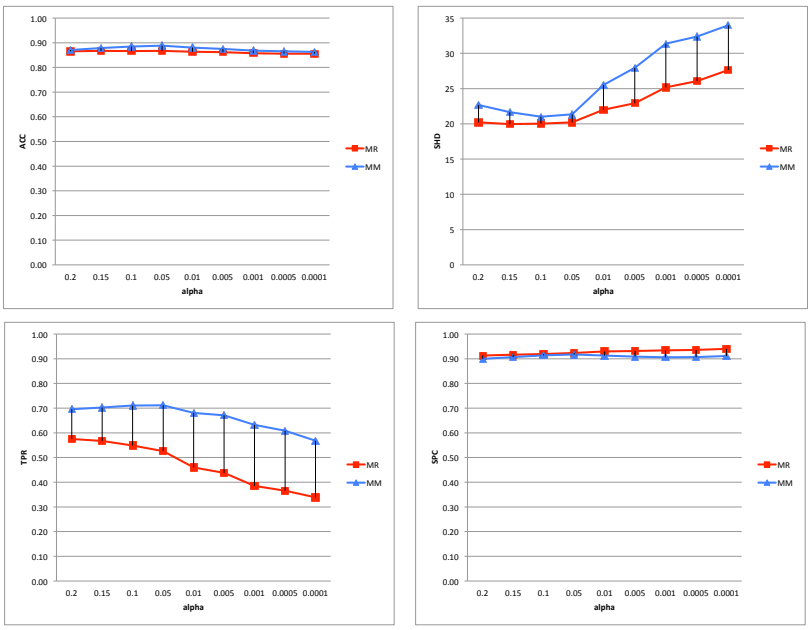


Figure 2.10: Comparison of MajRSC (MM) and Retry (MR) options to implement the PC-algorithm across values of α, $s = 500$. DGP generated using MajRSC option, $\alpha = 5\%$, $K = 16$, and $p = 1$. Mean ACC, TPR, SPC, and SHD reported.

possible directions of the ambiguous edges are tried, and the first combination that results in an extendable CPDAG is chosen.[40] In the simulation exercise, I compare the performance of the PC-algorithm using the MajRSC option versus the Retry option. Furthermore, I construct a DGP with each option and conduct simulations using Retry vs MajRSC in each case. The notation adopted in all the graphs (i.e., MM, MR, RM, RR) uses the first letter to refer to the DGP and the second letter to refer to the option used in the simulations.

Figure 2.10 displays the simulation results that uses the MajRSC option for the DGP (with $\alpha = 5\%$). Both MajRSC and Retry options perform quite similar regarding ACC and SPC. Although performance in terms of SHD is quite similar only for values of α between 0.2 and 0.05, these are the values of α that deliver the best (i.e., lower) values of SHD. However, the MasjRSC option dominates in terms of TPR. The gap between the two options is of at least ten percentage points, with a TPR in

[40] If no valid combination is found, an arbitrary DAG is generated on the skeleton as in the option "rand", and then converted into its CPDAG.

the MajRSC option of approximately 70% for values of α between 0.2 and 0.05 and a maximum at $\alpha = 0.05$. All the measures achieve their best values at $\alpha = 0.05$, which coincides with the α chosen to obtain the calibrated DGP, although values between 0.2 and 0.05 work almost as well. To assess whether it is always the case that the best α coincides with the one chosen to calibrate the DGP, I also constructed DGPs with $\alpha = \{0.10, 0.0001\}$ and repeated the simulation exercise. The results are presented in figures 2.11 and 2.12, and they also suggest the use of α equal to 0.05 setting aside the former concern.

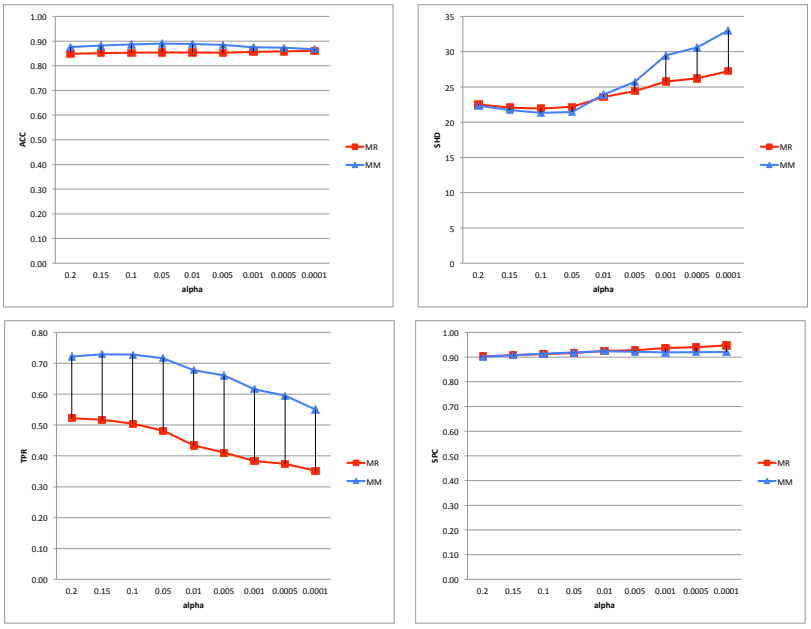


Figure 2.11: Comparison of MajRSC (MM) and Retry (MR) options to implement the PC-algorithm across values of α, $s = 500$. DGP generated using MajRSC option, $\alpha = 10\%$, $K = 16$, and $p = 1$. Mean ACC, TPR, SPC, and SHD reported.

Finally, in Figure 2.13 the DGP is instead constructed using the Retry option. In the simulations, the performance of both options in terms of ACC and SPC is again quite similar, while the Retry option does now a better job in terms of SHD. Since the DGP is constructed with the Retry option, one would expect this option to perform better now in terms of TPR. Even though the TPR gap shrank for values of α between 0.2 and 0.05, a surprising result is that the MajRSC option still dominates the Retry

option in terms of TPR. Summing up, both figures suggest the use of MajRSC and to choose α of 0.10 or 0.05.

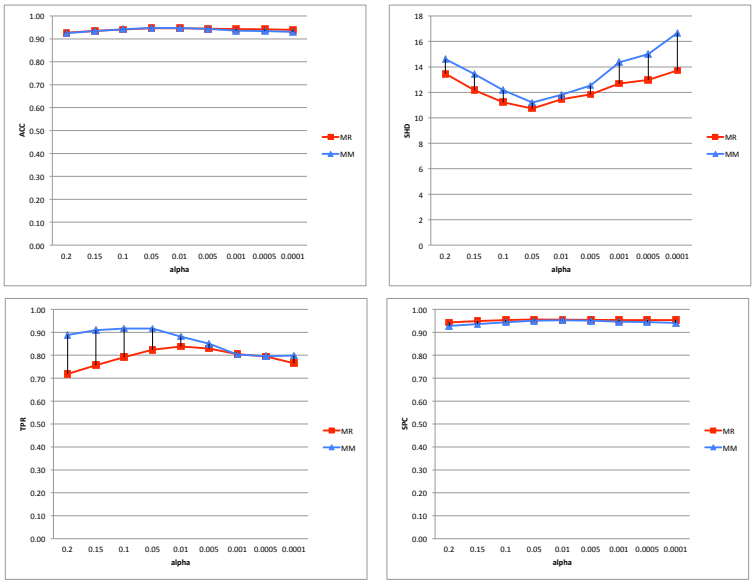


Figure 2.12: Comparison of MajRSC (MM) and Retry (MR) options to implement the PC-algorithm across values of α, $s = 500$. DGP generated using MajRSC option, $\alpha = 0.01\%$, $K = 16$, and $p = 1$. Mean ACC, TPR, SPC, and SHD reported.

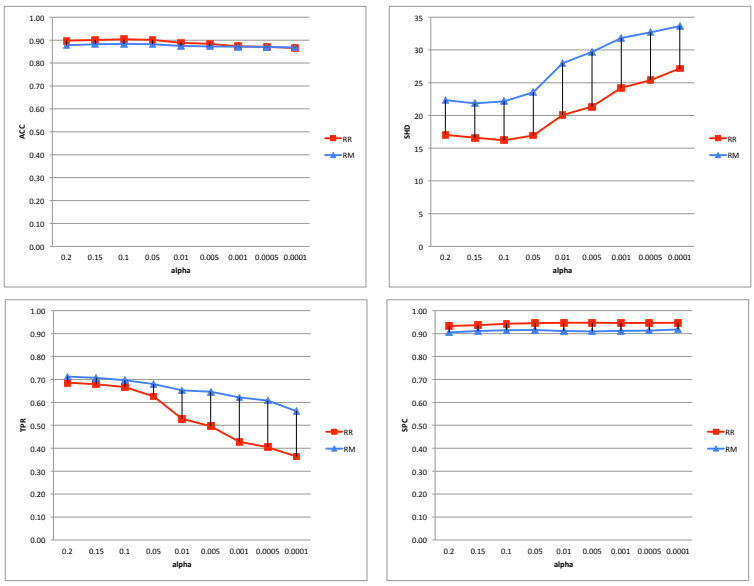


Figure 2.13: Comparison of MajRSC (RM) and Retry (RR) options to implement the PC-algorithm across values of α, $s = 500$. DGP generated using Retry option, $\alpha = 5\%$, and $K = 16$, and $p = 1$. Mean ACC, TPR, SPC, and SHD reported.

2.B.1.2 $K = 8$

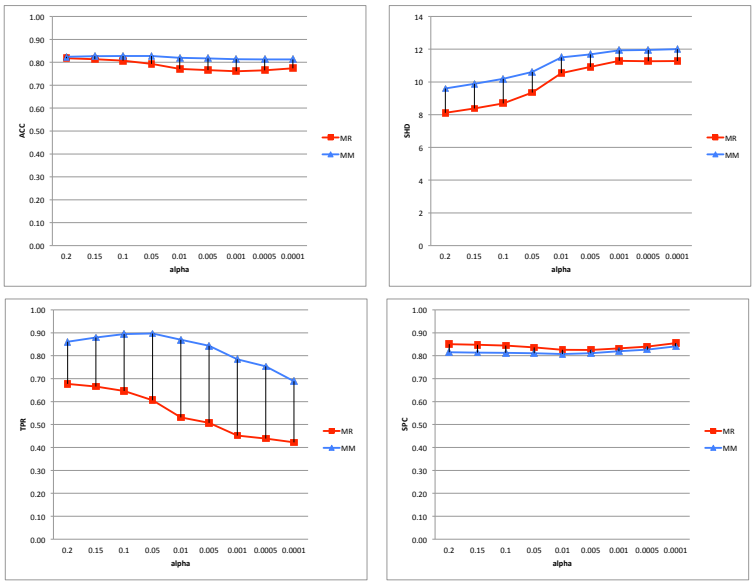


Figure 2.14: Comparison of MajRSC and Retry options to implement the PC-algorithm across values of α, $s = 500$. DGP generated using MajRSC option, $\alpha = 5\%$, $N = 8$, and and $p = 2$. Mean ACC, TPR, SPC, and SHD reported.

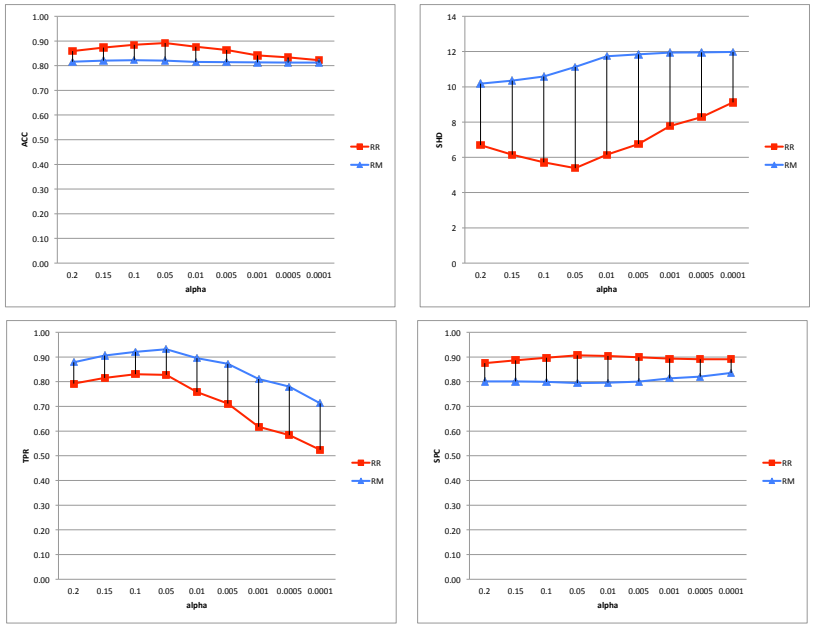


Figure 2.15: Comparison of MajRSC and Retry options to implement the PC-algorithm across values of α, $s = 500$. DGP generated using Retry option, $\alpha = 5\%$, $N = 8$, and and $p = 2$. Mean ACC, TPR, SPC, and SHD reported.

2.B.2 Application

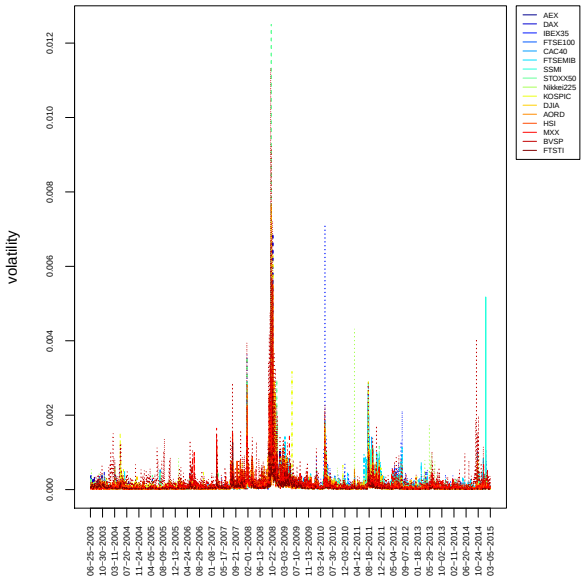


Figure 2.16: Time series plot of daily realized return volatility - Full period (06/25/2003 - 03/05/2015), K=16.

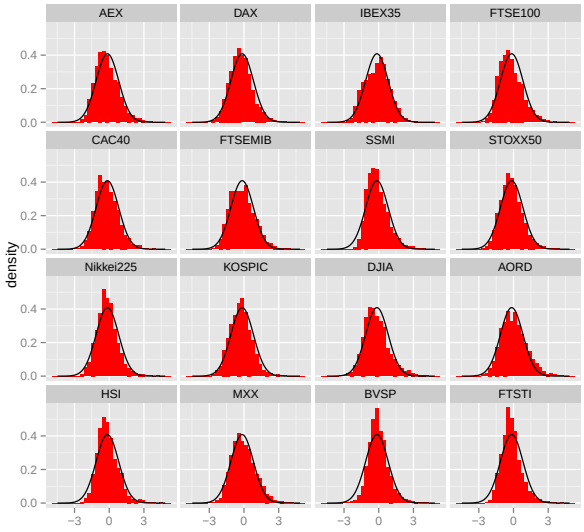


Figure 2.17: Distribution of daily realized return log-volatility (demeaned data) - Full period (06/25/2003 - 03/05/2015), K=16.

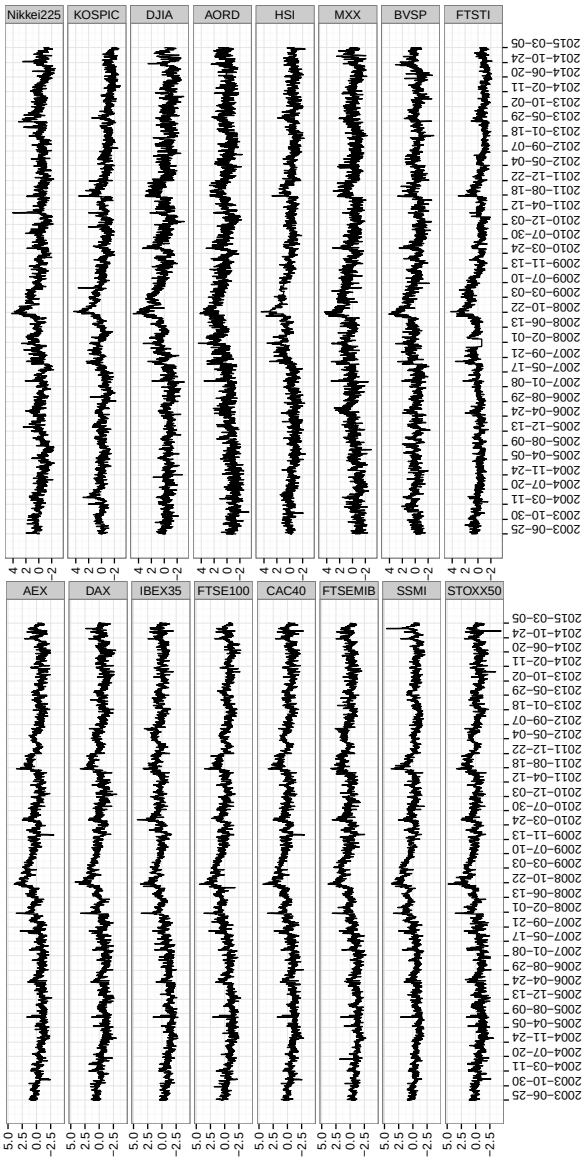
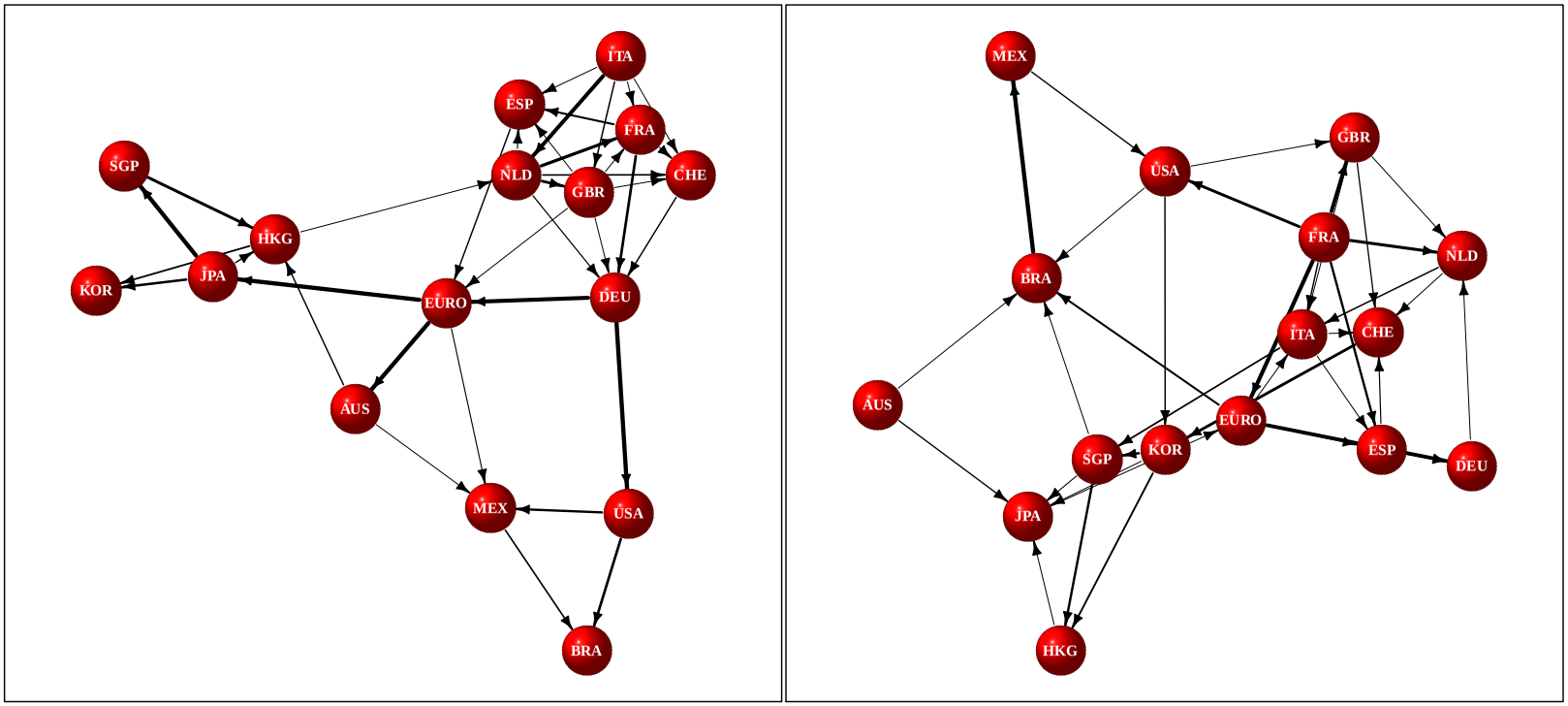
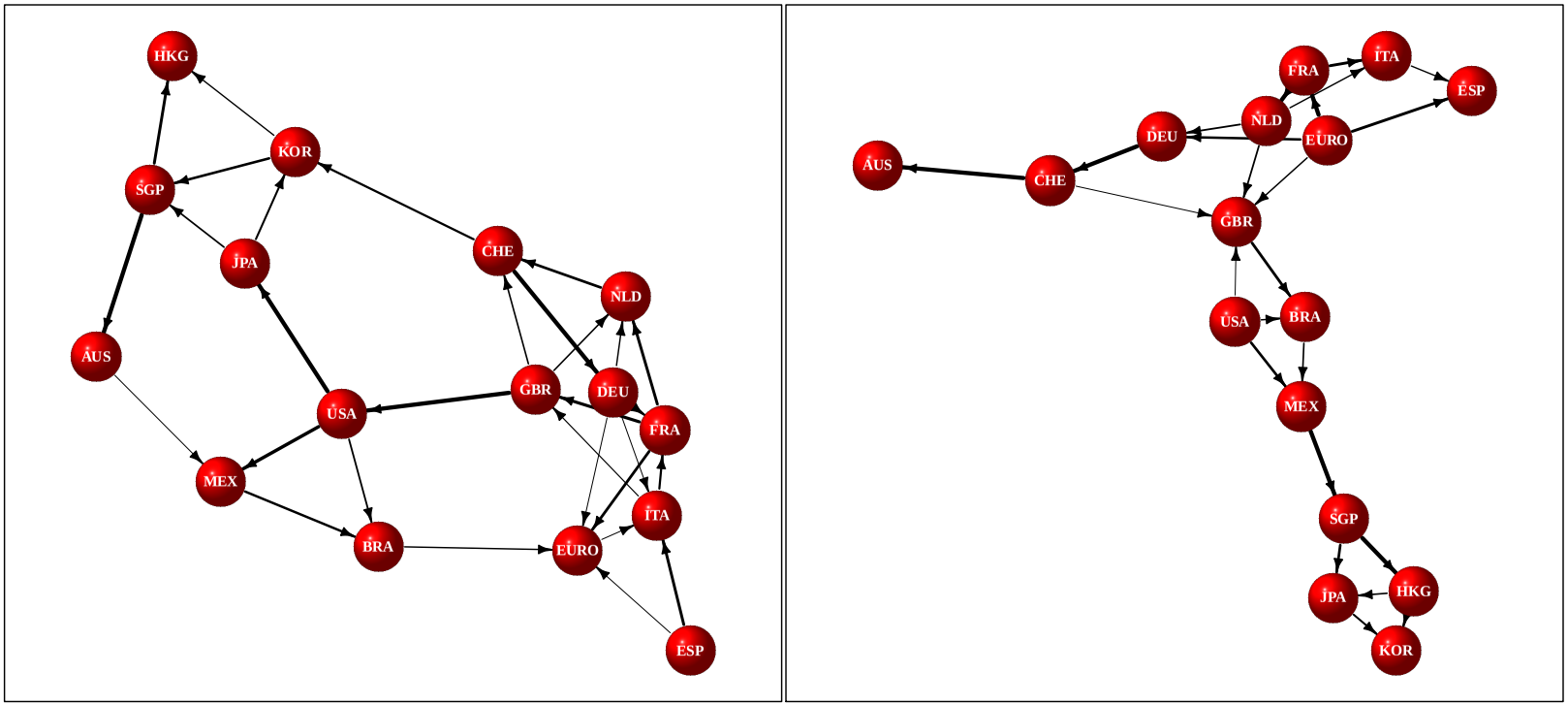


Figure 2.18: Time series plot of demeaned daily realized return log-volatility - Full period, K=16.



(a) Pre-crisis period (June 25, 2003 - March 1, 2007)

(b) Crisis period (March 2, 2007 - February 26, 2010)



(c) Post-crisis period (March 1, 2010 - December 28, 2012) (d) Euro-crisis period (January 2, 2013 - March 5, 2015)

Figure 2.20: Fruchterman-Reingold representation of the Network for each period. The sample is from June 25th, 2003 to March 5th, 2015, K=16. The analysis is divided in four periods: pre-crisis period (06/25/2003 - 03/01/2007), crisis period (03/02/2007 - 02/26/2010), post-crisis period (03/01/2010 - 12/28/2012), and euro-crisis period (01/02/2013 - 03/05/2015). Arrows widths are proportional to the strength of connection.

Chapter 3

GMM with Minimum Mean Square Error

3.1 Introduction

The theory of Generalized Method of Moments (GMM) has grown widely since Hansen's (1982) seminal paper. Among the contributions to this literature, we focus on redundancy. It is well known in the GMM estimation literature that incorporating an additional set of moment conditions to the initial set will not hurt asymptotically the GMM estimator in terms of efficiency.[1] However, even in that case, it is also well known from Breusch, Qian, Schmidt, and Wyhowski (1999) that these extra moments may not help improve efficiency either. When that is true, we say that these extra moments are *redundant* for the estimation of the full set of parameters, given the initial set of moment conditions. Similarly, the addition of extra moments may not help increase efficiency for only some of the parameters, in which case we say that they are *partially redundant* for the estimation of those parameters given the original set of moments.

[1] By "will not hurt asymptotically in terms of efficiency" we mean that it will not decrease efficiency, or, in other words, it will not increase the asymptotic variance.

The notions of redundancy and partial redundancy were first formally introduced by Breusch et al. (1999); in a later paper, Qian (2002) revisited the notion of partial redundancy more thoroughly. In this paper we revisit this literature and extend it to a more general setting. More specifically, we consider moments that are unbiased, i.e., equal to zero, but whose estimators are asymptotically biased. Think, for instance, of the "Kernel Moment Estimator" of Gagliardini, Gouriéroux, and Renault (2011). As a result, we can no longer talk about redundancy in terms of asymptotic variance since the asymptotic bias also comes into play. Therefore, we need a broader definition of redundancy and partial redundancy.

The contribution of this paper is twofold. First, we reassess what an optimal weighting matrix for a GMM estimator would be under the presence of asymptotic bias in the estimator of the moment conditions. Second, based on this new definition of optimal weighting matrix, we derive and reinterpret the necessary and sufficient conditions for redundancy and partial redundancy, which can now be interpreted in terms of the bias-variance trade off. The methodology we propose in this paper, though quite natural in both cases, is not obvious at first glance. For re-examining the definition of optimal weighting matrix we follow the ideas in the paper by Gagliardini et al. (2011), entitled "Efficient derivative pricing by the extended method of moments." This is one of the most related work in the extant literature. While their approach put together parametric and non-parametric rates of convergence, here we consider only parametric rates for sake of notational simplicity. However, an induced application of the methodology developed in this paper, still work in progress, is about misspecified asset pricing models. In regard to redundancy and partial redundancy, we follow quite closely the paper by Breusch et al. (1999). This leads us to discuss focused moment selection in the spirit of DiTraglia (2015), but the approach is different since we always consider AMSE-efficient GMM estimators. Another related work, is the paper by Cheng and Liao (2015). However, their objective is slightly differ-

ent. Although they do consider relevance, their aim is to avoid including redundant moment conditions after consistently eliminating invalid ones.

Notice that the optimal GMM weighting matrix, say W, is set in general such that it minimizes the asymptotic variance of the parameter's estimator, say $\ddot{\theta}_T$, of θ. This leads to setting $W = \Omega^{-1}$, where Ω is the asymptotic variance of the estimator of the moments conditions, call it $\bar{g}$. However, in the presence of asymptotic bias the natural choice for W will be the one that minimizes the Asymptotic Mean Square Error (AMSE) instead of asymptotic variance only. Intuitively, if we were to ignore the asymptotic bias, we would be obtaining a more precise estimator around a value that is not the true value θ_0. As a result, that choice of weighting matrix would not be adequate. Therefore, we propose to choose a weighting matrix that accounts for both asymptotic bias and variance. Namely, to choose $W = M^{-1}$, where M is the AMSE of the estimator of the moment conditions.

It is thus natural to redefine both redundancy and partial redundancy in terms of AMSE. That is, the aim is now to assess whether the extra moments help reduce the initial AMSE of the whole set of parameters or a subset of it, respectively. For simplicity of exposition, assume that the initial moment conditions are based on a function g_1, i.e., we have $E[g_1(\cdot;\theta_0)] = 0$, while the additional moments are based on g_2, with $g = (g_1' \quad g_2')'$. In this paper, we show that adding extra moments cannot hurt in terms of AMSE. Hence, redundancy and partial redundancy conditions can be derived by looking directly at the difference between two AMSEs: the one obtained when we use the initial set of moments and that obtained when we use the full set of moments. However, following Breusch et al. (1999), we know this direct approach can be algebraically demanding and hard to interpret.

Breusch et al. (1999) propose instead to exploit the idea that the GMM estimator derived from moments based on g_1 and g_2 is numerically equal, in general, to another GMM estimator based on moments from g_1 and the residual of a projection of g_2

on g_1. This approach was very useful in their paper to the extent that it made the variance-covariance matrix of the moment estimators block-diagonal. This was key for them since it allowed the delivery of an expression for the asymptotic variance directly comparable to the one obtained when only the initial set of moments was used. More precisely, the asymptotic variance was decomposed in two terms: one equal to the asymptotic variance when only the first set of moments was used, and the other term coming from the addition of the second set of moments.

However, in presence of asymptotic bias, the focus is not on the asymptotic variance matrix but rather on the AMSE matrix as it was explained above. As a result, their approach is no longer the best since it does not provide, in general, a block-diagonal AMSE matrix, making the comparison between AMSEs not straightforward. Accordingly, the question is how can we transform the additional moment conditions such that we achieve a block-diagonal AMSE matrix? After careful examination of the methodology, we believe that the answer we suggest is quite satisfactory in that it preserves the simplicity of the original method and can be seen as a natural generalization of the transformation of moments based on g_2 found in Breusch et al. (1999). Namely, we propose to replace in the expression for the projection of g_2 on g_1 the matrices based on blocks of Ω, i.e., Ω_{21} and Ω_{11}^{-1}, by the analogous matrices based on blocks of M, i.e., M_{21} and M_{11}^{-1}. This transformation, though it will not deliver a block-diagonal asymptotic variance matrix for the moment estimators, in general, it will deliver a block-diagonal AMSE matrix. As desired, the comparison between AMSEs becomes now more straightforward.

We may wonder why looking at redundancy or partial redundancy is of relevance since, as we show in the paper, adding moment conditions does not hurt in terms of AMSE. The natural question is why not simply add all the moments? The answer is twofold. On the one hand, examining conditions in this more general framework with asymptotic bias allows us to study more closely the variance-bias trade-off. In

particular, from the redundancy conditions we show that having some asymptotic bias is not unfavorable, in terms of AMSE, in certain situations. On the other hand, in some cases adding extra moment conditions, in presence of asymptotic bias, might lead to more complicated estimators, in the sense that the estimator of the parameter becomes computationally intensive. As a result, the conditions derived in this paper might help us assess the information contained in the extra moments before we unnecessarily complicate our estimation procedure.

The remainder of the paper is organized as follows. In Section 3.2, we derive the optimal choice of weighting matrix for GMM under a non-zero bias case for the estimator of the moment conditions. We give an intuitive explanation of why this is a natural choice in our more general framework. In Section 3.3, we derive conditions for redundancy of one set of moment conditions given a second set. We provide several equivalent forms to state the same redundancy condition, and explain the underlying intuition. In Section 3.4, we establish partial redundancy of one set of moment conditions, for the estimation of a subset of the parameters, given a second set of moments. As in the case of redundancy, we provide conditions for partial redundancy in several equivalent forms. We briefly compare what partial redundancy requires with respect to full redundancy. Section 3.5 concludes. Some of the proofs are relegated to Appendix 3.A.

3.2 Optimal Choice of Weighting Matrix W

In this section, we derive the optimal choice of weighting matrix for a GMM estimator in the more general framework of non-zero asymptotic bias of the estimator of the moment conditions. This section plays a key role to understand the results derived in the rest of the paper.

In this paper, we consider a simple setup, in which we estimate a $p \times 1$ vector

of true values of parameters θ_0 using different sets of moment conditions. The key difference with respect to standard GMM settings, is that we allow our estimator to have a non-zero asymptotic bias. More precisely, let the set of moment conditions be given by:

$$\begin{cases} E[g_1(y_t;\theta_0)] = 0 \\ E[g_2(y_t;\theta_0)] = 0 \qquad for \;\; t = 1, ..., T \end{cases} \tag{3.1}$$

where y_t is a scalar random variable for simplicity, and $g_1(y_t;\theta_0)$ and $g_2(y_t;\theta_0)$ have dimension k_1 and k_2 respectively.

Define, $g(y_t;\theta) = \begin{bmatrix} g_1(y_t;\theta)' & g_2(y_t;\theta)' \end{bmatrix}'$, for $t = 1, ..., T$, so that $E[g(y_t;\theta_0)] = 0$ is the vector that contains all the moment conditions $k = k_1 + k_2$.

We assume that:

Assumption 3.1. *The true value of the parameter, θ_0, is fully identified from the first set of moments only, and thus we take $k_1 \geq p$:*

$$E[g_1(y_t;\theta)] = 0 \quad \Leftrightarrow \quad \theta = \theta_0 \tag{3.2}$$

In addition, for simplicity of exposition, we assume the following:

Assumption 3.2. *The data consists of a finite number T of observations, $y_1, ..., y_T$. The process $\{y_t : t \in \mathbb{N}\}$ on $\mathcal{Y} \subset \mathbb{R}$, is strictly stationary and ergodic.*

Assumption 3.3. $\{g(y_t;\theta) : t \in \mathbb{N}\}$ *is a martingale difference sequence (m.d.s.).*

Remark 3.1. *On the one hand, Assumption 3.3 can be relaxed to allow for some more complicated time series settings, like mixing conditions for processes, instead of m.d.s. On the other hand, the main message can be transmitted by just focusing on the most restricted but simplest case: i.i.d.*

Let $\ddot{\theta}_T$ be a GMM estimator of θ based on the full set of moment conditions (3.1)

and a weighting matrix W. The minimization problem is the following:

$$\ddot{\theta}_T = arg \min_{\theta \in \Theta} Q_T(\theta),$$

with

$$\begin{aligned} Q_T(\theta) &= \bar{g}(\theta)' W \bar{g}(\theta) \\ \bar{g}(\theta) &= (\hat{E}[g_1(y_t;\theta)]' \quad \hat{E}[g_2(y_t;\theta)]')' \end{aligned} \tag{3.3}$$

where $\hat{E}[g_i(y_t;\theta)]$ denotes the sample average estimator of $E[g_i(y_t;\theta_0)]$, for $i = 1, 2$; and W is a positive definite (p.d.) weighting matrix.

Remark 3.2. *Notice that we need W in order to derive our estimator of θ. As it is standard in GMM settings we can first take $W = I$, where I is the identity matrix, obtain a consistent estimator of the parameter and hence a consistent estimator for W, say W_T. Next, estimate our parameter using the estimator of the weighting matrix W_T.*

We consider a GMM estimator that is subject to non-zero asymptotic bias due to the fact that the asymptotic distribution of the estimator of the moment conditions is such that[2]

Assumption 3.4. *The asymptotic distribution of the estimator of the moment conditions is given by*

$$\sqrt{T}\bar{g}(\theta_0) \xrightarrow{d} \mathcal{N}(b_0,\ \Omega) \tag{3.4}$$

where $\bar{g}(\theta_0)$ is the sample average estimator of $E[g(y_t;\theta_0)]$, b_0 is the asymptotic bias, and Ω is the asymptotic variance-covariance matrix.

[2] A more complicated version of this estimator arises in the so called "Kernel Moment Estimator" of Gagliardini, Gouriéroux, and Renault (2011), with the caveat that in addition they consider more complex identification settings due to the nature of their problem.

Remark 3.3. *It is important to understand the implications of Assumption 3.4—namely that, even though we consider moment conditions that are unbiased, the estimator of these moments is asymptotically biased and as a result the estimator of the parameter is also asymptotically biased. We should take some caution in what we mean by "asymptotically biased." This qualification should be understood in the following sense: we say the estimator of the moment conditions, $\bar{g}$, is asymptotically biased in the sense that $\sqrt{T}\bar{g}$ has a nonzero asymptotic bias, but the asymptotic bias of $\bar{g}$ alone which is $b_0/\sqrt{T}$ converges to zero as $T \to \infty$.*

We make the following standard assumption about Ω:

Assumption 3.5. *Ω is a finite and non-singular variance-covariance matrix.*

We also make standard assumptions on the parameter space and the function g:

Assumption 3.6. *Θ is an open subset of $\mathbb{R}^p$ that contains θ_0.*

Assumption 3.7. *$g(\cdot;\theta)$ and $\partial g/\partial\theta(\cdot;\theta)$ are Borel measurable for each $\theta \in \Theta$ and $\partial g/\partial\theta(y;\cdot)$ is continuous on Θ for each $y \in \mathbb{R}$.*

Assumption 3.8. *$\partial g(y_1;\theta)/\partial\theta$ is first moment continuous at θ_0 and $E[\partial g/\partial\theta(y_1;\theta_0)]$ exists, is finite, and has full rank.*

Assumption 3.9. *$\operatorname{plim}_{T\to\infty} \frac{\partial \bar{g}'}{\partial \theta'}(\ddot{\theta}_T) = J_0$, and $\operatorname{plim}_{T\to\infty} W_T = W$.*

Remark 3.4. *By Lemma 3.1 of Hansen (1982), under Assumptions 3.7 and 3.8, if $E[g(y_1;\theta_0)]$ exists and is finite, then $g(y_1;\theta)$ is first moment continuous at θ_0. This last property, is important to guarantee consistency (see consistency Theorem 2.1 of Hansen, 1982).*

In this situation, the asymptotic distribution of the GMM estimator using a weighting matrix W will be slightly modified to take into account the asymptotic

bias as follows:

$$\sqrt{T}(\ddot{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(B_\infty, (J_0'WJ_0)^{-1}J_0'W\Omega WJ_0(J_0'WJ_0)^{-1}) \tag{3.5}$$

where $B_\infty = -(J_0'WJ_0)^{-1}J_0'Wb_0$ is the asymptotic bias of $\sqrt{T}(\ddot{\theta}_T - \theta_0)$, and $J_0 = J(\theta_0)$ is the $(k \times p)$ limit Jacobian matrix as defined in (3.14) in Section 3.3. The details of the derivation of (3.5) are quite standard, and thus are relegated to the Appendix.

Notice that, the optimal weighting matrix W is set in general such that it minimizes the asymptotic variance of $\sqrt{T}(\ddot{\theta}_T - \theta_0)$, which leads to setting $W = \Omega^{-1}$, where Ω is the asymptotic variance of $\sqrt{T}\bar{g}(\theta_0)$. However, in the presence of asymptotic bias, the natural choice for W will be one that minimizes the AMSE instead of only asymptotic variance as in the standard case. Intuitively, if we were to ignore the asymptotic bias, we would be obtaining a more precise estimator around a value that is not θ_0. Hence, that choice of weighting matrix would not be adequate. The right approach is therefore, to choose a weighting matrix that accounts for both asymptotic bias and variance, that is:

$$\begin{aligned} AMSE\left[\sqrt{T}(\ddot{\theta}_T - \theta_0)\right] &= Avar\left[\sqrt{T}(\ddot{\theta}_T - \theta_0)\right] + Abias^2\left[\sqrt{T}(\ddot{\theta}_T - \theta_0)\right] \\ &= (J_0'WJ_0)^{-1}J_0'W\Omega WJ_0(J_0'WJ_0)^{-1} \\ &\quad + (J_0'WJ_0)^{-1}J_0'Wb_0b_0'WJ_0(J_0'WJ_0)^{-1} \\ &= (J_0'WJ_0)^{-1}J_0'W[\Omega + b_0b_0']WJ_0(J_0'WJ_0)^{-1} \\ &= (J_0'WJ_0)^{-1}J_0'WM_0WJ_0(J_0'WJ_0)^{-1} \end{aligned} \tag{3.6}$$

with

$$M_0 := \Omega + b_0b_0' \tag{3.7}$$

$$Abias^2 := Abias \;\; Abias' \tag{3.8}$$

where *Avar* and *Abias* stand for the asymptotic variance and bias respectively.

As a result, the optimal choice of the weighting matrix is $W = M_0^{-1}$ since in this case $AMSE$ reduces to $(J_0' M_0^{-1} J_0)^{-1}$.[3] Hence, we can define the following Optimal GMM estimator in presence of asymptotic bias:

Definition 3.1 (**Optimal GMM in presence of asymptotic bias**). *A GMM estimator is said to be optimal in presence of asymptotic bias when the weighting matrix used equals the inverse of the AMSE of the estimator of the moment functions, evaluated at the true value of the parameter.*

Let $\ddot{\theta}_T$ be the optimal GMM estimator of θ, then its asymptotic distribution is given by

$$\sqrt{T}(\ddot{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(B_\infty, (J_0' M_0^{-1} J_0)^{-1} J_0' M_0^{-1} \Omega M_0^{-1} J_0 (J_0' M_0^{-1} J_0)^{-1}) \tag{3.9}$$

with

$$B_\infty = -(J_0' M_0^{-1} J_0)^{-1} J_0' M_0^{-1} b_0 \tag{3.10}$$

where B_∞ is the asymptotic bias of $\sqrt{T}(\ddot{\theta}_T - \theta_0)$ when $W = M_0^{-1}$, Ω is defined in (3.4), J_0 is defined in (3.5), and M_0 is defined in (3.7).

In this paper, we will only consider optimal GMM estimators as framed in Definition 3.1.

[3] Notice, however, that under this choice of W the asymptotic variance will still take a sandwich form $(J_0' M_0^{-1} J_0)^{-1} J_0' M_0^{-1} \Omega M_0^{-1} J_0 (J_0' M_0^{-1} J_0)^{-1}$. It is clear from the above expression, that only in absence of asymptotic bias we will get the reduced formula for the asymptotic variance, since if $M_0 = \Omega$ then $Avar = (J_0' M_0^{-1} J_0)^{-1} J_0' M_0^{-1} \Omega M_0^{-1} J_0 (J_0' M_0^{-1} J_0)^{-1} = (J_0' \Omega^{-1} J_0)^{-1} J_0' \Omega^{-1} \Omega \Omega^{-1} J_0 (J_0' \Omega^{-1} J_0)^{-1} = (J_0' \Omega^{-1} J_0)^{-1}$.

3.3 Redundancy in presence of Asymptotic Bias

In this section, we introduce the concept of redundancy, and provide necessary and sufficient conditions under which it holds true. We first give a detailed explanation of how the standard methodology used to derive conditions for redundancy has to be altered in order to account for asymptotic bias in the estimator of the moment conditions. Next, we present the main result of the paper in Theorem 3.2, namely several equivalent necessary and sufficient conditions for redundancy. Finally, we provide a brief discussion on the bias-variance trade-off that takes place in our modified framework.

3.3.1 Redundancy for two sets of moment conditions

In this subsection, we derive necessary and sufficient conditions for redundancy of one set of moment conditions given a second set. We first explain the approach by which we derive these redundancy conditions, and then introduce the main result of the paper.

Consider the following partition of the asymptotic bias and variance of the estimator of the moment conditions,

$$b_0 = \begin{bmatrix} b_{01} \\ b_{02} \end{bmatrix} \tag{3.11}$$

and

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \tag{3.12}$$

where b_{0i} is the asymptotic bias of $\sqrt{T}\hat{E}[g_i(y_t;\theta_0)]$, for $i = 1, 2$; and Ω_{ij} is the asymptotic variance-covariance matrix between $\sqrt{T}\hat{E}[g_i(y_t;\theta_0)]$ and $\sqrt{T}\hat{E}[g_j(y_t;\theta_0)]$ with

$i, j = 1, 2$.

In Section 3.2, we derived that the optimal weighting matrix for GMM in presence of asymptotic bias is the inverse of

$$M_0 = \Omega + b_0 b_0' = \begin{bmatrix} \Omega_{11} + b_{01}b_{01}' & \Omega_{12} + b_{01}b_{02}' \\ \Omega_{21} + b_{02}b_{01}' & \Omega_{22} + b_{02}b_{02}' \end{bmatrix} := \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \tag{3.13}$$

where M_0 is finite and non-singular, given that Ω is assumed finite and non-singular.

Define

$$J_0 = E\left[\frac{\partial g(y_t;\theta_0)}{\partial \theta'}\right] = E\begin{bmatrix} \frac{\partial g_1(y_t;\theta_0)}{\partial \theta'} \\ \frac{\partial g_2(y_t;\theta_0)}{\partial \theta'} \end{bmatrix} := \begin{bmatrix} J_{01} \\ J_{02} \end{bmatrix} \tag{3.14}$$

We assume the following about the identification of the parameters:

Assumption 3.10. *J_{01} has full column rank.*

Remark 3.5. *Assumption 3.10 implies that θ_0 is fully identified by the first set of moment conditions only, and thus we require $k_1 \geq p$ as stated in Section 3.2. Notice that this is a local identification assumption, while assumption 3.1 is about global identification.*

Let $\hat{\theta}_T$ and $\ddot{\theta}_T$ be the GMM estimators of θ based on the moment conditions $E[g_1(y_t;\theta_0)] = 0$ only, and the full set of moments in (3.1) respectively. The optimal weighting matrices are M_{11}^{-1} and M_0^{-1} respectively. Under our assumptions, and following the derivations in Section 3.2, we have that the asymptotic distributions of our estimators are given by,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(B_{1\infty}, (J_{01}'M_{11}^{-1}J_{01})^{-1}J_{01}'M_{11}^{-1}\Omega_{11}M_{11}^{-1}J_{01}(J_{01}'M_{11}^{-1}J_{01})^{-1}) \tag{3.15}$$

and

$$\sqrt{T}(\ddot{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(B_\infty, (J_0' M_0^{-1} J_0)^{-1} J_0' M_0^{-1} \Omega M_0^{-1} J_0 (J_0' M_0^{-1} J_0)^{-1}) \tag{3.16}$$

with

$$B_{1\infty} = -(J_{01}' M_{11}^{-1} J_{01})^{-1} J_{01}' M_{11}^{-1} b_{01} \tag{3.17}$$

where B_∞ is defined in (3.10).

The extra moment conditions based on g_2 can never hurt asymptotically, in the sense that they cannot increase the asymptotic mean square error since the difference

$$AMSE[\sqrt{T}(\hat{\theta}_T - \theta_0)] - AMSE[\sqrt{T}(\ddot{\theta}_T - \theta_0)] = (J_{01}' M_{11}^{-1} J_{01})^{-1} - (J_0' M_0^{-1} J_0)^{-1} \tag{3.18}$$

is positive semi-definite (p.s.d).[4] Hence, we can state the following definition

Definition 3.2 (**Redundancy**). *We say g_2 is redundant given g_1 if the optimal GMM estimator of θ based on the first set of moments in (3.1) only, has the same AMSE as the optimal GMM estimator of θ based on the full set of moments in (3.1)—that is, if the difference in equation (3.18) is zero.*

3.3.1.1 Transforming the Moment Conditions

Here we discuss in detail how the moment conditions should be manipulated in order to work with more suitable expressions to derive the conditions for redundancy later in the paper.

Deriving conditions for redundancy of g_2 given g_1 can be done simply by finding conditions under which equation (3.18) is equal to zero. However, proceeding this

[4]The proof of this statement will become clear later in the paper, in Theorem 3.2, once we redefine the moment conditions based on g_2 and rewrite the expression for the AMSE.

way can be algebraically demanding and hard to interpret. The following alternative approach, based on Breusch et al. (1999), turns out to be much better in that it sheds some light on the underlyings of redundancy.

The basic idea is that the estimator derived from moments based on g_1 and g_2 is numerically equal to another estimator based on moments from g_1 and the residual of a projection of g_2 on g_1.

Let

$$EL[g_2(y_t;\theta_0) \mid g_1(y_t;\theta_0)] = \Omega_{21}\Omega_{11}^{-1}g_1(y_t;\theta_0) \tag{3.19}$$

and

$$r_2(y_t;\theta_0) = g_2(y_t;\theta_0) - EL[g_2(y_t;\theta_0) \mid g_1(y_t;\theta_0)] \tag{3.20}$$

where $EL[\cdot \mid \cdot]$ is the linear projection of $g_2(y_t;\theta_0)$ on $g_1(y_t;\theta_0)$, and $r_2(y_t;\theta_0)$ is the residual in this linear projection. From these definitions, we can consider GMM estimation based on the following set of moment conditions

$$E[\varphi(y_t;\theta_0)] := \begin{bmatrix} E[g_1(y_t;\theta_0)] \\ E[r_2(y_t;\theta_0)] \end{bmatrix} = \begin{bmatrix} E[g_1(y_t;\theta_0)] \\ E[g_2(y_t;\theta_0) - EL[g_2(y_t;\theta_0) \mid g_1(y_t;\theta_0)]] \end{bmatrix} = 0 \tag{3.21}$$

This new set differs from the original one in that $g_2(y_t;\theta_0)$ has been replaced by the residual in its projection on $g_1(y_t;\theta_0)$. Then,

$$\sqrt{T}\bar{\varphi}(y_t;\theta_0) \xrightarrow{d} \mathcal{N}(\tilde{b}_0, \tilde{\Omega}) \tag{3.22}$$

where

$$\tilde{b}_0 = \begin{bmatrix} b_{01} \\ b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01} \end{bmatrix} \tag{3.23}$$

and

$$\tilde{\Omega} = \begin{bmatrix} \Omega_{11} & 0 \\ 0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} \end{bmatrix} := \begin{bmatrix} \tilde{\Omega}_{11} & \tilde{\Omega}_{12} \\ \tilde{\Omega}_{21} & \tilde{\Omega}_{22} \end{bmatrix} \tag{3.24}$$

Hence, the weighting matrix is given by the inverse of

$$\begin{aligned} \tilde{M}_0 &= \tilde{\Omega} + \tilde{b}_0\tilde{b}_0' \\ &:= \begin{bmatrix} \tilde{M}_{11} & \tilde{M}_{12} \\ \tilde{M}_{21} & \tilde{M}_{22} \end{bmatrix} \end{aligned} \tag{3.25}$$

with

$$\begin{aligned} \tilde{M}_{11} &= \Omega_{11} + b_{01}b_{01}' \\ \tilde{M}_{12} &= b_{01}[b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01}]' \\ \tilde{M}_{21} &= [b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01}]b_{01}' \\ \tilde{M}_{22} &= [\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}] + [b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01}][b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01}]' \end{aligned}$$

and the limit Jacobian matrix now is

$$\tilde{J}_0 = E\left[\frac{\partial\varphi(y_t;\theta_0)}{\partial\theta'}\right] = \begin{bmatrix} J_{01} \\ J_{02} - \Omega_{21}\Omega_{11}^{-1}J_{01} \end{bmatrix} := \begin{bmatrix} \tilde{J}_{01} \\ \tilde{J}_{02} \end{bmatrix} \tag{3.26}$$

This approach was very useful in Breusch et al. (1999) to the extent that it made the variance-covariance matrix of the moment estimators block-diagonal. This was

key in their paper since it allowed the delivery of an expression for the asymptotic variance directly comparable to the one obtained when only the initial set of moments was used. More precisely, the asymptotic variance was decomposed in two terms: one equal to the asymptotic variance when only the first set of moments was used, and the other term coming from the addition of the second set of moments. However, in presence of asymptotic bias, the focus is not on the asymptotic variance matrix but on the AMSE matrix as explained in the previous section. As a result, this approach is not the best since it does not provide, in general, a block-diagonal AMSE matrix, making the comparison between AMSEs not straightforward.

However, we say "in general" since there is an exception to this last statement. From the expression of $\tilde{M}_0$ it is clear that if $\tilde{M}_{12} = 0$ we will still get a block-diagonal AMSE matrix. Since $\tilde{M}_{12} = b_{01}[b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01}]'$, $\tilde{M}_{12}$ will be zero if $b_{01} = 0$ or if $b_{02} = \Omega_{21}\Omega_{11}^{-1}b_{01}$. The first condition simply means that the asymptotic bias coming from the estimator of moments based on g_1 is zero; while the second condition means that the asymptotic bias coming from the estimator of moments based on $r_2(y_t;\theta_0)$ is zero, where $E[r_2(y_t;\theta_0)]$ are the transformed moments that deliver a block-diagonal Ω. Summing up, if after the transformation that makes Ω block-diagonal we have zero asymptotic bias from the estimator of either set of moment conditions in (3.21), then $\tilde{M}_{12} = 0$ and, as a result, we will also have a block-diagonal AMSE matrix.

A natural generalization of Breusch et al.'s (1999) transformation of moments based on g_2 would be to replace in the expression for the projection of g_2 on g_1 the matrices Ω_{21} and Ω_{11}^{-1} by M_{21} and M_{11}^{-1}. This transformation, though it will not deliver, in general, a block-diagonal asymptotic variance matrix for the moment estimators, it will deliver a block-diagonal AMSE matrix. The underlying rationale can be easily justified as follows:

Analogous to having $C^* = \Omega_{21}\Omega_{11}^{-1}$ as the solution to

$$C^* = arg\min_C Avar[g_2 - Cg_1]$$

where $Avar[X]$ is the asymptotic variance of X; we can think there exists a matrix B^* such that

$$B^* = arg\min_B AMSE[g_2 - Bg_1]$$

It is straightforward to show our conjecture that $B^* = M_{21}M_{11}^{-1}$:

Proof. *Let $B^* = arg\min_B AMSE[g_2 - Bg_1]$, we have that:*

$$\begin{aligned} AMSE[g_2 - Bg_1] &= Avar[g_2 - Bg_1] + [Abias(g_2 - Bg_1)][Abias(g_2 - Bg_1)]' \\ &= \Omega_{22} + B\Omega_{11}B' - \Omega_{21}B' - B\Omega_{12} + [b_{02} - Bb_{01}][b_{02} - Bb_{01}]' \end{aligned}$$

where we have used the notation for asymptotic variance and bias provided in (3.12) *and* (3.11) *respectively.*

Thus the FOC is given by:

$$\begin{aligned} & 2B^*\Omega_{11} - 2\Omega_{21} - 2(b_{02} - B^*b_{01})b_{01}' = 0 \\ \Leftrightarrow \quad & B^*[\Omega_{11} + b_{01}b_{01}'] - [\Omega_{21} + b_{02}b_{01}'] = 0 \\ \Leftrightarrow \quad & B^*M_{11} - M_{21} = 0 \\ \Leftrightarrow \quad & B^* = M_{21}M_{11}^{-1} \end{aligned}$$

□

Therefore, let

$$EL_M[g_2(y_t;\theta_0) \mid g_1(y_t;\theta_0)] = M_{21}M_{11}^{-1}g_1(y_t;\theta_0) \tag{3.27}$$

and

$$r_{2M}(y_t;\theta_0) = g_2(y_t;\theta_0) - EL_M[g_2(y_t;\theta_0) \mid g_1(y_t;\theta_0)] \tag{3.28}$$

where $EL_M[\cdot \mid \cdot]$ is the modified linear projection of $g_2(y_t;\theta_0)$ on $g_1(y_t;\theta_0)$, and $r_{2M}(y_t;\theta_0)$ is the residual in this modified linear projection. From these definitions, we can consider GMM estimation based on the following set of moment conditions:

$$\begin{aligned} E[\psi(y_t;\theta_0)] :&= \begin{bmatrix} E[g_1(y_t;\theta_0)] \\ E[r_{2M}(y_t;\theta_0)] \end{bmatrix} \\ &= \begin{bmatrix} E[g_1(y_t;\theta_0)] \\ E[g_2(y_t;\theta_0) - EL_M[g_2(y_t;\theta_0) \mid g_1(y_t;\theta_0)]] \end{bmatrix} = 0 \end{aligned} \tag{3.29}$$

This new set differs from the original one in that $g_2(y_t;\theta_0)$ has been replaced by the residual in the modified projection on $g_1(y_t;\theta_0)$. Then,

$$\sqrt{T}\bar{\psi}(y_t;\theta_0) \xrightarrow{d} \mathcal{N}(\breve{b}_0, \breve{\Omega}) \tag{3.30}$$

where

$$\breve{b}_0 = \begin{bmatrix} b_{01} \\ b_{02} - M_{21}M_{11}^{-1}b_{01} \end{bmatrix} \tag{3.31}$$

and

$$\breve{\Omega} :: = \begin{bmatrix} \breve{\Omega}_{11} & \breve{\Omega}_{12} \\ \breve{\Omega}_{21} & \breve{\Omega}_{22} \end{bmatrix} \tag{3.32}$$

with

$$\breve{\Omega}_{11} = \Omega_{11}$$

$$\breve{\Omega}_{12} = \Omega_{12} - \Omega_{11} M_{11}^{-1} M_{12}$$

$$\breve{\Omega}_{21} = (\Omega_{12} - \Omega_{11} M_{11}^{-1} M_{12})'$$

$$\breve{\Omega}_{22} = \Omega_{22} + M_{21} M_{11}^{-1} \Omega_{11} M_{11}^{-1} M_{12} - \Omega_{21} M_{11}^{-1} M_{12} - (\Omega_{21} M_{11}^{-1} M_{12})'$$

Hence, after some algebra, the weighting matrix is given by the inverse of

$$\begin{aligned}
\breve{M}_0 &= \breve{\Omega} + \breve{b}_0 \breve{b}_0' \\
&= \begin{bmatrix} \Omega_{11} + b_{01} b_{01}' & 0 \\ 0 & (\Omega_{22} + b_{02} b_{02}') - M_{21} M_{11}^{-1} M_{12} \end{bmatrix} \\
&:= \begin{bmatrix} M_{11} & 0 \\ 0 & M_{22} - M_{21} M_{11}^{-1} M_{12} \end{bmatrix} \\
&:= \begin{bmatrix} \breve{M}_{11} & \breve{M}_{12} \\ \breve{M}_{21} & \breve{M}_{22} \end{bmatrix}
\end{aligned} \tag{3.33}$$

Finally, the limit Jacobian matrix is now given by

$$\begin{aligned}
\breve{J}_0 &= E\left[\frac{\partial \psi(y_t; \theta_0)}{\partial \theta'}\right] \\
&= \begin{bmatrix} J_{01} \\ J_{02} - M_{21} M_{11}^{-1} J_{01} \end{bmatrix} := \begin{bmatrix} \breve{J}_{01} \\ \breve{J}_{02} \end{bmatrix}
\end{aligned} \tag{3.34}$$

Remark 3.6. *Though this approach is more straightforward and adequate for our general framework, it can be shown, with some more algebra, that the redundancy results below can also be obtained from the transformed moments* (3.21).

Remark 3.7. *Two special cases are worth noticing. If* $b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01} = 0$*, or put in a different way* $b_{02} = \Omega_{21}\Omega_{11}^{-1}b_{01}$*:*

$$\begin{aligned} M_{21}M_{11}^{-1} &= (\Omega_{21} + b_{02}b_{01}')(\Omega_{11} + b_{01}b_{01}')^{-1} \\ &= (\Omega_{21} + \Omega_{21}\Omega_{11}^{-1}b_{01}b_{01}')(\Omega_{11} + b_{01}b_{01}')^{-1} \\ &= \Omega_{21}(I + \Omega_{11}^{-1}b_{01}b_{01}')(I + \Omega_{11}^{-1}b_{01}b_{01}')^{-1}\Omega_{11}^{-1} \\ &= \Omega_{21}\Omega_{11}^{-1} \end{aligned}$$

Similarly, if $b_{01} = 0$ *it is straightforward to see that* $M_{21}M_{11}^{-1} = \Omega_{21}\Omega_{11}^{-1}$*. This means that, if* $b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01} = 0$ *or* $b_{01} = 0$*, transforming the moment conditions such that the AMSE matrix is block-diagonal, reduces to transforming them just to make* Ω *block-diagonal.*

This is in line with our previous conclusion: "if after the transformation that makes Ω *block-diagonal we have zero asymptotic bias from the estimator of either set of moment conditions in* (3.21)*, then* $\tilde{M}_{12} = 0$ *and, as a result, we will also have a block-diagonal AMSE matrix." Overall, this tell us that in either of these two cases to analyze redundancy we can simply rely on the transformation that makes* Ω *block-diagonal. This is the foundation for our first result, Theorem 3.1, in the next subsection.*

Before moving to the results, it is useful to summarize the main notation introduced so far and that will be used henceforward in this paper.

Notation. *Throughout this paper we use the following notation:*

Compact notation: $g := g(y_t;\cdot)$*; similarly* $g_i := g_i(y_t;\cdot)$ *for* $i = 1, 2$*. The same compact notation applies to* $r_2(y_t;\cdot)$ *and* $r_{2M}(y_t;\cdot)$*.*

b_0*: asymptotic bias (column) vector of the estimator of the full set of moment conditions as defined in* (3.11)*.*

Ω*: asymptotic variance matrix of the estimator of the full set of moment conditions as defined in* (3.12)*.*

$M_0 = \Omega + b_0 b_0'$*: AMSE matrix of the estimator of the full set of moment conditions as defined in* (3.13)*.*

J_0*: limit Jacobian matrix as defined in* (3.14)*.*

These objects are based on the moment conditions $E[g(y_t;\theta_0)] = 0$ *defined in* (3.1)*.*

When the above objects have the " $\tilde{}$ *" symbol on top, they are based instead on the modified set of moment conditions,* $E[\varphi(y_t;\theta_0)] = 0$*, defined in* (3.21)*. That is, the projection using blocks of* Ω*.*

When the above objects have instead the " $\breve{}$ *" symbol on top, they are based instead on the modified set of moment conditions,* $E[\psi(y_t;\theta_0)] = 0$*, defined in* (3.29)*. That is, the projection using blocks of* M*.*

3.3.1.2 Redundancy Results

The main results of the paper are presented here. The theorems introduced in this subsection are key to understanding the central message delivered in this paper. Theorem 3.2 presents the more general result. The proof of this theorem is also given here since it is simple and it provides some insights on the result.

Before establishing our first result, we introduce the following lemma that justifies

using either the original or the modified set of moment conditions.

Lemma 3.1. *Let Assumptions 3.1 - 3.10 hold. Let* $\ddot{\theta}_T$ *be the GMM estimator of* θ *based on the full set of moment conditions* (3.1)*, using weighting matrix* M_0^{-1}*; and let* $\breve{\theta}_T$ *be the GMM estimator of* θ *based on the modified set of moment conditions* (3.29)*, using the weighting matrix* $\breve{M}_0^{-1}$*. Then* $\ddot{\theta}_T$ *is numerically equal to* $\breve{\theta}_T$*.*

The intuition is that since the moment conditions based on $\psi(y_t;\theta_0)$ in (3.29) contain the same information as those based on $g(y_t;\theta_0)$ in (3.1), the GMM estimators based on each of these sets are expected to be numerically equal.[5] The proof is given in the Appendix.

Remember that redundancy in our more general framework will be established in terms of AMSE and not asymptotic variance as in the standard case. The intuition is that, in our modified setting, we could also have a gain in terms of asymptotic bias, which would be overlooked if we were to focus only on the asymptotic variance. Thus, we need next to derive the expression for the AMSE.

The AMSE, for a choice of weighting matrix $\breve{M}_0^{-1}$, is given by

$$\begin{aligned} AMSE &= (\breve{J}_0'\breve{M}_0^{-1}\breve{J}_0)^{-1} \\ &= \left\{ \begin{bmatrix} \breve{J}_{01}' & \breve{J}_{02}' \end{bmatrix} \begin{bmatrix} \breve{M}_{11}^{-1} & 0 \\ 0 & \breve{M}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \breve{J}_{01} \\ \breve{J}_{02} \end{bmatrix} \right\}^{-1} \\ &= \left(\breve{J}_{01}'\breve{M}_{11}^{-1}\breve{J}_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02} \right)^{-1} \\ &= \left(J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'(M_{22} - M_{21}M_{11}^{-1}M_{12})^{-1}\breve{J}_{02} \right)^{-1} \end{aligned} \tag{3.35}$$

where we have used that $\breve{M}_{11}^{-1} = M_{11}^{-1}$ and that $\breve{J}_{01} = J_{01}$.

This expression facilitates the comparison with a situation where only moments based on $g_1(y_t;\theta_0)$ are used, since the inverse of the first term, $(J_{01}'M_{11}^{-1}J_{01})^{-1}$, corre-

[5] They will be numerically equal as long as the first step consistent estimators used for Ω and J_0 are the same. Otherwise, they will just be asymptotically equivalent since both are consistent.

sponds to the AMSE for that case. Consequently, we can think that the other term arises from including also moments based on $g_2(y_t;\theta_0)$.

Moreover, this expression provides a proof for our earlier statement that the extra moment conditions given by g_2 can never hurt asymptotically, in the sense that they cannot increase the asymptotic mean square error. That is, using equation (3.35), it is now straightforward to check that the difference

$$\begin{aligned} &AMSE[\sqrt{T}(\hat{\theta}_T - \theta_0)] - AMSE[\sqrt{T}(\breve{\theta}_T - \theta_0)] \\ &= (J_{01}' M_{11}^{-1} J_{01})^{-1} - \left(J_{01}' M_{11}^{-1} J_{01} + \breve{J}_{02}' \breve{M}_{22}^{-1} \breve{J}_{02} \right)^{-1} \end{aligned} \tag{3.36}$$

is positive semi-definite.

Based on Lemma 1, the following two theorems state the conditions for g_2 to be redundant given g_1.

Theorem 3.1 (Case of $\tilde{M}_{12} = 0$: zero asymptotic bias from g_1 or r_2)**.** *Under assumptions 3.1 - 3.10, if $b_{01} = 0$ or $b_{02} - \Omega_{21}\Omega_{11}^{-1} b_{01} = 0$, then $\tilde{M}_{12} = 0$, and as a result the following statements are equivalent:*

(i) g_2 *is redundant given* g_1.

(ii) $\tilde{J}_{02} := E[\partial r_2(y_t, \theta_0)/\partial \theta'] = E\{\partial [g_2(y_t, \theta_0) - \Omega_{21}\Omega_{11}^{-1} g_1(y_t, \theta_0)]/\partial \theta'\} = 0$.

(iii) $J_{02} = \Omega_{21}\Omega_{11}^{-1} J_{01}$.

(iv) There exists a $k_1 \times p$ *matrix* A *such that* $J_{01} = \Omega_{11} A$ *and* $J_{02} = \Omega_{21} A$.

Remark 3.8. *Let $\tilde{\theta}_T$ be the GMM estimator of θ based on the modified set of moment conditions (3.21) using the weighting matrix $\tilde{M}_0^{-1}$. Notice that Lemma 3.1 holds with $\tilde{\theta}_T$ instead of $\breve{\theta}_T$ and $\tilde{M}_0$ instead of $\breve{M}_0$ since as we showed in the previous subsection when $b_{01} = 0$ or $b_{02} - \Omega_{21}\Omega_{11}^{-1} b_{01} = 0$ we have that $M_{21} M_{11}^{-1} = \Omega_{21}\Omega_{11}^{-1}$.*

Remark 3.9. *Theorem 3.1 is simply a generalization of Theorem 3.1 in Breusch et al. (1999). In their paper, they do not consider asymptotic bias in any of the moment estimators. Our Theorem shows that the same result can be obtained if we do not allow for asymptotic bias in either the estimator of moments based on g_1 or on r_2. That is, their result can be seen as a particular case of our Theorem 3.1.*

Proof. *If the asymptotic bias coming from g_1 is zero, that is $b_{01} = 0$, or the asymptotic bias coming from r_2 is zero, that is $b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01} = 0$, we showed in the previous subsection that $M_{21}M_{11}^{-1} = \Omega_{21}\Omega_{11}^{-1}$, which means that we can use the transformation that makes block-diagonal Ω since it also makes the AMSE matrix block-diagonal. Therefore, the limit Jacobian Matrix reduces to the same as in Breusch et al. (1999), that is matrix G in equation (14) of their paper. Thus, $\breve{J}_{02} = \tilde{J}_{02} = J_{02} - \Omega_{21}\Omega_{11}^{-1}J_{01} = G_2$. As a result:*

$$
\begin{aligned}
&AMSE[\sqrt{T}(\hat{\theta}_T - \theta_0)] - AMSE[\sqrt{T}(\breve{\theta}_T - \theta_0)] \\
&= (J_{01}'M_{11}^{-1}J_{01})^{-1} - \left(J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02}\right)^{-1} \\
&= (J_{01}'\Omega_{11}^{-1}J_{01})^{-1} - \left(J_{01}'\Omega_{11}^{-1}J_{01} + \tilde{J}_{02}'\tilde{M}_{22}^{-1}\tilde{J}_{02}\right)^{-1}
\end{aligned}
$$

where we have that if $b_{01} = 0$, then $\tilde{M}_{22} = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} + b_{02}b_{02}'$; while if $b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01} = 0$ then $\tilde{M}_{22} = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}$.

Since J_{01} and $\tilde{J}_{02}$ are equal to Breusch et al. (1999) D_1 and G_2 matrices respectively, the expression for the difference of AMSE is exactly equal to that for the difference in asymptotic variance in their paper up to $\tilde{M}_{22}$ which in their paper is equal to $\Sigma_{22} = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}$. Hence, the proof follows from the proof of their Theorem 1, since to establish the result the actual expression for Σ_{22} does not matter, we only need it to be non-singular. □

The next theorem presents a more general redundancy result, where we allow for non-zero asymptotic bias coming from both g_1 and g_2, and from the estimator of the

modified moments.

Theorem 3.2 (Case of non-zero bias)**.** *Under assumptions 3.1 - 3.10, the following statements are equivalent:*

(i) g_2 *is redundant given* g_1*.*

(ii) $\breve{J}_{02} := E[\partial r_{2M}(y_t, \theta_0)/\partial\theta'] = E\{\partial[g_2(y_t, \theta_0) - M_{21}M_{11}^{-1}g_1(y_t, \theta_0)]/\partial\theta'\} = 0.$

(iii) $J_{02} = M_{21}M_{11}^{-1}J_{01}$*.*

(iv) There exists a $k_1 \times p$ *matrix* A *such that* $J_{01} = M_{11}A$ *and* $J_{02} = M_{21}A$*.*

Remark 3.10. *From condition (iii), it is clear that when the asymptotic bias coming from* g_1 *is zero, that is* $b_{01} = 0$*, or the asymptotic bias coming from* r_2 *is zero, that is* $b_{02} - \Omega_{21}\Omega_{11}^{-1}b_{01} = 0$*, since we showed that* $M_{21}M_{11}^{-1} = \Omega_{21}\Omega_{11}^{-1}$ *we are back to Theorem 3.1 and thus to Breusch et al. (1999)'s result.*

Proof. *(i)* $\Leftrightarrow$ *(ii): The condition for redundancy of* g_2 *given* g_1 *is*

$$(J_{01}'M_{11}^{-1}J_{01})^{-1} = (J_0'M_0^{-1}J_0)^{-1},$$

or

$$J_{01}'M_{11}^{-1}J_{01} = J_0'M_0^{-1}J_0,$$

From equation (3.35) this is equivalent to

$$J_{01}'M_{11}^{-1}J_{01} = J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02}.$$

and therefore to

$$\breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02} = 0$$

This is possible if and only if $\breve{J}_{02} = 0$ *because* $\breve{J}'_{02}\breve{M}_{22}^{-1}\breve{J}_{02}$ *is p.s.d. and* $\breve{M}_{22}^{-1}$ *is nonsingular. Thus (i) and (ii) are equivalent.*

(ii) $\Leftrightarrow$ *(iii): This is straightforward from the definition of* $\breve{J}_{02}$ *in (3.34):* $\breve{J}_{02} = J_{02} - M_{21}M_{11}^{-1}J_{01} = 0$ *iff* $J_{02} = M_{21}M_{11}^{-1}J_{01}$.

(iii) $\Leftrightarrow$ *(iv): If (iii) holds, then (iv) holds with* $A = M_{11}^{-1}J_{01}$. *Conversely, if (iv) holds, then* $A = M_{11}^{-1}J_{01}$ *and* $J_{02} = M_{21}A = M_{21}M_{11}^{-1}J_{01}$. □

Remark 3.11. *Notice that condition (iii) can equivalently be stated in terms of* Ω *and* b_0 *as* $J_{02} = [\Omega_{21} + b_{02}b'_{01}][\Omega_{11} + b_{01}b'_{01}]^{-1}J_{01}$. *This is trivial, since by definition* $M_{21} = [\Omega_{21} + b_{02}b'_{01}]$ *and* $M_{11} = [\Omega_{11} + b_{01}b'_{01}]$, *therefore* $J_{02} = M_{21}M_{11}^{-1}J_{01}$ *iff* $J_{02} = [\Omega_{21} + b_{02}b'_{01}][\Omega_{11} + b_{01}b'_{01}]^{-1}J_{01}$.

3.3.2 Discussion: Two cases of Interest

In this subsection, we present a brief discussion on redundancy when we try to disentangle asymptotic variance and bias reduction. Since redundancy is established in terms of AMSE, it is natural to wonder whether we can say something solely in terms of variance or bias.

3.3.2.1 Zero Variance Reduction

Theorem 3.2 establishes the conditions under which g_2 is redundant given g_1. As explained before, in presence of asymptotic bias, this is done in terms of AMSE and not just variance. However, since in many situations we might wonder about the effect on the asymptotic variance in particular, it is then natural to wonder whether we can say something only in terms of variance reduction.

Imagine we are singularly interested in variance reduction, and we want to understand what the impact of adding a second set of moments is in the asymptotic variance, instead of focusing on the AMSE as a whole. We might be inclined to think that in order to get a zero variance reduction in this more general case, we should apply

the redundancy condition for the case where there is no bias, i.e., $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$.[6] The rationale for this thinking is that we would then be focusing on variance. However, we will see this is not in general the case. The following corollary provides a first insight:

Corollary 3.1. *$J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$ does not preclude variance reduction in general. If in addition, the asymptotic bias of the modified moments is zero, $b_{02} - M_{21}M_{11}^{-1}b_{01} = 0$, or, more generally, $\Omega_{11}^{-1}\Omega_{12} = M_{11}^{-1}M_{12}$, the variance reduction is zero. But this also delivers a zero AMSE reduction.*

The proof of this corollary is presented in the Appendix. The reader is encouraged to go through the details of the proof to get further insights on this result.

The implications of condition $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$ are worthy of further comment. Under this condition, we have that $\breve{J}_{02} = (\Omega_{21}\Omega_{11}^{-1} - M_{21}M_{11}^{-1})J_{01}$. Since $M_{21} = [\Omega_{21} + b_{02}b_{01}']$ and $M_{11} = [\Omega_{11} + b_{01}b_{01}']$ it is clear that in the case of zero asymptotic bias, or just $b_{01} = 0$ or $b_{02} = \Omega_{21}\Omega_{11}^{-1}b_{01}$, as discussed in Theorem 3.1, $\breve{J}_{02} = 0$. As a result, there will be zero reduction in variance, and also in AMSE.

Moreover, as will be shown later on, it is sufficient to have zero asymptotic bias in the modified moment conditions r_{2M} to achieve this result. This has the same flavor as Theorem 1 of Breusch et al. (1999) in the sense that there is no variance reduction from adding moments based on g_2 if under zero bias of the transformed moments, r_{2M}, we have that $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01} = 0$. This makes sense, since under zero bias of the transformed moments, the asymptotic covariance matrix $\breve{\Omega}$ becomes block-diagonal.

More precisely, notice that $\breve{\Omega}_{21} = \Omega_{21} - M_{21}M_{11}^{-1}\Omega_{11}$. Therefore, the condition $\Omega_{21} - M_{21}M_{11}^{-1}\Omega_{11} = 0$ amounts to having $\breve{\Omega}_{21} = 0$ and thus also $\breve{\Omega}_{12} = 0$. Remember that $\breve{\Omega}_{21}$ and $\breve{\Omega}_{12}$ are the covariance matrices between the moments based on g_1 and

[6] This is the same condition as in the zero bias case from g_1 or r_2 from Theorem 3.1, or the full zero bias case from Breusch et al. (1999).

those based on the modified moment conditions r_{2M}. In other words, what this result is telling us, is that if the modified moment conditions are orthogonal to the set of moments based on g_1, then we get redundancy under the simpler condition $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$.

3.3.2.2 Bias reduction

In the same way that we might be interested in variance reduction alone, it is also natural to wonder whether we can say something just in terms of bias reduction. Unfortunately, as suspected based on the above results for the variance reduction case, this is a difficult task. The corollary below provides a result in that direction:

Corollary 3.2. *A zero asymptotic bias of the estimator of the modified moment conditions* $E[r_{2M}(y_t;\theta)]$*, i.e.,* $\breve{b}_{02} = b_{02} - M_{21}M_{11}^{-1}b_{01} = 0$*, does not preclude, in general, a bias reduction.*

The proof of this corollary is given in the Appendix. The reader is again encouraged to go through the details of the proof to get further insights on this result.

Remark 3.12. *Notice, however, that if* $J_{02} = M_{21}M_{11}^{-1}J_{01}$ *then there will be not only no bias reduction, but also no AMSE reduction as shown in Theorem 3.2.*

We may also wonder if there is any particular situation in which we can get full bias reduction, i.e., $\breve{B}_\infty = 0$. From the proof of Corollary 3.2, we can write $\breve{B}_\infty = -(J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02})^{-1}(J_{01}'M_{11}^{-1}b_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{b}_{02})$. Given that the first factor of $\breve{B}_\infty$ is non-singular, $\breve{B}_\infty = 0$ if and only if the second factor is zero. Using, that $\breve{J}_{02} = J_{02} - M_{21}M_{11}^{-1}J_{01}$, and that $\breve{b}_{02} = b_{02} - M_{21}M_{11}^{-1}b_{01}$ we get that $\breve{B}_\infty = 0$ if

and only if

$$
\begin{aligned}
& J_{01}' M_{11}^{-1} b_{01} + \breve{J}_{02}' \breve{M}_{22}^{-1} \breve{b}_{02} \\
& = J_{01}' M_{11}^{-1} b_{01} + (J_{02} - M_{21} M_{11}^{-1} J_{01})' \breve{M}_{22}^{-1} (b_{02} - M_{21} M_{11}^{-1} b_{01}) \\
& = 0
\end{aligned}
$$

Clearly, $J_{01}' M_{11}^{-1} b_{01} \neq 0$, since this is the second factor for the asymptotic bias when we use only moments based on g_1. Hence, if $\breve{b}_{02} = b_{02} - M_{21} M_{11}^{-1} b_{01} = 0$, i.e., zero asymptotic bias from the modified moment conditions, then we won't be able to achieve full bias reduction. As a result, in order to achieve full bias reduction we will require the two terms above to have opposite signs.

From all the discussion in this subsection, it remains to say that in both Corollary 3.1 and Corollary 3.2 we can appreciate how the usual variance-bias trade-off plays a key role in making these corollaries hold.

3.4 Partial Redundancy in presence of Asymptotic Bias

This section can be seen as an extension of the redundancy results for the case when the focus is on a particular subset of the parameter set. Though we present here the concept of partial redundancy and some interesting additional results, they are immaterial to the main message of the paper which was presented in the previous section. Hence it is upon the reader's interests to go through the details of this section.

So far we have derived conditions for redundancy focusing on the entire parameter set. However, in many situations we are interested in the estimation of only a subset of the parameters, and we might even think of the rest as nuisance parameters. It

is then natural to wonder how the redundancy conditions would adapt in this case. That is, when is one set of moment conditions redundant given the other set but looking at only a subset of the parameters? This is known as partial redundancy.

In this section, we provide necessary and sufficient conditions for partial redundancy of one set of moments given the other. The basic approach consists of partitioning the parameter vector, and the corresponding Jacobian matrices in order to disentangle the conditions for one subset of parameters only.

Consider the following partition of the $p \times 1$ parameter vector θ

$$\theta = (\theta_1' \;\; \theta_2')' \tag{3.37}$$

Next, consider the corresponding partition for the limit Jacobian matrices J_0 and $\breve{J}_0$ from equation (3.14) and (3.34) respectively,

$$\begin{aligned} J_0 = E\left[\frac{\partial g(y_t;\theta_0)}{\partial \theta'}\right] &= E\begin{bmatrix} \dfrac{\partial g_1(y_t;\theta_0)}{\partial \theta_1'} & \dfrac{\partial g_1(y_t;\theta_0)}{\partial \theta_2'} \\ \dfrac{\partial g_2(y_t;\theta_0)}{\partial \theta_1'} & \dfrac{\partial g_2(y_t;\theta_0)}{\partial \theta_2'} \end{bmatrix} \\ := \begin{bmatrix} J_{01} \\ J_{02} \end{bmatrix} &:= \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \end{aligned} \tag{3.38}$$

and

$$\begin{aligned} \breve{J}_0 = E\left[\frac{\partial \psi(y_t;\theta_0)}{\partial \theta'}\right] &= \begin{bmatrix} J_{11} & J_{12} \\ J_{21} - M_{21}M_{11}^{-1}J_{11} & J_{22} - M_{21}M_{11}^{-1}J_{12} \end{bmatrix} \\ := \begin{bmatrix} \breve{J}_{01} \\ \breve{J}_{02} \end{bmatrix} &:= \begin{bmatrix} \breve{J}_{11} & \breve{J}_{12} \\ \breve{J}_{21} & \breve{J}_{22} \end{bmatrix} \end{aligned} \tag{3.39}$$

For now, we continue to assume that $\theta_0 = (\theta_{01}' \;\; \theta_{02}')'$ is identified by the first set of moments only. Hence we assume as before that $J_{01} = [J_{11} \;\; J_{12}]$ has a full column

rank, and thus we require $k_1 \geq p$. We will slightly relax this assumption later in this Section to allow only part of the parameter vector to be identified by the first set of moments.

In order to establish the partial redundancy results, we will focus on θ_1. First, we need to derive partitioned expressions for the AMSE and extract from there the AMSE for θ_1.

From Section 3.3, the AMSE of the GMM estimator $\hat{\theta}_T$ of θ based on the first set of moment conditions only, using weighting matrix M_{11}^{-1} is given by

$$\begin{aligned} AMSE[\sqrt{T}(\hat{\theta}_T - \theta_0)] &= (J_{01}' M_{11}^{-1} J_{01})^{-1} \\ &= \begin{bmatrix} J_{11}' M_{11}^{-1} J_{11} & J_{11}' M_{11}^{-1} J_{12} \\ J_{12}' M_{11}^{-1} J_{11} & J_{12}' M_{11}^{-1} J_{12} \end{bmatrix}^{-1} \end{aligned} \tag{3.40}$$

Then, the AMSE for $\hat{\theta}_{1T}$ will be given by the first block-partition of the inverse of the above matrix. That is,

$$\begin{aligned} AMSE[\sqrt{T}(\hat{\theta}_{1T} - \theta_{01})] &= (J_{11}' M_{11}^{-1} J_{11} - J_{11}' M_{11}^{-1} J_{12} (J_{12}' M_{11}^{-1} J_{12})^{-1} J_{12}' M_{11}^{-1} J_{11})^{-1} \\ &:= (\Sigma_{\theta_1, k_1})^{-1} \end{aligned} \tag{3.41}$$

where we have applied the partitioned-matrix inverse rule to get the result.

Similarly, from Section 3.3, the AMSE of the GMM estimator $\breve{\theta}_T$ of θ based on

the full set of moment conditions (3.29), using weighting matrix $\breve{M}_0^{-1}$ is given by

$$\begin{aligned} AMSE[\sqrt{T}(\breve{\theta}_T - \theta_0)] &= (\breve{J}_0' \breve{M}_0^{-1} \breve{J}_0)^{-1} \\ &= (J_{01}' M_{11}^{-1} J_{01} + \breve{J}_{02}' \breve{M}_{22}^{-1} \breve{J}_{02})^{-1} \\ &= \begin{bmatrix} J_{11}' M_{11}^{-1} J_{11} + \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{21} & J_{11}' M_{11}^{-1} J_{12} + \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{22} \\ J_{12}' M_{11}^{-1} J_{11} + \breve{J}_{22}' \breve{M}_{22}^{-1} \breve{J}_{21} & J_{12}' M_{11}^{-1} J_{12} + \breve{J}_{22}' \breve{M}_{22}^{-1} \breve{J}_{22} \end{bmatrix}^{-1} \end{aligned} \tag{3.42}$$

where we have used the expression for AMSE in (3.35) from Section 3.3. Therefore, the AMSE for $\breve{\theta}_{1T}$ will be given by the first block-partition of the inverse of the above matrix. That is,

$$\begin{aligned} &AMSE[\sqrt{T}(\breve{\theta}_{1T} - \theta_{01})] \\ &= \Big[(J_{11}' M_{11}^{-1} J_{11} + \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{21}) - (J_{11}' M_{11}^{-1} J_{12} + \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{22}) \\ &\quad \times (J_{12}' M_{11}^{-1} J_{12} + \breve{J}_{22}' \breve{M}_{22}^{-1} \breve{J}_{22})^{-1} (J_{12}' M_{11}^{-1} J_{11} + \breve{J}_{22}' \breve{M}_{22}^{-1} \breve{J}_{21}) \Big]^{-1} \\ &:= (\Sigma_{\theta_1,k})^{-1} \end{aligned} \tag{3.43}$$

where we have applied again the partitioned-matrix inverse rule.

From the above expressions for AMSE, (3.41) and (3.43), a necessary and sufficient condition for g_2 to be partially redundant given g_1 for θ_1, is that $\Sigma_{\theta_1,k_1} = \Sigma_{\theta_1,k}$. This gives rise to our next definition,

Definition 3.3 (**Partial Redundancy for θ_1**). *We say that g_2 is partially redundant for the estimation of θ_1 given g_1, if the optimal GMM estimator of θ_1 based on the first set of moments in (3.29) only, has the same AMSE as the optimal GMM estimator of θ_1 based on the full set of moments in (3.29). That is, if $\Sigma_{\theta_1,k_1} = \Sigma_{\theta_1,k}$.*

Before stating the result, we will need the following lemma

Lemma 3.2. *Let Assumptions 3.1 - 3.10 hold. Define*

$$A := J_{12}' M_{11}^{-1} J_{12} + \breve{J}_{22}' \breve{M}_{22}^{-1} \breve{J}_{22} \tag{3.44a}$$

$$B := \breve{M}_{22}^{-1} - \breve{M}_{22}^{-1} \breve{J}_{22} A^{-1} \breve{J}_{22}' \breve{M}_{22}^{-1} \tag{3.44b}$$

$$C := \breve{J}_{21} - \breve{J}_{22} (J_{12}' M_{11}^{-1} J_{12})^{-1} (J_{12}' M_{11}^{-1} J_{11}). \tag{3.44c}$$

Then B is positive definite, and $\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} = C'BC$.

The proof of this lemma is given in the Appendix.

Now we are in the position to present our partial redundancy result,

Theorem 3.3. *Under Assumptions 3.1 - 3.10, the following statements are equivalent*

(i) g_2 *is partially redundant, for the estimation of* θ_1*, given* g_1*.*

(ii) $\breve{J}_{21} = \breve{J}_{22} (J_{12}' M_{11}^{-1} J_{12})^{-1} (J_{12}' M_{11}^{-1} J_{11})$.

(iii) $J_{21} - M_{21} M_{11}^{-1} J_{11} = (J_{22} - M_{21} M_{11}^{-1} J_{12})(J_{12}' M_{11}^{-1} J_{12})^{-1} (J_{12}' M_{11}^{-1} J_{11})$

Proof. *By definition, g_2 is partially redundant, for the estimation of θ_1, given g_1 when $\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} = 0$. From Lemma 2, $\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} = C'BC$ and B is positive definite. Hence, $\Sigma_{\theta_1,k} = \Sigma_{\theta_1,k_1}$ if and only if $C = 0$. Conditions (ii) and (iii) are just restatements of $C = 0$.* □

Remark 3.13. *It is important to realize the relationship between redundancy and partial redundancy. From the comparison of Theorem 3.3 and Theorem 3.2, we can notice that if g_2 is redundant given g_1, we must have that g_2 is partially redundant given g_1, for any subset of θ. This is the case since, g_2 is redundant given g_1 if $\breve{J}_{02} = [\breve{J}_{21} \;\; \breve{J}_{22}] = 0$ which clearly implies the partial redundancy condition (ii).*

3.4.1 Special Case of Partial Redundancy

So far we have always maintained the assumption that the full set of true values of the parameter, θ_0, is identified from the first set of moments only. However, there are situations in which this is not the case. That is, the first set of moments could only identify part of the set of parameters. For instance, it could only depend on part of the parameter set, say θ_1. In this situation, our previous results will not hold. Hence, it is interesting to study what it entails to achieve partial redundancy when this assumption is relaxed.

For this purpose, we will focus on the simple case in which the first set of moment conditions depends only on θ_1, while the second set of moment conditions depends on both θ_1 and θ_2. That is, the set of moment conditions is now given by

$$\begin{cases} E[g_1(y_t;\theta_{01})] = 0 \\ E[g_2(y_t;\theta_{01},\theta_{02})] = 0 \qquad for \;\; t = 1,...,T \end{cases} \tag{3.45}$$

Remark 3.14. *This analysis could alternatively be carried out by using the GMM equivalence of estimators when we concentrate out nuisance parameters (e.g., by replacing them with an estimate). For instance, Crepon, Kramarz, and Trognon (1997) presents a good discussion on elimination of nuisance parameters, though in a case with no asymptotic bias. This alternative approach goes beyond the scope of this paper, hence the extension of their results is left for future research.*

We continue with the notation from the previous section, with the caveat that now $J_{12} = 0$ and as a result $\breve{J}_{22} = J_{22}$. Our identification assumptions are now slightly modified in the following manner:

Assumption 3.11. *The true value of the parameters θ_{01} and θ_{02} are fully identified from the first and second set of moments only respectively.*

Assumption 3.12. *J_{11} and J_{22} are of full column rank.*

The definition below states what partial redundancy entails in this modified setup,

Definition 3.4 (**Partial Redundancy for** θ_1)**.** *We say that* g_2 *is partially redundant for the estimation of* θ_1 *given* g_1*, if the optimal GMM estimator of* θ_1 *based on the first set of moments in* (3.45) *only, has the same AMSE as the optimal GMM estimator of* θ_1 *based on the full set of moments in* (3.45)*.*

Remark 3.15. *If* g_1 *and* g_2 *are of dimension* k_1 *and* k_2 *respectively, and* θ_1 *and* θ_2 *are of dimension* p_1 *and* p_2 *respectively, the identification assumption above requires that* $k_2 \geq p_2$*. In particular, if* $k_2 = p_2$ *we could think, as in the case with no asymptotic bias (e.g., Ahu and Schmidt, 1995), that adding* g_2 *does not affect the GMM estimate of* θ_1*, that is,* g_2 *is partially redundant given* g_1*. This can be easily shown since from equation* (3.43)*:*

$$\begin{aligned} AMSE[\sqrt{T}(\breve{\theta}_{1T} - \theta_{01})] = [&J_{11}' M_{11}^{-1} J_{11} + \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{21} - \breve{J}_{21}' \breve{M}_{22}^{-1} J_{22} (J_{22}' \breve{M}_{22}^{-1} J_{22})^{-1} \\ &\times J_{22}' \breve{M}_{22}^{-1} \breve{J}_{21}]^{-1} \end{aligned}$$

where we have used that in this case $J_{12} = 0$ *and* $\breve{J}_{22} = J_{22}$*.*

Now if $k_2 = p_2$ *then* J_{22} *is a non-singular square matrix. As a result, we can apply distributive property of the inverse of a product of non-singular matrices to* $(J_{22}' \breve{M}_{22}^{-1} J_{22})^{-1}$ *in the above expression:*

$$\begin{aligned} AMSE[\sqrt{T}(\breve{\theta}_{1T} - \theta_{01})] &= [J_{11}' M_{11}^{-1} J_{11} + \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{21} - \breve{J}_{21}' \breve{M}_{22}^{-1} J_{22} J_{22}^{-1} \breve{M}_{22} (J_{22}')^{-1} \\ &\quad \times J_{22}' \breve{M}_{22}^{-1} \breve{J}_{21}]^{-1} \\ &= [J_{11}' M_{11}^{-1} J_{11} + \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{21} - \breve{J}_{21}' \breve{M}_{22}^{-1} \breve{J}_{21}]^{-1} \\ &= [J_{11}' M_{11}^{-1} J_{11}]^{-1} \end{aligned}$$

which coincides with the AMSE in the case when we use moments based on g_1 *only. Hence, as expected,* g_2 *would be partially redundant, for the estimation of* θ_1*, in this*

case.

However, when $k_2 > p_2$, adding the second set of moments could improve the estimation of θ_1 or not. We are interested in recognizing circumstances in which such an improvement does not take place.

Theorem 3.4. *Under Assumptions 3.2-3.9, 3.11, and 3.12, the following statements are equivalent:*

(i) g_2 is partially redundant for the estimation of θ_1 given g_1.

(ii) $\breve{J}_{21} = J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}$.

(iii) There exists a $p_2 \times p_1$ matrix R such that $\breve{J}_{21} = J_{22}R$.

where $\breve{J}_{21} = J_{21} - M_{21}M_{11}^{-1}J_{11}$ and $\breve{M}_{22} = M_{22} - M_{21}M_{11}^{-1}M_{12}$ as defined in equations (3.39) and (3.33) respectively.

Proof. *The AMSE of the GMM estimator of θ_1 based on the first set of moment conditions in (3.45) is given by*

$$\begin{aligned} AMSE[\sqrt{T}(\hat{\theta}_{1T} - \theta_{01})] &= (J_{11}'M_{11}^{-1}J_{11})^{-1} \\ &:= (\Sigma_{\theta_1,k_1})^{-1} \end{aligned} \tag{3.46}$$

since $J_{12} = 0$.

The AMSE of the GMM estimator of θ_1 based on the full set of moment conditions in (3.45) is given by

$$\begin{aligned} AMSE[\sqrt{T}(\breve{\theta}_{1T} - \theta_{01})] &= \Big[J_{11}'M_{11}^{-1}J_{11} + \breve{J}_{21}'\breve{M}_{22}^{-1}\breve{J}_{21} - \breve{J}_{21}'\breve{M}_{22}^{-1}J_{22} \\ &\quad \times (J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}\Big]^{-1} \\ &:= (\Sigma_{\theta_1,k})^{-1} \end{aligned} \tag{3.47}$$

where we have used again that $J_{12} = 0$*, and that* $\breve{J}_{22} = J_{22}$*.*

Therefore, we get that

$$\begin{aligned}
\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} &= \breve{J}_{21}'\breve{M}_{22}^{-1}\breve{J}_{21} - \breve{J}_{21}'\breve{M}_{22}^{-1}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21} \\
&= \breve{J}_{21}'\left[\breve{M}_{22}^{-1} - \breve{M}_{22}^{-1}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\right]\breve{J}_{21} \\
&= \breve{J}_{21}'\breve{M}_{22}^{-1/2}\left[I - \breve{M}_{22}^{-1/2}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1/2}\right]\breve{M}_{22}^{-1/2}\breve{J}_{21}
\end{aligned} \tag{3.48}$$

Since $L = [I - \breve{M}_{22}^{-1/2}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1/2}] = [I - P]$ *is a projection matrix onto the space orthogonal to* $\breve{M}_{22}^{-1/2}J_{22}$ *, it is symmetric and idempotent. As a result, we can use that* $L = LL$ *in the above expression to obtain*

$$\begin{aligned}
&\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} \\
&= \breve{J}_{21}'\breve{M}_{22}^{-1/2}LL\breve{M}_{22}^{-1/2}\breve{J}_{21} \\
&= \breve{J}_{21}'\breve{M}_{22}^{-1/2}\left[I - \breve{M}_{22}^{-1/2}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1/2}\right] \\
&\quad \times \left[I - \breve{M}_{22}^{-1/2}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1/2}\right]\breve{M}_{22}^{-1/2}\breve{J}_{21}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} \\
&= \left[\breve{J}_{21}'\breve{M}_{22}^{-1/2} - \breve{J}_{21}'\breve{M}_{22}^{-1}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1/2}\right] \\
&\quad \times \left[\breve{M}_{22}^{-1/2}\breve{J}_{21} - \breve{M}_{22}^{-1/2}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}\right] \\
&= \left[\breve{J}_{21}' - \breve{J}_{21}'\breve{M}_{22}^{-1}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\right]\breve{M}_{22}^{-1}\left[\breve{J}_{21} - J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}\right] \\
&= \left[\breve{J}_{21} - J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}\right]'\breve{M}_{22}^{-1}\left[\breve{J}_{21} - J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}\right]
\end{aligned} \tag{3.49}$$

which is positive semi-definite. Therefore, adding the second set of moments cannot hurt the AMSE of the estimation of θ_1.

(i) $\Leftrightarrow$ *(ii): The condition for partial redundancy of* g_2 *for the estimation of* θ_1 *given* g_1 *is*

$$\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} = 0$$

From equation (3.49) this occurs if and only if

$$\left[\breve{J}_{21} - J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}\right] = 0$$

because the expression in equation 3.49 is p.s.d. and $\breve{M}_{22}^{-1}$ *is non-singular. This is equal to condition (ii). Thus (i) and (ii) are equivalent.*

(ii) $\Leftrightarrow$ *(iii): If (ii) holds, then (iii) holds with* $R = (J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}$. *Conversely, if (iii) holds, then* $R = (J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}$ *and* $\breve{J}_{21} = J_{22}R = J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}$. □

Remark 3.16. *Notice that there is also a direct interpretation to the equivalence between (i) and (iii).*

From (3.49) we also have that

$$\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} = \breve{J}_{21}'\breve{M}_{22}^{-1/2}[I - P]\breve{M}_{22}^{-1/2}\breve{J}_{21}$$

where $P = [\breve{M}_{22}^{-1/2}J_{22}(J_{22}'\breve{M}_{22}^{-1}J_{22})^{-1}J_{22}'\breve{M}_{22}^{-1/2}]$ *is the projection onto the space spanned by the columns of* $\breve{M}_{22}^{-1/2}J_{22}$.

Hence, $\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} = 0$ *if and only if* $\breve{M}_{22}^{-1/2}\breve{J}_{21}$ *is in the column space of* $\breve{M}_{22}^{-1/2}J_{22}$, *or* $\breve{J}_{21}$ *is in the column space of* J_{22}, *which is condition (iii).*

3.5 Concluding Remarks

The literature on redundancy until now has been based on asymptotically unbiased estimators of the moment conditions, leading to a concept of redundancy and partial redundancy defined in terms of efficiency of the estimator of the parameters only. However, in the last decade, asymptotically biased estimators have started to gain some support in estimation procedures. Hence the need to revisit these concepts within this broader class of estimators, which includes asymptotically unbiased estimators as a particular case.

In this paper, we focused on valid moment conditions under the caveat that its estimators were asymptotically biased. Therefore we extended the redundancy literature to contemplate these cases. We revisited the concept of redundancy and partial redundancy using AMSE as a criterion. First, we argued that the GMM optimal weighting matrix should now be such that it minimizes AMSE instead of asymptotic variance. As a result, we redefined this matrix to be equal to the inverse of the AMSE, instead of asymptotic variance, of the estimator of the moment conditions. Next, we showed that in this more general framework adding valid moments cannot hurt in terms of AMSE. This parallels the well known result, in standard GMM settings, that adding valid moments can never hurt the asymptotic efficiency of the optimal GMM estimator.

Finally, we derived conditions for *redundancy* and *partial redundancy* in this broader framework to assess the information that extra valid moment conditions add to the estimation. We intended to deliver the idea that in certain situations we can take advantage of asymptotic bias, and hence it is not always detrimental in terms of AMSE reduction. In any case, we believe that extending the existent theory to work with asymptotically biased estimators is a challenging topic for future research.

3.A Appendix

In this appendix we present the proof of Lemmas and Theorems that were immaterial to the main message of this paper and thus were relegated to the appendix.

3.A.1 GMM Estimator under Non-zero Asymptotic Bias

In this subsection we present a sketch of the proof for the derivation of the asymptotic distribution of the GMM estimator under non-zero asymptotic bias.

Proof (Sketch of the proof)**.** *Let $\ddot{\theta}_T$ be the GMM estimator of θ based on the full set of moment conditions* (3.1)*. The minimization problem from Section 3.2 is given by*

$$\ddot{\theta}_T = arg \min_{\theta \in \Theta} Q_T(\theta),$$

with

$$\begin{aligned} Q_T(\theta) &= \bar{g}(\theta)' W \bar{g}(\theta) \\ \bar{g}(\theta) &= (\hat{E}[g_1(y_t;\theta)]' \quad \hat{E}[g_2(y_t;\theta)]')' \end{aligned}$$

where $\hat{E}[g_i(y_t;\theta)]$ denotes the sample average estimator of $E[g_i(y_t;\theta_0)]$, for $i = 1, 2$, and W is a p.d. weighting matrix.

Notice that this would deliver an unfeasible estimator since the weighting matrix W is a population object as it is standard in GMM. As a result, we replace W with a consistent estimator W_T and derive the asymptotic distribution for the feasible estimator as it is standard in the literature.

From the first order condition $\dfrac{\partial Q_T}{\partial \theta}(\ddot{\theta}_T) = 0$ and a mean value expansion of $\bar{g}(\ddot{\theta}_T)$

around θ_0 *we have that*

$$\frac{\partial Q_T}{\partial \theta}(\ddot{\theta}_T) = \frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T \bar{g}(\ddot{\theta}_T) = 0$$

$$\Rightarrow \frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T \bar{g}(\theta_0) + \frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T \frac{\partial \bar{g}}{\partial \theta'}(\bar{\theta}_T)(\ddot{\theta}_T - \theta_0) = 0 \tag{3.50}$$

where $\bar{\theta}_T$ *is between* $\ddot{\theta}_T$ *and* θ_0 *component–wise.*

We pre-multiply (3.50) *by* $\sqrt{T}$ *:*

$$\frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T \sqrt{T} \bar{g}(\theta_0) + \frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T \frac{\partial \bar{g}}{\partial \theta'}(\bar{\theta}_T) \sqrt{T}(\ddot{\theta}_T - \theta_0) = 0 \tag{3.51}$$

We rewrite the above expression as

$$\sqrt{T}(\ddot{\theta}_T - \theta_0) = -\left[\frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T \frac{\partial \bar{g}}{\partial \theta'}(\bar{\theta}_T)\right]^{-1} \frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T \sqrt{T} \bar{g}(\theta_0) \tag{3.52}$$

By assumption 3.4 and 3.5 we have that $\sqrt{T}\bar{g}(\theta_0) \xrightarrow{d} \mathcal{N}(b_0,\ \Omega)$.

To conclude the proof we need the following additional lemma:

Lemma 3.3. *Suppose Assumptions 3.1, 3.2, and 3.10 are satisfied. If (i)* $\operatorname{plim}_{T\to\infty} \ddot{\theta} = \theta_0$*; and (ii)* $\operatorname{plim}_{T\to\infty} W_T = W$*; then* $\operatorname{plim}_{T\to\infty} \frac{\partial \bar{g}}{\partial \theta'}(\ddot{\theta}_T) = \operatorname{plim}_{T\to\infty} \frac{\partial \bar{g}}{\partial \theta'}(\bar{\theta}_T) = J_0$ *and* $\operatorname{plim}_{T\to\infty} \frac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T) W_T = J_0' W$*, where* J_0 *is the limit Jacobian matrix as defined in* (3.14) *in Section 3.3.*

Proof (Sketch of the proof of Lemma A1)**.** *It is enough to show that* $\operatorname{plim}_{T\to\infty} \frac{\partial \bar{g}}{\partial \theta'}(\bar{\bar{\theta}}_T) = J_0$ *for any* $\bar{\bar{\theta}}_T$ *such that* $\| \bar{\bar{\theta}}_T - \theta_0 \| = O_p(\frac{1}{\sqrt{T}})$*. Since* $\bar{g}$ *is a sample average, under assumptions 3.1, 3.2, and 3.10, by the LLN for stationary and ergodic processes,*

$$\frac{\partial \bar{g}}{\partial \theta'}(\bar{\bar{\theta}}_T) = \begin{bmatrix} \hat{E}\left(\frac{\partial g_1}{\partial \theta'}(\bar{\bar{\theta}}_T)\right) \\ \hat{E}\left(\frac{\partial g_2}{\partial \theta'}(\bar{\bar{\theta}}_T)\right) \end{bmatrix} \xrightarrow{p} \begin{bmatrix} E\left(\frac{\partial g_1}{\partial \theta'}(\theta_0)\right) \\ E\left(\frac{\partial g_2}{\partial \theta'}(\theta_0)\right) \end{bmatrix} = E\left[\frac{\partial g}{\partial \theta'}(\theta_0)\right] = J_0 \tag{3.53}$$

Therefore, since W_T is a consistent estimator of W, it is straightforward to show that $\text{plim}_{T\to\infty} \dfrac{\partial \bar{g}'}{\partial \theta}(\ddot{\theta}_T)W_T = J_0'W$. □

Hence, by Lemma 3.3, under assumptions 3.1 - 3.10 and applying a CLT for ergodic stationary m.d.s. processes (e.g., Hayashi, 2000, p. 106) the result follows. □

3.A.2 Proof of Lemma 3.1

Proof. *Let*

$$Q = \begin{bmatrix} I_{k_1} & 0 \\ -M_{21}M_{11}^{-1} & I_{k_2} \end{bmatrix} \tag{3.54}$$

Notice that we can write $\psi(y_t;\theta_0) = Q\ g(y_t;\theta_0)$. Since Q is non-singular, it follows that the optimal GMM estimator for θ based on $E[g(y_t,\theta_0)] = 0$ is the same as the optimal GMM estimator for θ based on $E[\psi(y_t;\theta_0)] = 0$. □

3.A.3 Proof of Corollary 3.1

Proof. *The asymptotic variance of $\breve{\theta}_T$ is given by*

$$\begin{aligned}
&Avar[\sqrt{T}(\breve{\theta}_T - \theta_0)] \\
&:= (\breve{J}_0'\breve{M}_0^{-1}\breve{J}_0)^{-1}\breve{J}_0'\breve{M}_0^{-1}\breve{\Omega}\breve{M}_0^{-1}\breve{J}_0(\breve{J}_0'\breve{M}_0^{-1}\breve{J}_0)^{-1} \\
&:= (J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02})^{-1}\Big[J_{01}'M_{11}^{-1}\Omega_{11}M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{\Omega}_{21}M_{11}^{-1}J_{01} \\
&\quad + J_{01}'M_{11}^{-1}\breve{\Omega}_{12}\breve{M}_{22}^{-1}\breve{J}_{02} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{\Omega}_{22}\breve{M}_{22}^{-1}\breve{J}_{02}\Big](J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02})^{-1}
\end{aligned} \tag{3.55}$$

where we have used the expressions for $\breve{M}_0^{-1}$, $\breve{J}_0$, and $\breve{\Omega}$ given in equations (3.33), (3.34), (3.32) respectively; we have also used the following equivalences: $\breve{M}_{11}^{-1} = M_{11}^{-1}$, $\breve{J}_{01} = J_{01}$, and $\breve{\Omega}_{11} = \Omega_{11}$.

If $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$, *then*

$$\begin{aligned}\breve{J}_{02} &= J_{02} - M_{21}M_{11}^{-1}J_{01} \\ &= \Omega_{21}\Omega_{11}^{-1}J_{01} - M_{21}M_{11}^{-1}J_{01} \\ &= (\Omega_{21}\Omega_{11}^{-1} - M_{21}M_{11}^{-1})J_{01} \end{aligned} \tag{3.56}$$

We want to compare the asymptotic variance in equation (3.55)*, under the condition in* (3.56)*, to the one obtained when we only use moments based on* g_1*. This last asymptotic variance is given from equation* (3.15) *by*

$$Avar[\sqrt{T}(\hat{\theta}_T - \theta_0)] = (J_{01}'M_{11}^{-1}J_{01})^{-1}J_{01}'M_{11}^{-1}\Omega_{11}M_{11}^{-1}J_{01}(J_{01}'M_{11}^{-1}J_{01})^{-1} \tag{3.57}$$

For no variance reduction from adding moments based on g_2*, we need that*

$$Avar[\sqrt{T}(\hat{\theta}_T - \theta_0)] - Avar[\sqrt{T}(\breve{\theta}_T - \theta_0)] = 0 \tag{3.58}$$

However, in general, it is not obvious that these two expressions for the asymptotic variance will coincide. In fact, each term in (3.55) *is a quadratic form, and thus this expression will not reduce to* (3.57) *unless the extra terms are equal to zero. If we impose the condition* $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$*, this will only imply that* $\breve{J}_{02} = (\Omega_{21}\Omega_{11}^{-1} - M_{21}M_{11}^{-1})J_{01}$*, which in general will not be zero. As a result, in contrast with the zero bias case, we cannot preclude variance reduction even when* $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$*. It seems quite obvious that it would certainly be the case if* $\breve{J}_{02} = 0$*, but in this situation we would also have zero gain in AMSE based on Theorem 3.2.*

Hence, one would like to find additional conditions to $J_{02} = \Omega_{21}\Omega_{11}^{-1}J_{01}$ *under which* (3.58) *holds. It is straight forward to show that if in addition* $b_{02} - M_{21}M_{11}^{-1}b_{01} = 0$ *or* $\Omega_{12} - \Omega_{11}M_{11}^{-1}M_{12} = 0$*, then there will be zero variance reduction.*

If in addition the asymptotic bias of the modified moments, $b_{02} - M_{21}M_{11}^{-1}b_{01}$*, is*

zero, then using the fact that $\breve{M}_{12} = 0$ *we get*

$$\breve{M}_{12} = \Omega_{12} - \Omega_{11}M_{11}^{-1}M_{12} + b_{02} - M_{21}M_{11}^{-1}b_{01} = 0$$

As a result, $\Omega_{12} - \Omega_{11}M_{11}^{-1}M_{12} = 0$*, which can also be written as* $\Omega_{21}\Omega_{11}^{-1} = M_{21}M_{11}^{-1}$*. If we substitute this in the expression for* $\breve{J}_{02}$ *we get:*

$$\begin{aligned}\breve{J}_{02} &= (\Omega_{21}\Omega_{11}^{-1} - M_{21}M_{11}^{-1})J_{01} \\ &= 0\end{aligned}$$

Hence, we get zero variance reduction. However, we are back to the zero AMSE reduction case from Theorem 3.2. That is, we also have no bias reduction. □

3.A.4 Proof of Corollary 3.2

Proof. *On the one hand, mimicking equation (3.10) it is clear that the asymptotic bias using the modified moments is given by*

$$\begin{aligned}\breve{B}_\infty &= -(\breve{J}_0'\breve{M}_0^{-1}\breve{J}_0)^{-1}\breve{J}_0'\breve{M}_0^{-1}\breve{b}_0 \\ &= -(J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02})^{-1}(J_{01}'M_{11}^{-1}b_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{b}_{02})\end{aligned} \tag{3.59}$$

where we have used that $\breve{J}_{01} = J_{01}$*,* $\breve{b}_{01} = b_{01}$*,* $\breve{M}_0$ *is block-diagonal with* $\breve{M}_{11} = M_{11}$*, and the partition for* $\breve{b}_0$*.*

On the other hand, from equation (3.17) the asymptotic bias, when we use moments based on g_1 *alone, is given by* $B_{1\infty} = -(J_{01}'M_{11}^{-1}J_{01})^{-1}J_{01}'M_{11}^{-1}b_{01}$*.*

If we compare $B_{1\infty}$ *and* $\breve{B}_\infty$*, we see that even if the asymptotic bias from the modified moments is zero, i.e.,* $\breve{b}_{02} = b_{02} - M_{21}M_{11}^{-1}b_{01} = 0$*, there is still hope for bias*

reduction:

$$\begin{aligned}\breve{B}_\infty &= -(J_{01}'M_{11}^{-1}J_{01} + \breve{J}_{02}'\breve{M}_{22}^{-1}\breve{J}_{02})^{-1}J_{01}'M_{11}^{-1}b_{01} \\ &= -[J_{01}'M_{11}^{-1}J_{01} + (J_{02} - M_{21}M_{11}^{-1}J_{01})'\breve{M}_{22}^{-1}(J_{02} - M_{21}M_{11}^{-1}J_{01})]^{-1}J_{01}'M_{11}^{-1}b_{01}\end{aligned}$$

where we have used that $\breve{J}_{02} = J_{02} - M_{21}M_{11}^{-1}J_{01}$.

As a result, since $(J_{02} - M_{21}M_{11}^{-1}J_{01})'\breve{M}_{22}^{-1}(J_{02} - M_{21}M_{11}^{-1}J_{01})$ *is a quadratic form and* $\breve{M}_{22}^{-1}$ *is non-singular, unless* $J_{02} = M_{21}M_{11}^{-1}J_{01}$ *and hence* $\breve{J}_{02} = 0$*, this term won't be zero. In other words, even if the asymptotic bias of the modified moment conditions is zero, we will get in general a bias reduction.* □

3.A.5 Proof of Lemma 3.2

Proof. *For the proof of Lemma 2 we will first state the following result about inversion of sum of matrices*[7] *(e.g., Leamer, 1978, p. 324).*

Lemma 3.4 (Matrix Inversion Lemma)**.** *If* H*,* K *and* $H + FKF'$ *are non-singular, then*

$$(H + FKF')^{-1} = H^{-1} - H^{-1}F(F'H^{-1}F + K^{-1})^{-1}F'H^{-1}. \tag{3.60}$$

We first prove that B *in equation* (3.44b) *is nonsingular, which implies that it is p.d. Note that* $\breve{M}_{22}$ *and* A *are non-singular, and that* J_{12} *is of full column rank so* $J_{12}'M_{11}^{-1}J_{12}$ *is non-singular. Then, by the Matrix Inversion Lemma,*

$$\begin{aligned}B &:= \breve{M}_{22}^{-1} - \breve{M}_{22}^{-1}\breve{J}_{22}A^{-1}\breve{J}_{22}'\breve{M}_{22}^{-1} \\ &= \breve{M}_{22}^{-1} - \breve{M}_{22}^{-1}\breve{J}_{22}(\breve{J}_{22}'\breve{M}_{22}^{-1}\breve{J}_{22} + J_{12}'M_{11}^{-1}J_{12})^{-1}\breve{J}_{22}'\breve{M}_{22}^{-1} \\ &= [\breve{M}_{22} + \breve{J}_{22}(J_{12}'M_{11}^{-1}J_{12})^{-1}\breve{J}_{22}']^{-1}\end{aligned} \tag{3.61}$$

[7] This lemma is a special case of the Woodbury matrix identity, also known as Sherman-Morrison-Woodbury formula or just Woodbury formula.

Hence B is non-singular, and therefore p.d.

In order to prove that $\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} = C'BC$ we will write both expressions and establish the equality term by term.

On the one hand, from (3.41) and (3.43) we have

$$\begin{aligned}
&\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1} \\
&= \Big[(J_{11}'M_{11}^{-1}J_{11} + \breve{J}_{21}'\breve{M}_{22}^{-1}\breve{J}_{21}) - (J_{11}'M_{11}^{-1}J_{12} + \breve{J}_{21}'\breve{M}_{22}^{-1}\breve{J}_{22}) \\
&\quad \times (J_{12}'M_{11}^{-1}J_{12} + \breve{J}_{22}'\breve{M}_{22}^{-1}\breve{J}_{22})^{-1}(J_{12}'M_{11}^{-1}J_{11} + \breve{J}_{22}'\breve{M}_{22}^{-1}\breve{J}_{21})\Big] \\
&\quad - \left[J_{11}'M_{11}^{-1}J_{11} - J_{11}'M_{11}^{-1}J_{12}(J_{12}'M_{11}^{-1}J_{12})^{-1}J_{12}'M_{11}^{-1}J_{11}\right] \\
&= \left[\breve{J}_{21}'\breve{M}_{22}^{-1}\breve{J}_{21} - \breve{J}_{21}'\breve{M}_{22}^{-1}\breve{J}_{22}A^{-1}\breve{J}_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}\right] \\
&\quad + \left[J_{11}'M_{11}^{-1}J_{12}(J_{12}'M_{11}^{-1}J_{12})^{-1}J_{12}'M_{11}^{-1}J_{11} - J_{11}'M_{11}^{-1}J_{12}A^{-1}J_{12}'M_{11}^{-1}J_{11}\right] \\
&\quad - \breve{J}_{21}'\breve{M}_{22}^{-1}\breve{J}_{22}A^{-1}J_{12}'M_{11}^{-1}J_{11} \\
&\quad - J_{11}'M_{11}^{-1}J_{12}A^{-1}\breve{J}_{22}'\breve{M}_{22}^{-1}\breve{J}_{21}
\end{aligned} \tag{3.62}$$

where A is defined in (3.44a).

On the other hand, we have that

$$\begin{aligned}
&C'BC \\
&= \left[\breve{J}_{21} - \breve{J}_{22}(J_{12}'M_{11}^{-1}J_{12})^{-1}(J_{12}'M_{11}^{-1}J_{11})\right]' B \left[\breve{J}_{21} - \breve{J}_{22}(J_{12}'M_{11}^{-1}J_{12})^{-1}(J_{12}'M_{11}^{-1}J_{11})\right] \\
&= \breve{J}_{21}'B\breve{J}_{21} \\
&\quad + J_{11}'M_{11}^{-1}J_{12}(J_{12}'M_{11}^{-1}J_{12})^{-1}\breve{J}_{22}'B\breve{J}_{22}(J_{12}'M_{11}^{-1}J_{12})^{-1}J_{12}'M_{11}^{-1}J_{11} \\
&\quad - \breve{J}_{21}'B\breve{J}_{22}(J_{12}'M_{11}^{-1}J_{12})^{-1}J_{12}'M_{11}^{-1}J_{11} \\
&\quad - J_{11}'M_{11}^{-1}J_{12}(J_{12}'M_{11}^{-1}J_{12})^{-1}\breve{J}_{22}'B\breve{J}_{21}
\end{aligned} \tag{3.63}$$

where B and C are defined in equations (3.44b) and (3.44c) respectively.

The first term of (3.62) is equal to the first term of (3.63), since it can be rewritten

as

$$\begin{aligned}
&\left[\breve{J}'_{21}\breve{M}_{22}^{-1}\breve{J}_{21} - \breve{J}'_{21}\breve{M}_{22}^{-1}\breve{J}_{22}A^{-1}\breve{J}'_{22}\breve{M}_{22}^{-1}\breve{J}_{21}\right] \\
&= \breve{J}'_{21}\left[\breve{M}_{22}^{-1} - \breve{M}_{22}^{-1}\breve{J}_{22}A^{-1}\breve{J}'_{22}\breve{M}_{22}^{-1}\right]\breve{J}_{21} \\
&= \breve{J}'_{21}B\breve{J}_{21}
\end{aligned}$$

The second terms are also equal, since the second term of (3.62) *can be rewritten as*

$$\begin{aligned}
&J'_{11}M_{11}^{-1}J_{12}(J'_{12}M_{11}^{-1}J_{12})^{-1}J'_{12}M_{11}^{-1}J_{11} - J'_{11}M_{11}^{-1}J_{12}A^{-1}J'_{12}M_{11}^{-1}J_{11} \\
&= J'_{11}M_{11}^{-1}J_{12}LJ'_{12}M_{11}^{-1}J_{11} \qquad (3.64)
\end{aligned}$$

where $L := (J'_{12}M_{11}^{-1}J_{12})^{-1} - A^{-1}$. *Now, using the Matrix Inversion Lemma in* (3.60) *we can rewrite* A^{-1} *as*

$$\begin{aligned}
A^{-1} &= \left[J'_{12}M_{11}^{-1}J_{12} + \breve{J}'_{22}\breve{M}_{22}^{-1}\breve{J}_{22}\right]^{-1} \\
&= (J'_{12}M_{11}^{-1}J_{12})^{-1} - (J'_{12}M_{11}^{-1}J_{12})^{-1}\breve{J}'_{22}N^{-1}\breve{J}_{22}(J'_{12}M_{11}^{-1}J_{12})^{-1} \qquad (3.65)
\end{aligned}$$

where $N := \breve{J}_{22}(J'_{12}M_{11}^{-1}J_{12})^{-1}\breve{J}'_{22} + \breve{M}_{22}$.

Substituting A^{-1} *back in* L *we get*

$$L = (J'_{12}M_{11}^{-1}J_{12})^{-1}\breve{J}'_{22}N^{-1}\breve{J}_{22}(J'_{12}M_{11}^{-1}J_{12})^{-1} \qquad (3.66)$$

Now we use the Matrix Inversion Lemma again to yield $N^{-1} = B$ *as was done in* (3.61) *above. Substituting this equality in the expression for* L *in* (3.66), *and this*

new expression for L back into (3.64), we get

$$\begin{aligned}
&J'_{11}M_{11}^{-1}J_{12}LJ'_{12}M_{11}^{-1}J_{11} \\
&= J'_{11}M_{11}^{-1}J_{12}(J'_{12}M_{11}^{-1}J_{12})^{-1}\breve{J}'_{22}B\breve{J}_{22}(J'_{12}M_{11}^{-1}J_{12})^{-1}J'_{12}M_{11}^{-1}J_{11} \qquad (3.67)
\end{aligned}$$

which is exactly equal to the second term of $C'BC$ in (3.63).

Finally, we show that the third terms are also equal. This implies that the fourth terms are equal since the third and fourth terms are the transpose of each other. The third term in (3.63) can be rewritten as

$$\begin{aligned}
&-\breve{J}'_{21}B\breve{J}_{22}(J'_{12}M_{11}^{-1}J_{12})^{-1}J'_{12}M_{11}^{-1}J_{11} \\
&= -\breve{J}'_{21}[\breve{M}_{22}^{-1} - \breve{M}_{22}^{-1}\breve{J}_{22}A^{-1}\breve{J}'_{22}\breve{M}_{22}^{-1}]\breve{J}_{22}(J'_{12}M_{11}^{-1}J_{12})^{-1}J'_{12}M_{11}^{-1}J_{11} \\
&= -\breve{J}'_{21}\breve{M}_{22}^{-1}\breve{J}_{22}[(J'_{12}M_{11}^{-1}J_{12})^{-1} - A^{-1}\breve{J}'_{22}\breve{M}_{22}^{-1}\breve{J}_{22}(J'_{12}M_{11}^{-1}J_{12})^{-1}]J'_{12}M_{11}^{-1}J_{11} \qquad (3.68)
\end{aligned}$$

The term in brackets in the last expression is of the form $P^{-1}-(P+Q)^{-1}QP^{-1} = (P+Q)^{-1}$ if we take $P = (J'_{12}M_{11}^{-1}J_{12})$ and $Q = \breve{J}'_{22}\breve{M}_{22}^{-1}\breve{J}_{22}$; this equality can be verified by noticing that $(P+Q)[P^{-1}-(P+Q)^{-1}QP^{-1}] = I$. With this substitution, (3.68) becomes

$$\begin{aligned}
&\breve{J}'_{21}\breve{M}_{22}^{-1}\breve{J}_{22}[(J'_{12}M_{11}^{-1}J_{12}) + \breve{J}'_{22}\breve{M}_{22}^{-1}\breve{J}_{22}]^{-1}J'_{12}M_{11}^{-1}J_{11} \\
&= \breve{J}'_{21}\breve{M}_{22}^{-1}\breve{J}_{22}A^{-1}J'_{12}M_{11}^{-1}J_{11} \qquad (3.69)
\end{aligned}$$

which is the same as the third term of $\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1}$ in (3.62).

Therefore, the equality between $\Sigma_{\theta_1,k} - \Sigma_{\theta_1,k_1}$ and $C'BC$ follows. □

Bibliography

Ahu, S. C. and P. Schmidt (1995). A separability result for gmm estimation, with applications to gls prediction and conditional moment tests. *Econometric Reviews 14*(1), 19–34.

Anselin, L. (1988). A test for spatial autocorrelation in seemingly unrelated regressions. *Economics Letters 28*(4), 335–341.

Berk, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics 2*(3), 489–502.

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*(1), 387–422.

Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics 104*(3), 535–559.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks 27*(1), 55–71.

Borovkova, S. and R. Lopuhaa (2012). Spatial garch: A spatial approach to multivariate volatility modeling. *Available at SSRN 2176781*.

Boudjellaba, H., J.-M. Dufour, and R. Roy (1992). Testing causality between two vec-

tors in multivariate autoregressive moving average models. *Journal of the American Statistical Association 87*(420), 1082–1090.

Bouissou, M. B., J.-J. Laffont, and Q. H. Vuong (1986). Tests of noncausality under Markov assumptions for qualitative panel data. *Econometrica 54*(2), 395–414.

Breusch, T., H. Qian, P. Schmidt, and D. Wyhowski (1999). Redundancy of moment conditions. *Journal of Econometrics 91*(1), 89–111.

Chamberlain, G. (1982). The general equivalence of Granger and Sims causality. *Econometrica 50*(3), 569–581.

Cheng, X. and Z. Liao (2015). Select the valid and relevant moments: An information-based LASSO for GMM with many moments. *Journal of Econometrics 186*(2), 443–464.

Cohen-Cole, E., A. Kirilenko, and E. Patacchini (2013). Strategic interactions on financial networks for the analysis of systemic risk. *Handbook on Systemic Risk 1*, 306–326.

Colombo, D. and M. H. Maathuis (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research 15*(1), 3741–3782.

Contessi, S., P. De Pace, and M. Guidolin (2014). How did the financial crisis alter the correlations of u.s. yield spreads?. *Journal of Empirical Finance 28*, 362–385.

Cooley, T. F. and S. F. Leroy (1985). Atheoretical macroeconometrics: A critique. *Journal of Monetary Economics 16*(3), 283–308.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological) 24*(2), 406–424.

Crepon, B., F. Kramarz, and A. Trognon (1997). Parameters of interest, nuisance parameters and orthogonality conditions an application to autoregressive error component models. *Journal of Econometrics 82*(1), 135–156.

Dastoor, N. K. (1981). A note on the interpretation of the Cox procedure for non-nested hypotheses. *Economics Letters 8*(2), 113–119.

Davis, R. A., P. Zang, and T. Zheng (2012). Sparse vector autoregressive modeling. *arXiv preprint arXiv:1207.0520*.

Demiralp, S. and K. D. Hoover (2003). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics 65*(s1), 745–767.

Diebold, F. X. and K. Yilmaz (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics 182*(1), 119–134.

DiTraglia, F. (2015). Using invalid instruments on purpose: Focused moment selection and averaging for GMM, second version.

Dufour, J.-M., D. Pelletier, and E. Renault (2006). Short run and long run causality in time series: inference. *Journal of Econometrics 132*(2), 337–362.

Dufour, J.-M. and E. Renault (1998). Short run and long run causality in time series: Theory. *Econometrica 66*(5), 1099–1125.

Dufour, J.-M. and A. Taamouti (2010). Short and long run causality measures: Theory and inference. *Journal of Econometrics 154*(1), 42–58.

Dufour, J.-M. and D. Tessier (1993). On the relationship between impulse response analysis, innovation accounting and Granger causality. *Economics Letters 42*(4), 327–333.

Dungey, M., G. Milunovich, S. Thorp, and M. Yang (2015). Endogenous crisis dating and contagion using smooth transition structural garch. *Journal of Banking & Finance 58*, 71–79.

Dungey, M. and E. Renault (2013). Identifying contagion. *working paper*.

Fisher, G. R. and M. McAleer (1981). Alternative procedures and associated tests of significance for non-nested hypotheses. *Journal of Econometrics 16*(1), 103–119.

Florens, J., M. Mouchart, and J. Rolin (1993). Noncausality and marginalization of Markov processes. *Econometric Theory 9*(02), 241–262.

Florens, J.-P. and M. Mouchart (1982). A note on noncausality. *Econometrica 50*(3), 583–591.

Gagliardini, P., C. Gouriéroux, and E. Renault (2011). Efficient derivative pricing by the extended method of moments. *Econometrica 79*(4), 1181–1232.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association 77*(378), 304–313.

Gottschalk, J. (2001). An introduction into the svar methodology: Identification, interpretation and limitations of svar models. *Kiel Institute of World Economics, working paper 1072*.

Gourieroux, C. and A. Monfort (1989a). A general framework for testing a null hypothesis in a 'mixed' form. *Econometric Theory 5*(1), 63–82.

Gourieroux, C. and A. Monfort (1989b). *Statistique et modèles économétriques: Notions générales, estimation, prévision, algorithmes*, Volume 1. Economica.

Gouriéroux, C. and A. Monfort (1994). Testing non-nested hypotheses. *In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics 4*(Chapter 44), 2583–2637.

Gouriéroux, C. and A. Monfort (1995). *Statistics and Econometric Models*, Volume 1. Cambridge University Press.

Gouriéroux, C., A. Monfort, and E. Renault (1987). Kullback causality measures. *Annales d'Economie et de Statistique 6/7*, 369–410.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Hearon, J. Z. (1972). Compartmental matrices with single root and nonnegative nilpotent matrices. *Mathematical Biosciences 14*(1), 135–142.

Heber, G., A. Lunde, N. Shephard, and K. Sheppard (2009). Oxford-man institute's realized library, library version: 0.2. *Oxford-Man Institute, University of Oxford*.

Hoel, P. G. (1947). On the choice of forecasting formulas. *Journal of the American Statistical Association 42*(240), 605–611.

Hoover, K. D., S. Demiralp, and S. J. Perez (2009). Empirical identification of the vector autoregression: The causes and effects of us m2. *In The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry. Oxford University Press* (2), 37–58.

Horn, R. and C. R. Johnson (1994). *Topics in matrix analysis*. Cambridge University Press.

Hsiao, C. (1979). Autoregressive modeling of canadian money and income data. *Journal of the American Statistical Association 74*(367), 553–560.

Hsiao, C. (1981). Autoregressive modelling and money-income causality detection. *Journal of Monetary Economics 7*(1), 85–106.

Jackson, M. O. (2008). *Social and economic networks*, Volume 3. Princeton University Press, Princeton.

Kalisch, M. and P. Bühlmann (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research 8*, 613–636.

Kalisch, M., M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann (2012). Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software 47*(11), 1–26.

Keating, J. W. (2000). Macroeconomic modeling with asymmetric vector autoregressions. *Journal of Macroeconomics 22*(1), 1–28.

Kilian, L. (2013). Structural vector autoregressions. *In "Handbook of Research Methods and Applications in Empirical Macroeconomics"* (22), 515–554.

Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications, Inc., Mineola, New York.

Lam, C. and P. C. Souza (2015). Estimating the spatial weight matrix using the adaptive lasso. *working paper*.

Leamer, E. E. (1978). *Specification searches: ad hoc inference with nonexperimental data*. Wiley New York.

Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. Springer-Verlag, Berlin Heidelberg.

Manresa, E. (2015). Estimating the structure of social interactions using panel data. *MIT Sloan, working paper*.

Moneta, A. (2008). Graphical causal models and vars: An empirical assessment of the real business cycles hypothesis. *Empirical Economics 35*(2), 275–300.

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*(5), 705–730.

Moody, J. and D. R. White (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review 68*(1), 103–127.

Nijman, T. and E. Sentana (1996). Marginalization and contemporaneous aggregation in multivariate GARCH processes. *Journal of Econometrics 71*(1/2), 71–87.

Noureldin, D., N. Shephard, and K. Sheppard (2014). Multivariate rotated ARCH models. *Journal of Econometrics 179*(1), 16–30.

Opsahl, T., F. Agneessens, and J. Skvoretz (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks 32 (3)*(3), 245–251.

Pearl, J. (2000). *Causality: Models, reasoning and inference*, Volume 29. Cambridge University Press.

Pesaran, M. H. (1974). On the general problem of model selection. *The Review of Economic Studies 41*(2), 153–171.

Pesaran, M. H. (1982). Comparison of local power of alternative tests of non-nested regression models. *Econometrica 50*(5), 1287–1305.

Pesaran, M. H. and M. Weeks (2001). Non-nested hypothesis testing: an overview. *A Companion to Theoretical Econometrics* B.H. Baltagi (Ed.), 279–309.

Polasek, W. (1994). *Temporal Causality Measures Based on AIC*. In: Bozdogan, H. (Ed.), Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Springer Netherlands.

Polasek, W. (2002). Bayesian causality measures for multiple ARCH models using marginal likelihoods. *In: Technical Report. Institute of Statistics and Econometrics, Univeristy of Basel, Switzerland*.

Qian, H. (2002). Partial redundancy of moment conditions. *Econometric Theory 18*(02), 531–539.

Richardson, T. (2012). Workshop on causal inference for high-dimensional data. *Atlantic Causal Conference at Johns Hopkins University*.

Robins, J. M., R. Scheines, P. Spirtes, and L. Wasserman (2003). Uniform consistency in causal inference. *Biometrika 90*(3), 491–515.

Rubio-Ramírez, J. F., D. F. Waggoner, and T. Zha (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies 77*(2), 665–696.

Saaty, T. L. and R. G. Busacker (1965). *Finite graphs and networks: An introduction with applications*. McGraw-Hill Book Company.

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters 85*(2), 461–464.

Sims, C. A. (1972). Money, income, and causality. *The American Economic Review 62*(4), 540–552.

Spirtes, P., C. N. Glymour, and R. Scheines (2000). *Causation, prediction, and search*, Volume 81. MIT press.

Stock, J. H. and M. W. Watson (2001). Vector autoregressions. *The Journal of Economic Perspectives 15*(4), 101–115.

Swanson, N. R. and C. W. J. Granger (1997). Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association 92*(437), 357–367.

Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning 65*(1), 31–78.

Uhler, C., G. Raskutti, P. Bühlmann, and B. Yu (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics 41*(2), 436–463.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference 49*(1), 137–162.

Yang, J. and Y. Zhou (2013). Credit risk spillovers among financial institutions around the global credit crisis: Firm-level evidence. *Management Science 59*(10), 2343–2359.

Zhang, J. and P. Spirtes (2002). Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pp. 632–639. Morgan Kaufmann Publishers Inc.