Three Essays on Model Selection in Time Series Econometrics: Model Averaging, Causal Graphs, and Structural Identification

INAUGURAL-DISSERTATION

zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaft
doctor rerum politicarum
(Dr. rer. pol.)

am Fachbereich Wirtschaftswissenschaft
der Freien Universität Berlin

vorgelegt von
Niels Mariano Aka

Berlin
2020

Erstgutachter:	Prof. Dr. Helmut Lütkepohl *Freie Universität Berlin und DIW Berlin*
Zweitgutachter:	Prof. Dr. Rolf Tschernig *Lehrstuhl für Ökonometrie* *Universität Regensburg*

Tag der Disputation: 9. Dezember 2020

In gratitude for the excellent companionship of my fellow graduate school members, for the continued support of family and friends, and for the commitment and inspiration provided by my teachers.

Acknowledgments

Much of this thesis has been written at the German Institute for Economic Research (DIW Berlin). Besides funding, teaching, and equipment, the DIW Graduate Center has provided a wonderfully open space for academic exchange and debate among graduate students and between students and faculty, which I am very thankful for. On the administrative side, I would like to thank the Promotionsbüro at the Freie Universität Berlin (FU Berlin) and the Graduate Center team at the DIW, especially Juliane Metzner, for making things run smoothly.

During the course of preparing this thesis, the various simulations and computations have kept quite a few processing units busy. Luckily, both the DIW Berlin and the FU Berlin have excellent and well supported computing facilities so that I could let the code run on multiple cores on DIW's *crunch* and on the Free University's *soroban* and *curta* computing clusters.

The following work has been presented, either by my self or by co-authors, at a number of conferences and seminars. These include the IMS China meeting, Statistische Woche, VfS Ausschuss für Ökonometrie, Econometrics Society European Meeting, Computational and Financial Econometrics as well as seminars at the DIW Berlin, FU Berlin, Humboldt Universtität Berlin, Soochow University, Nanjing University of Aeronautics and Astronautics, and Universität Bielefeld. I am thankful for the opportunities to present my work to the scientific community and appreciate the feedback received. I am also grateful for conference funding provided by DIW Berlin, Universität Bielefeld, and Statistische Woche.

I am deeply indebted to my two supervisors Helmut Lütkepohl and Rolf Tschernig. I thank both for their advice and support, for the generous amount of time they have made available, for the many aspects of time series analysis that I have learned from them, and for making these past years enjoyable.

Last but not least, I thank my family, and especially my partner Sandra, for supporting me throughout all this time with an open ear and a kind heart.

Berlin, August 2020 *Niels Aka*

Erklärung zu Ko-Autorenschaften

Diese Dissertation besteht aus drei (Arbeits-)Papieren, von denen eines in Zusammenarbeit mit einem Koautor entstanden ist. Der Eigenanteil an Konzeption, Durchführung und Berichtsabfassung der Kapitel lässt sich folgendermaßen zusammenfassen:

- Niels Aka und Rolf Tschernig:

  "How to Use Model Confidence Sets for Forecasting and Impulse Response Estimation and the Value of Model Averaging"

  Eigenanteil: 60 Prozent

- Niels Aka:

  "Connecting the Dots: Structural VARs and Causal Graphs"

  Eigenanteil: 100 Prozent

- Niels Aka:

  "Sign Restrictions and Causal Learning in Structural VARs: A First Case Study Using Oil Market Data"

  Eigenanteil: 100 Prozent

Contents

List of Abbreviations

ATE	average treatment effect
CIG	conditional independence graph
CPDAG	completed partially directed acyclic graph
DAG	directed acyclical graph
DGP	data generating process
FDR	false discovery rate
FWER	family-wise error rate
ICA	independent component analysis
IV	instrumental variable
MCS	model confidence set
MSEP	mean square error of prediction
OLS	ordinary least squares
PC-1	Peter and Clark algorithm, version 1
PDAG	partially directed acyclic graph
PSM	propensity score matching
RCT	randomised control trial
RDD	regression discontinuity design
RMSE	root mean square error
SEM	simultaneous equation model

SVAR	structural vector autoregressive
VAR	vector autoregressive
VECM	vector error correction model

Summary

This dissertation is concerned with model uncertainty and model selection in macroeconometric time series analysis, covering model choice in both reduced-form as well as structural models. The former deals with lag selection, whereas the latter focuses on the appropriate specification of contemporaneous interactions between variables. In both areas, quantifying and limiting the degree to which model uncertainty affects conclusions in applied work is important in safeguarding science from 'data snooping' or making bad model choices.

Chapter 1 compares and evaluates a range of model selection methods in the context of univariate autoregressive processes. Among these methods are the model confidence set, model averaging techniques, shrinkage estimators, and standard information criteria. It is found that for forecasting and impulse response analysis, incorporating model uncertainty through model averaging offers substantial reductions in mean square error when model uncertainty is high. When this uncertainty is low, standard approaches that pick a single model outperform other competitors.

Chapters 2 and 3 turn to structural time series analysis. Both chapters address the use of causal graphs in structural vector autoregressive (SVAR) analysis, which serves two purposes. Graphs succinctly summarise key modelling assumptions and, secondly, formalise the search for assumptions that are likely to be in accord with the data. Chapter 2 elaborates specific properties of these graphs in the context of SVAR models. In particular, it clarifies when a graph represents a SVAR model identified through short-run exclusion restrictions and highlights caveats when using causal graphs to learn about suitable restrictions from the data.

Chapter 3 showcases the usefulness of combining causal graphs with sign restrictions to identify SVAR models for the global crude oil market. Such a combination yields economically interpretable and meaningful results. The restrictions inspired by causal graph analysis replace a set of rather ad hoc assumptions in the literature. The results therefore add robustness to the finding that demand shocks play a more significant role for oil price movements than supply shocks.

Zusammenfassung

Diese Dissertation beschäftigt sich mit Modellunsicherheit und Modellauswahl in der makroökonomischen Zeitreihenanalyse. Die Arbeit befasst sich sowohl mit der Wahl von Modellen in reduzierter Form als auch in strukturellen Modellen. Erstere Form besteht aus der Wahl der inkludierten vergangenen Beobachtungen (engl. lags), letztere betrifft die geeignete Spezifikation der zeitgleichen Interaktionen zwischen den Modellvariablen. In beiden Bereichen ist die Quantifizierung und Limitierung der Effekte auf die Ergebnisse angewandter Arbeit, die durch Modellunsicherheit entstehen, von Bedeutung. So gilt es, irreführende Ergebnisse aufgrund von Überspezifizierung sowie die Wahl schlechter Modelle im Allgemeinein zu verhindern oder zu begrenzen.

Das erste Kapitel vergleicht und evaluiert eine Reihe an Modellselektionsverfahren im Kontext univariater autoregressiver Prozesse. Zu diesen Methoden gehören das Model Confidence Set, Modellmittelungsverfahren, sowie Shrinkage und Informationskriterien. Für die Prognose und Impulsantwortanalyse wird festgestellt, dass die Berücksichtigung von Modellunsicherheit durch Modellmittelung zur Reduzierung quadrierter Fehler beiträgt, wenn diese Unsicherheit hoch ist. Bei geringerer Unsicherheit sind Standardverfahren, die ein einzelnes Modell wählen, von Vorteil.

Kapitel 2 und 3 wenden sich der strukturellen Zeitreihenanalyse zu. Beide Kapitel beschäftigen sich mit Kausalgraphen für strukturelle vektorautoregressive (SVAR) Modelle. Kausalgraphen veranschaulichen wichtige Modellierungsannahmen und können darüber hinaus die Suche nach Annahmen formalisieren, so dass diese den Dateneigenschaften Rechnung tragen. Kapitel 2 eruiert Eigenschaften dieser Graphen im Zusammenhang mit SVAR-Modellen. Insbesondere erörtert das Kapitel inwiefern ein Graph ein identifiziertes VAR-Modell widerspiegelt. Es zeigt auch Probleme bei der datengetriebenen Suche nach Restriktionen auf.

Kapitel 3 veranschaulicht die Vorzüge einer Kombination von Kausalgraphen mit Vorzeichenrestriktionen zur Identifizierung von SVAR-Modellen. Eine solche Kombination liefert sowohl ökonomisch interpretierbare als auch aussagekräftige Ergebnisse. Die durch Kausalgraphen inspirierten Restriktionen ersetzen ad-hoc Restriktionen aus der Literatur. Mit dieser Veränderungen kann die Robustheit von Ergebnissen, die die

Bedeutsamkeit von Nachfrageschocks auf dem globalen Rohölmarkt gegenüber Angebotsschocks hervorheben, bekräftigt werden.

Introduction and Overview

For most empirical work in macroeconomics, modelling choices have to be made. But as with any choice that is based on limited information or knowledge, this choice has to be made under uncertainty. There is, therefore, always the possibility that a suboptimal model has been chosen, which could adversely effect subsequent analyses and possibly even broader conclusions about the state of the economy or its inner workings. In macroeconomics, such limited information can come about because only sampled data from a larger population is available, because of competing theories about the nature of economic phenomena, or because the relevant processes are too complex to be accurately represented by simplifying model assumptions. Acknowledging, and possibly incorporating, the uncertainty about model choice is thus an important step in scientific discovery.

This thesis contributes to the field of model selection and model uncertainty by assessing the merits of a number of statistical methods in time series analysis in offering robust model choices. Many quantities in macroeconomics are repeatedly and regularly observed over time. Incorporating the specific characteristics of such data is the domain of time series analysis. In macroeconomics, one particular tool that has proven informative and adaptable in explaining the development of multiple economic variables are vector autoregressive (VAR) models. This thesis will focus on this type of model, and its simplified version for univariate processes. An introduction to time series and VAR analysis can be found in Stock and Watson (2007, ch. 14–16). A more detailed and advanced treatment is given by Lütkepohl (2005) and Kilian and Lütkepohl (2017). General topics in time series analysis are covered by Brockwell and Davis (2002) at an introductory level and by Hamilton (1994) and Brockwell and Davis (1991) at a more advanced stage.

Vector autoregressive models are useful as a forecasting tool, but they can also provide substantive insight on economic mechanisms. In both areas, model choice is important. The thesis will evaluate the properties of model selection methods in both types of exercises: reduced-form, non-structural models which are suitable for forecasting on the one hand, and structural models capable of informing theory on the

other hand. Specifically, Chapter 1, which is joint work with Rolf Tschernig, compares the performance of a range of methods that select the lag structure of simple, non-structural univariate autoregressive models. The performance is judged with respect to forecasting, but also to impulse response estimates (that lack structural meaning in this case). Chapters 2 and 3, in turn, are concerned with model choices in structural vector autoregressive (SVAR) analysis. In both these chapters so-called causal graphs are discussed in relation to adequately forming and evaluating SVAR models.

The methods in Chapter 1 include model confidence sets, frequentist model averaging estimators, frequentist shrinkage methods, and standard information criteria. All of these methods allow to select one or more model(s) which are judged most suitable among a whole range of candidate models. The methods have been developed and applied in the econometric and statistical literature. Some, like information criteria, have existed for almost half a century. Others, such as model averaging and model confidence sets, are more recent. Their relative merits in safeguarding empirical analyses from making bad model choices are therefore less well studied in comparison. The chapter addresses this point by comparing the different approaches in a controlled environment, where artificial data is created by simulating specific time series.

A major difference between the scrutinised methods is that model confidence sets and model averaging techniques specifically allow for the possibility of many equally good models to exist and be used, whereas information criteria and shrinkage methods typically select one model for further analysis. The latter procedures disregard the possibility that other models might be equally well suited to fit the data, but with differing implications for subsequent analyses, or even that an inferior model has been chosen due to noise. Whether this disregard makes a difference in analysing certain time series is the focus of Chapter 1. As so often, the answer is 'it depends.' When the data is noisy, uncertainty is high and separating the wheat form the chaff is difficult. Under such circumstances, there are substantial advantages in hedging model choice and to allow several models to influence subsequent outcomes, such as forecasts. When the data is more informative, accounting for model uncertainty is too costly. Researchers would be better off picking a single model in that instance.

The crux in practice is how to obtain knowledge of whether the data is informative or not in the first place. Preliminary analyses using in-sample or out-of-sample goodness of fit measures may be indicative and thus lead to a first rough, heuristic assessment. In general, however, more research on the efficacy and feasibility of choosing among different weighting or selection schemes in applied work would be advantageous.

Chapters 2 and 3 turn to structural analysis and combine SVAR models with causal graphs. The latter are just like usual directed graphs, consisting of nodes and directed

edges, but infused with causal meaning. Thus, the nodes represent the same variables as in the SVAR model and the edges represent relations between cause and effect. A distinct requirement for conducting structural analysis is that we can tell certain cause-effect relations apart. That is often difficult in a macroeconomic setting, where variables are highly interrelated and endogenous. In this regard, causal graphs can be helpful by succinctly summarising and visualising the essential restrictions that have been placed on the data to identify the model at large.

A key question is under what conditions a graph represents a model that has been successfully identified. Or, putting it the other way round, what restrictions have to be placed on the associated graph such that the underlying model becomes identified. As causal graphs are a fairly uncommon tool to be used in combination with SVAR models, this question has been rarely studied in the literature. Chapter 2 addresses the issue of causal graphs and SVAR identification, clarifies other properties of causal graphs in relation to SVAR models, and provides an introduction to the topic by reviewing the current state of research.

Causal graphs have been largely developed by computer scientists in the literature on machine learning and artificial intelligence. That is why causal graphs have also been used in a different way, and in SVAR models predominantly so, to inductively infer causal relations in the data without prior subject-matter knowledge. These methods can therefore help to identify SVAR models by relating a set of statistical independence relations that were inferred from the data with causal relations. However, the statistical properties of such an approach, as with machine learning tools in general, are still being studied, improved, and evaluated. In the case of causal graphs there are important statistical caveats that researchers should be aware of when applying them for causal discovery. These caveats are also discussed in Chapter 2.

Chapter 3 highlights a new use case of causal discovery. The chapter combines causal (machine) learning with the more conventional method of identifying SVARs through sign restrictions. One weakness of identification through causal learning is that economic interpretation of the results may still be difficult as there is no information on the precise mechanisms at play. Sign restrictions, on the other hand, are usually motived by economic theory (or conventional wisdoms) and provide this intuition. On the downside, the results of sign identified SVAR analysis are frequently to loose to meaningfully inform researchers. Chapter 3 shows that a combination of both methods can provide interpretable and sharpened results. As a case study, the focus is on disentangling supply and demand in the global market for crude oil.

In summary, causal graphs are an exciting and fairly new tool that can improve communication of modelling assumptions and systematise discovery of those assumptions

which are likely supported by the data. It is not, however, a panacea to solve model uncertainty in structural VAR analysis. While it does offer a new way to relate key modelling assumptions to the data, it is accompanied by its own problems. First and foremost is the lack of an appropriate measure of uncertainty. This measure is hard to come by due to the procedure's inductive nature.

Nonetheless, the procedures outlined in this dissertation may help to formalise the often rather informal process of model choice and model discovery in macroeconometrics for both reduced-form and structural time series analysis. Consequently, the methods may guard researchers from falling into 'data snooping' (White, 2000) and improve the reproducibility of applied research, which after all is a hallmark of the scientific method and which has lately received renewed and critical attention across many disciplines.

One area that is not directly touched upon in this thesis is post-model-selection inference. Incorporating the process of model choice in evaluating the sampling properties of parameter estimators is an important yet arduous topic. Nonetheless, the impossibility results stated in Leeb and Pötscher (2005) and related studies have nurtured efforts to still conduct valid inference post-model-selection under specific circumstances and with regard to specific objects. These efforts have borne fruit for reduced-form modelling in classic low-dimensional settings (e.g. Charkhi and Claeskens, 2018) as well as for more policy-oriented targets, like treatment effects, in high-dimensional models (e.g. Chernozhukov et al., 2018). Incorporating these developments for additional robustness into applied macroeconometric research may be a worthwhile task for future research.

CHAPTER 1

How to Use Model Confidence Sets for Forecasting and Impulse Response Estimation and the Value of Model Averaging[1]

1.1 Introduction

We explore the merits of using model confidence sets (MCS) to handle model uncertainty in forecasting exercises and impulse response analysis. We do so by comparing those confidence sets to several other model selection procedures in the framework of a Monte Carlo simulation study. The study therefore sheds light on how practitioners may address the detrimental impact that model uncertainty can have on empirical findings.

Model selection, and its impact on estimation and inference, has been a long standing topic in econometric modelling (see Theil, 1957; Leamer, 1978). The reason being that the most parsimonious model containing the data generating process (DGP) is generally unknown in empirical work. This so-called true model therefore has to be selected from a collection of models that the econometrician has initially chosen. Frequently, the true model is too complex for reliable estimation in finite samples. In this case, a more parsimonious model has to be selected that approximates the DGP reasonably well. This selection automatically implies a trade-off between the approximation error and estimation uncertainty.

In applied work, the current standard approach to address this trade-off is to select a single model that minimizes a selection criterion such as the Akaike information criterion (AIC), Hannan-Quinn criterion (HQC) or Schwarz information criterion (SIC). While popular and straightforward to apply, this simple approach has its drawbacks. First, different model selection criteria may select differently. Second, by selecting a

[1] This chapter is based on joint work with Rolf Tschernig.

single model, the practitioner ignores models that are ranked close to the preferred one. Such models may be equally good or even better but were not selected due to noise. In such a case, any subsequent analysis may benefit from also considering closely ranked models. Third, classic frequentist inference conditions on a given model and will therefore suffer from size distortions if models were actually selected beforehand. This result has been established theoretically (Leeb and Pötscher, 2005) and empirically (Demetrescu, Hassler, and Kuzin, 2011).

While we do not focus on inference, we do address the first two concerns by using procedures that avoid a binary selection and instead broaden the model choice. Foremost among the procedures to relax the all-or-nothing approach of standard information criteria is the model confidence set. The MCS was suggested by Hansen, Lunde, and Nason (2011) to estimate a set of superior models from an initially given set, where superiority is defined by a user-specified loss function. Importantly, the estimation procedure developed by Hansen, Lunde, and Nason (2011) confers the property that the estimated set will (asymptotically) cover all superior models at a pre-specified significance level. A key challenge in achieving this property is to control the family-wise error rate (FWER) of a large set of hypothesis tests. Over the past two decades, methods for controlling the FWER have become available that can effectively take into account dependence among the tests by using bootstrap methods (e.g. Romano and Wolf, 2005). Hansen, Lunde, and Nason (2011) succeeded in adapting this framework to the model selection problem.

In practical work, however, it is unclear how best to continue with an empirical analysis if the estimated MCS contains more than one model. In this paper, we focus on h-step ahead forecasts and impulse responses as the relevant quantities of interest, but similar issues would arise if the focus lies on parameter estimates or other quantities. With different models describing the data equally well, we suggest averaging as one solution to reduce the plurality of models to a final answer. Averaging raises two further issues: what quantities to average and which weights to use. One may average cross model parameters or, alternatively, across (non-linear) functions of those parameters. It is not clear ex-ante which of these actions will yield better results, and we therefore study model averaging, by which we describe functions of averages of model parameters, and forecast combinations, which denote averages of functions of parameters.

Both model averaging and forecast combinations have been intensively studied in the literature on (frequentist) averaging. Claeskens and Hjort (2008) provide the first comprehensive book on frequentist model averaging and Moral-Benito (2015) provides a recent survey. Creating forecast combinations is a very active field of research by itself and relevant surveys are provided by Aiolfi, Capistrán, and Timmermann (2011)

and Timmermann (2006). A recent comprehensive treatment is found in the book by Elliot and Timmermann (2016).

It may be noted that in Bayesian econometrics, model averaging has a much longer tradition. However, in this paper our aim is to shed light on a specific set of frequentist methods, most of which have been designed with the explicit issue of model uncertainty in mind. We compare these to standard approaches that may serve as reference points for many different methods. Moreover, we evaluate the performance of methods in terms of point forecast accuracy. Focusing on point predictors instead of distributional aspects may also be less meaningful to Bayesians. For an empirical comparison of methods in the context of autoregressive processes that includes Bayesian approaches, see Kascha and Trenkler (2015).

As regards weights, we investigate the following schemes. First, we average all models in the MCS with equal weights, taking the null hypothesis of equal performance seriously. Alternatively, we estimate weights using jackknife model averaging (JMA) and apply it to the initial model set as well as to the models contained in the estimated MCS. JMA has the advantage that the weights are asymptotically optimal in various situations, as shown by Hansen and Racine (2012) and Zhang, Wan, and Zou (2013).

In the simulation, we further apply a range of competitive procedures to identify suitable autoregressive specifications within a larger pool of candidate specifications. Among them are lasso, post-lasso, and ridge regression. Taking into account all the different approaches, including the differences between model averaging and forecast combinations, we thus explore and compare the merits of 18 selection procedures.

The task in the simulation is to compute h-step ahead forecasts up to horizon 15 and impulse responses up to 20 periods ahead. As data generating processes (DGPs) we take three different univariate autoregressive processes of order eight with some of the parameters set to zero. The AR(8) processes differ in terms of their signal-to-noise ratios, frequency characteristics and degrees of persistence. We focus on univariate processes to keep the analysis as simple as possible and to focus on the relative merits of the methods that we compare in a basic setting. Building on the baseline results in this study, further investigations may explore to what extent the results change for more complex model classes. There are six sample sizes ranging from 40 to 500 observations. The initial set of models, from which suitable specifications will be selected, includes a total of 256 autoregressive processes based on all lag combinations up to lag eight. We use this particular setup to conduct a full subset specification search.

Based on mean squared error, the results suggest that using the Schwarz criterion works well for model selection in larger samples and across DGPs, but may perform poorly in small samples, in particular for impulse response estimation. In the latter

case, applying JMA to the models inside the MCS turns out to be a robust strategy. This combination is found among the best strategies in small samples and performs comparable to the best competitors besides Schwarz in larger samples. For computing impulse responses, model averaging is found to be superior to combining impulse responses of each model in the MCS.

The paper is organized as follows. Section 1.2 describes all relevant methods with a particular emphasis on model averaging and model confidence sets. The setup of the Monte Carlo simulation is laid out in Section 1.3. Section 1.4 reports the results and Section 1.5 briefly summarizes.

1.2 Methods for Model Selection and Model Averaging

In this section we briefly describe all methods used in our simulation study. We first outline the general setup of model averaging and forecast combination. Then we describe Jackknife model averaging in more detail and show how to use the MCS for model averaging. Next we sketch shrinkage methods, in particular ridge and lasso estimation. Finally we propose two new combinations of existing methods.

1.2.1 Setup

We only consider dynamic regression models for a scalar dependent variable y_t, where the number of regressors is smaller than the sample size. All approaches considered in this paper aim at estimating one or several quantities of interest such as the conditional mean, marginal effects, h-step ahead predictions or impulse response functions. They all require the user to specify an initial collection of models for further consideration. Denote this set of models by $\mathcal{M}^0$ and index all models in the set by $i = 1, \ldots, m_0$. To some extent we follow the notation of Hansen, Lunde, and Nason (2011), hereafter HLN. To facilitate the presentation, consider the estimation of the conditional mean $\mu_t \equiv E[y_t|\boldsymbol{x}_t]$ based on an observed sample $(y_t, \boldsymbol{x}_t)$, $t = 1, 2, \ldots, n$, where the $(1 \times k_{max})$ vector $\boldsymbol{x}_t$ denotes all explanatory variables available in the sample and may include lagged y_t's. We assume that $\boldsymbol{x}_t$ belongs to the information set Ω_t which includes all potential explanatory variables that are predetermined w.r.t. the error term of the data-generating process (DGP) of y_t. Note that at this point we allow for the possibility that μ_t is misspecified and that there are further relevant, possibly unobserved, variables in Ω_t. In that case, there exists a vector of explanatory variables $\boldsymbol{x}_t^+ \in \Omega_t$ such that $\mu_t = E[y_t|\boldsymbol{x}_t] \neq E[y_t|\boldsymbol{x}_t, \boldsymbol{x}_t^+]$ and μ_t exhibits an approximation error.

We have by definition

$$y_t = \mu_t + u_t, \tag{1.1}$$

$$E[u_t|\boldsymbol{x}_t] = 0. \tag{1.2}$$

The conditional variance is denoted by $\sigma_t^2 \equiv E[u_t^2|\boldsymbol{x}_t]$.

If the set of available regressors $\boldsymbol{x}_t$ contains irrelevant ones, then estimation efficiency can be increased by excluding them. The subset of regressors is one dimension in which the m_0 models in $\mathcal{M}^0$ may vary and on which we focus in this paper. In the case of linear models one may write a subset model as

$$y_t = \boldsymbol{x}_{t,i}\boldsymbol{\beta}_i + u_{t,i}, \tag{1.3}$$

where the $(1 \times k_i)$ vector $\boldsymbol{x}_{t,i}$ and the $(k_i \times 1)$ vector $\boldsymbol{\beta}_i$ denote the regressors and the parameters of model i. Note that we index the error term also by the model to explicitly indicate that its properties depend on the selected model. For example, if relevant regressors are omitted from $\boldsymbol{x}_{t,i}$, then $u_{t,i}$ contains u_t and the omitted regressor. In matrix notation we have

$$\boldsymbol{y} = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{u}_i, \quad i = 1, 2, \ldots, m_0, \tag{1.4}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{X}_i$ denote the sample vector of the dependent variable and the $(n \times k_i)$ sample matrix of explanatory observations used in model i, respectively. The $(n \times 1)$ sample error vector is denoted by $\boldsymbol{u}_i$.

The set of models $\mathcal{M}^0$ may not contain all possible combinations of regressors based on $\boldsymbol{x}_t$. We denote this complete set of possible models by $\mathcal{M}^{all}$, indexed by $s = 1, 2, \ldots, m_{all}$, where $m_{all} = 2^{k_{max}}$ denotes the model which includes all available k_{max} regressors $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n')'$. The model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{m_{all}} + \boldsymbol{u} \tag{1.5}$$

can therefore be called the encompassing model. Note that in this setup $\mathcal{M}^0 \subseteq \mathcal{M}^{all}$ holds. That also means that the encompassing model may not be included in the initial model set $\mathcal{M}^0$. In general, we assume that the elements of $\mathcal{M}^0$ are indexed with increasing complexity. The model indexed by m_0 therefore has more or equally many regressors than any other model in $\mathcal{M}^0$. For a generic set of models we simply write $\mathcal{M}$. We assume that $k_{max} < n$ holds, that $\boldsymbol{X}_i'\boldsymbol{X}_i$ is invertible for every i, and thus that all $\boldsymbol{\beta}_i$ are identifiable.

For each of the m_0 models (1.4) in the initial model set $\mathcal{M}^0$, one can produce an estimate $\hat{\mu}_{t,i}$ of the conditional expectation μ_t, e. g. by OLS, and therefore also of the observation y_t. In order to have a common notation we will use $\hat{y}_{t,i}$ in place of $\hat{\mu}_{t,i}$ and h-step ahead forecasts, if available, will be denoted by $\hat{y}_{t+h|t,i}$, $h = 1, \ldots, H$, where (sample) information up to time t is used.

Model selection corresponds to making some choice out of the m_0 available estimated models for μ_t. If the models allow for computing h-step ahead predictions, as for example in case of autoregressive models of finite order, then model selection also implies a choice from the h-step ahead predictions for y_{t+h}, $h = 1, \ldots, H$, given by the set $\{\hat{y}_{t+h|t,1}, \hat{y}_{t+h|t,2}, \ldots, \hat{y}_{t+h|t,m_0}\}$.

All methods for model selection are in one way or another based on an expected loss. For predictions, the loss caused by the deviation of $\hat{y}_{t+h|t,i}$ from the observation y_{t+h} is measured by the user-specified loss function $\hat{L}_{t,h,i} = L(y_{t+h}, \hat{y}_{t+h|t,i})$. In principle, splitting the available sample into an estimation and evaluation part $n = n_{est} + n_{evl}$ allows to obtain for each horizon h and each model i a range of values for $\hat{L}_{t,h,i}$, $t = n_{est}, \ldots, n - h$. More relevant is the expected loss or risk for model i defined as $E[\hat{L}_{t,h,i}]$, where the expectation is taken over the estimation and the evaluation sample both w.r.t. the DGP. Most frequent is the mean squared error (of prediction) (MSEP) based on quadratic loss,

$$MSEP(y_{t+h}, i) \equiv E\left[(y_{t+h} - \hat{y}_{t+h|t,i})^2\right]. \tag{1.6}$$

Optimally, one would like to choose the model(s) with the lowest risk. For an arbitrary loss function $L_{t,h,i}$ and fixed h, Hansen, Lunde, and Nason (2011) suggest to evaluate differences between models in terms of their expected loss differences $\Delta_{ij} \equiv E[L_{t,h,i} - L_{t,h,j}]$ for all $i, j \in \mathcal{M}^0$, where it is assumed that all expected loss differences are finite and independent of t such that a ranking of models is possible. Furthermore, we will suppose for the moment that the ordering of loss differences is unaffected by h to simplify notation. Note that a time-varying expected value of each loss function is still allowed for. Hansen, Lunde, and Nason (2011, Definition 1) define a set of superior models $\mathcal{M}^*$ as

$$\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \Delta_{ij} \leq 0 \text{ for all } j \in \mathcal{M}^0\}. \tag{1.7}$$

Note that $\mathcal{M}^*$ may depend on sample size. For example, when using the MSEP, the best model(s) show the optimal trade-off between squared bias and estimation variance.

In practice, the MSEPs of each model are unknown and have to be estimated. Due to the estimation error, the model exhibiting minimal MSEP may not be chosen. In

the Monte Carlo simulations in Section 1.4 where we consider h-step ahead forecasts, we compare procedures based on model selection, model averaging, forecast combinations, shrinkage estimation, and combinations thereof. Some of these approaches have optimality properties under certain conditions, and all are frequently used in empirical work to weed out bad models. It is unclear, however, which approach excels at accounting for model uncertainty and to what degree the results in forecasting and impulse response analysis are affected by competing procedures. Furthermore, it is unclear how different these approaches are when model uncertainty is either large or small and whether any approach is dominating regardless of sample properties. With our simulation, we hope to shed more light on this issue. Before we evaluate differences between methods, the remainder of this section will briefly present each of the considered procedures.

1.2.2 Model Selection

After estimating the MSEP it is common in practice to select a single model by choosing the one with the lowest estimated MSEP. The MSEP for $h = 1$ can be estimated by AIC (Akaike, 1973, 1974) or cross-validation. While not exactly estimating the MSEP, the Hannan-Quinn criterion (HQC) by Hannan and Quinn (1979) or the Schwarz information criterion (SIC or BIC) by Schwarz (1978) are equally common for choosing the model dimension. All three approaches are based on minimising a criterion that trades off model fit with model dimension. In our regression context, the criteria can be expressed as $Cr_q = \log \hat{\sigma}^2_{u,i} + c_q(n)k_i$, where $\hat{\sigma}^2_{u,i} = \frac{1}{n}\sum_{t=1}^{n} \hat{u}^2_{t,i}$ is the estimated residual variance after fitting (1.4), k_i is the dimension of model i, and the weight function $c_q(n)$ varies with $q = \text{AIC}, \text{HQC}, \text{SIC}$. With usual sample sizes, AIC penalises model dimension the least with $c_{\text{AIC}}(n) = \frac{2}{n}$; HQC increases the weight to $c_{\text{HQC}}(n) = \frac{2 \log\log n}{n}$; finally, SIC uses $c_{\text{SIC}}(n) = \frac{\log n}{n}$. In Section 1.4 we use AIC, HQC, and SIC to estimate the conditional mean

$$\hat{y}_{t,\hat{i}_q} = \boldsymbol{X}_{\hat{i}_q}\hat{\boldsymbol{\beta}}_{\hat{i}_q}, \quad q = \text{AIC}, \text{HQC}, \text{SIC}, \quad \hat{i}_q \in \mathcal{M}^0, \tag{1.8}$$

and related quantities, such as the h-step ahead forecasts $\hat{y}_{t+h|t,\hat{i}_q}$, if the models in $\mathcal{M}^0$ allow for this.

Instead of explicitly selecting a model in $\mathcal{M}^0$ one may use shrinkage estimation. Depending on the type of the regularization term, shrinkage implicitly does model selection (lasso, post-lasso) or does not (ridge). In the latter case, the largest model m_0 is always used albeit with reduced flexibility due to the shrinkage parameter. Shrinkage

methods require a proper choice of the shrinkage parameter. All shrinkage methods used in Section 1.4 are briefly presented in Section 1.2.7.

If the set of superior models $\mathcal{M}^*$ given by (1.7) contains more than one model, all procedures selecting a single model fail to estimate $\mathcal{M}^*$. The more general case is allowed for by the MCS model selection procedure described in Section 1.2.6. The MCS procedure also allows to control the size in the underlying sequence of tests.

1.2.3 Model Averaging

Selecting models always implies a discrete choice which can be avoided if all models of $\mathcal{M}^0$ are considered by properly averaging across all of them. Using continuous weights, a continuous model choice is available.

We will adopt the following notation. Let $\boldsymbol{b}_i$ denote $(k_{max} \times 1)$ parameter vectors with constant length regardless of model dimension. The vector $\boldsymbol{b}_i$ contains the entries of $\boldsymbol{\beta}_i$ at those rows where the explanatory variables in $\boldsymbol{X}_i$ correspond to the columns in $\boldsymbol{X}$. All other entries in $\boldsymbol{b}_i$ are zero. For the encompassing model (1.5) one has $\boldsymbol{b}_{m_{all}} = \boldsymbol{\beta}_{m_{all}}$. Let $\mathcal{M} = \{1, 2, \ldots, m\}$ be a generic set of models.

Model averaging computes the weighted parameter average across all models in $\mathcal{M}$

$$\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} \hat{\boldsymbol{b}}_i w_i, \tag{1.9}$$

where the index ma indicates model averaging and $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)'$ denotes the vector of weights which sum to unity, $\sum_{i=1}^{m} w_i = 1$ (Claeskens and Hjort, 2008, Section 7). In this paper we follow Hansen and Racine (2012) and impose the stronger condition of non-negative weights bounded by one, $w_i \in [0, 1]$. We denote the set of possible weight vectors by

$$\mathcal{H}_n = \{\boldsymbol{w} \in [0, 1]^m : \sum_{i=1}^{m} w_i = 1\}. \tag{1.10}$$

When applied to $\mathcal{M}^0$, the averaging estimator (1.9) can be viewed as a restricted estimator of $\boldsymbol{\beta}_{m_{all}}$ of the encompassing model of $\mathcal{M}^0$. In this sense, model averaging can be viewed as an estimator of the single model m_{all} with a particular way of regularization.

Due to the linearity in parameters of the regression setup, averaging across parameters is equivalent to averaging across estimated conditional means

$$\hat{\boldsymbol{y}}_{ma}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} \boldsymbol{X}\hat{\boldsymbol{b}}_i w_i = \boldsymbol{X}\sum_{i=1}^{m} \hat{\boldsymbol{b}}_i w_i = \boldsymbol{X}\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}). \tag{1.11}$$

The equivalence between averaging across parameters or across the quantity of interest no longer holds if the latter is a nonlinear function of the parameters. A prominent case are h-step ahead predictions and impulse responses for $h > 1$. If the models in $\mathcal{M}^0$ allow for computing h-step ahead forecasts, applying model averaging leads to computing a single h-step ahead forecast using the averaged parameter estimate $\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w})$ given by (1.9). We denote this forecast by $\hat{y}_{t+h|t,ma}(\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}))$. An alternative is discussed in the next section.

Note that the model selection procedures mentioned in Section 1.2.2 are a special form of model averaging with weight 1 assigned to that model which is selected and weight 0 to all other models. We denote the corresponding weight vectors as $\hat{\boldsymbol{w}}_{AIC}$, $\hat{\boldsymbol{w}}_{HQC}$, and $\hat{\boldsymbol{w}}_{SIC}$, etc.

1.2.4 Forecast Combinations and Combinations of Impulse Responses

An alternative to computing h-step ahead forecasts $\hat{y}_{t+h|t,ma}(\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}))$ by model averaging is to compute the h-step ahead forecasts for each of the m models in $\mathcal{M}$ using $\hat{\boldsymbol{\beta}}_i$ and then average across all m individual h-step forecasts

$$\hat{y}_{t+h|t,fc}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} w_i \hat{y}_{t+h|t,i} = \sum_{i\in\mathcal{M}} w_i \hat{y}_{t+h|t,i}. \tag{1.12}$$

The procedure (1.12) is called forecast averaging which is indicated by the index fc. This approach can also be used in more general settings where various forecasts are available but not the data underlying some of the forecasts (e.g. Aiolfi, Capistrán, and Timmermann, 2011).

Analogously to combining forecasts, one may combine estimated impulse responses $\hat{\phi}_{h,i}$ computed for each of the m models delivering

$$\hat{\phi}_{h,fc}(\boldsymbol{w}) \equiv \sum_{i\in\mathcal{M}} w_i \hat{\phi}_{h,i}. \tag{1.13}$$

One important question is whether forecast combinations are superior to forecasts that are computed with averaged parameters and similarly for impulse response estimation. We will investigate these issues in our Monte Carlo study in Section 1.4. For

both model averaging and forecast combinations it is crucial to select the weights $\boldsymbol{w}$ in some sense optimally. One approach, also applicable to time series, is presented next.

1.2.5 Jackknife Model Averaging

Initially, as Hansen and Racine (2012) mention, Wolpert (1992) and Breiman (1996) introduced the idea of jackknife model averaging (JMA), which uses leave-one-out cross-validation to choose the weights $\boldsymbol{w}$. Hastie, Tibshirani, and Friedman (2009, Section 8.8) call this procedure stacking. However, only recently Hansen and Racine (2012) showed the asymptotic optimality of JMA "in the sense of achieving the lowest possible expected squared error over the class of linear estimators constructed from a countable set of weights" (Hansen and Racine, 2012, p. 36). Their procedure requires independent observations, but in contrast to alternative procedures allows for "bounded heteroscedasticity of unknown form" and an unbounded number of models. Zhang, Wan, and Zou (2013) showed the asymptotic optimality for a wider class of data generating processes including stochastic processes. It is for these reasons that we have chosen JMA for representing model averaging. As a side remark, Zhang, Wan, and Zou (2013) do no longer require the weights to be taken from a discrete grid of points.

Next, we briefly describe the algorithm of JMA. Hansen and Racine (2012), hereafter HR, consider linear estimators for which $\hat{\boldsymbol{\mu}}_i = \boldsymbol{P}_i\boldsymbol{y}$ holds and where the $n \times n$ matrix $\boldsymbol{P}_i$ does not depend on $\boldsymbol{y}$. For least-squares estimation $\boldsymbol{P}_i = \boldsymbol{X}_i \left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1} \boldsymbol{X}_i'$ is a projection matrix. Using this notation, the model averaging estimator for the conditional mean given by (1.11) for a given weight vector $\boldsymbol{w}$ is

$$\hat{\boldsymbol{y}}_{ma}(\boldsymbol{w}) = \sum_{i=1}^{m} w_i \boldsymbol{P}_i \boldsymbol{y} = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}, \quad \boldsymbol{P}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} w_i \boldsymbol{P}_i. \tag{1.14}$$

In order to estimate the weight vector $\boldsymbol{w}$, HR use jackknife estimation by applying leave-one-out cross-validation or n-fold cross-validation (Hastie, Tibshirani, and Friedman, 2009, Section 7.10.1). This estimator, denoted by $\widetilde{y}_{t,i}$, estimates the conditional mean μ_t without using the observation $(y_t, \boldsymbol{x}_t)$. As noted by HR, and others, the projection matrix $\widetilde{\boldsymbol{P}}_i$ corresponding to leave-one-out cross-validation is identical to $\boldsymbol{P}_i$ except for zeros on the diagonal. For a given weight vector $\boldsymbol{w}$, the Jackknife estimator is then given by

$$\widetilde{\boldsymbol{y}}_{ma}(\boldsymbol{w}) = \sum_{i=1}^{m} w_i \widetilde{\boldsymbol{P}}_i \boldsymbol{y} = \widetilde{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{y}, \quad \widetilde{\boldsymbol{P}}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} w_i \widetilde{\boldsymbol{P}}_i. \tag{1.15}$$

In order to determine $\boldsymbol{w}$, HR estimate the MSEP by

$$CV_n(\boldsymbol{w}) = ||\boldsymbol{y} - \widetilde{\boldsymbol{y}}_{ma}(\boldsymbol{w})||^2/n \tag{1.16}$$

and minimize it w.r.t. the weight vector $\boldsymbol{w}$

$$\widehat{\boldsymbol{w}}_{jma} = \underset{\boldsymbol{w}\in\mathcal{H}_n}{\arg\min}\ CV_n(\boldsymbol{w}), \tag{1.17}$$

where the set of possible weight vectors $\mathcal{H}_n$ is given by (1.10). Note that (1.16) is a quadratic function in $\boldsymbol{w}$ since $CV_n(\boldsymbol{w}) = \boldsymbol{w}^T\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}}\boldsymbol{w}/n$ with $\widetilde{\boldsymbol{U}} = (\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_m)$ and the jackknife residuals $\widetilde{\boldsymbol{u}}_i = \boldsymbol{y} - \widetilde{\boldsymbol{P}}_i\boldsymbol{y}$. Due to the inequality constraints on $\mathcal{H}_n$ this is a quadratic programming problem, which in the R language may be solved by the quadprog package (see Turlach, Weingessel, and Moler, 2019).

Zhang, Wan, and Zou (2013, Section 3) derive conditions for dependent processes that guarantee the optimality of the JMA procedure where optimality is defined as

$$\frac{L_n(\widehat{\boldsymbol{w}}_{jma})}{\inf_{\boldsymbol{w}\in\mathcal{H}_n} L_n(\boldsymbol{w})} \xrightarrow{p} 1, \tag{1.18}$$

and the quadratic loss is given by $L_n(\boldsymbol{w}) = (\boldsymbol{\mu} - \widehat{\boldsymbol{y}}_{ma}(\boldsymbol{w}))^T(\boldsymbol{\mu} - \widehat{\boldsymbol{y}}_{ma}(\boldsymbol{w}))$. These conditions include stationary homoskedastic finite-order AR processes.

Due to the quadratic nature of the loss function, no single weight will be estimated as exactly zero and therefore jackknife model averaging is not designed for model selection. This is in contrast to the approach of the next subsection.

1.2.6 MCS-Based Model Selection

The original aim of the MCS procedure is to estimate the set of superior models $\mathcal{M}^*$ given by (1.7) and thus to eliminate all inferior models from $\mathcal{M}^0$. In contrast to standard model selection procedures, it additionally allows to asymptotically control the family-wise error rate in the sequence of tests constituting the MCS procedure, and thus to estimate the set $\mathcal{M}^*$ with a certain level of confidence, at least asymptotically. With large enough samples and under repeated sampling, the MCS procedure allows the conclusion that all superior models are part of the estimated set $\widehat{\mathcal{M}}^*$ in at least $(1-\alpha)\times 100\,\%$ of cases, where α was fixed in advance.

To eliminate inferior objects from the set $\mathcal{M}^0$, a sequential testing procedure is used based on (1.7) with the following pair of hypotheses[2]

$$H_{0,\mathcal{M}} : \Delta_{ij} \leq 0 \text{ for all } i,j \in \mathcal{M}, \mathcal{M} \subseteq \mathcal{M}^0, \tag{1.19}$$

$$H_{A,\mathcal{M}} : \Delta_{ij} > 0 \text{ for some } i,j \in \mathcal{M}, \mathcal{M} \subseteq \mathcal{M}^0. \tag{1.20}$$

The hypotheses in (1.19) and (1.20) are indexed by $\mathcal{M}$ which emphasises the fact that both hypotheses refer to some generic set $\mathcal{M}$ that is a subset of $\mathcal{M}^0$ and may have been obtained by previous applications of the MCS testing procedure. The hypothesis $H_{0,\mathcal{M}}$ is true when $\mathcal{M} = \mathcal{M}^*$. The alternative is true when inferior models, as measured by expected loss differences, are still members of $\mathcal{M}$. These two hypotheses can be used to find an estimate of $\mathcal{M}^*$ in a sequential manner: if $H_{0,\mathcal{M}}$ can be rejected, then remove a model from $\mathcal{M}$ and apply the hypothesis test again to the remaining models until the null hypothesis can no longer be rejected. The remaining set of models estimates $\mathcal{M}^*$ and is called a model confidence set (MCS) and denoted by $\widehat{\mathcal{M}}^*_{1-\alpha}$.

The algorithm to obtain the MCS is a sequential testing procedure in which two alternating tests are carried out. Let $\delta_{\mathcal{M}}$ be a binary variable that is associated with a suitable test for the null hypothesis (1.19) and which equals 1 if $H_{0,\mathcal{M}}$ is rejected and 0 if it is not rejected. Further, let $e_{\mathcal{M}}$ be the model that is removed if $\delta_{\mathcal{M}} = 1$. Hansen, Lunde, and Nason (2011) call $\delta_{\mathcal{M}}$ "equivalence test" and $e_{\mathcal{M}}$ "elimination rule". Equipped with these tools the model confidence set (MCS) procedure can be stated as follows.

Algorithm 1.2.1. ***(MCS procedure)***

0. Start with $\mathcal{M} = \mathcal{M}^0$.

1. Test $H_{0,\mathcal{M}}$ using $\delta_{\mathcal{M}}$ at level α.

2. If $\delta_{\mathcal{M}} = 0$, set $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$ and stop.
If $\delta_{\mathcal{M}} = 1$, use $e_{\mathcal{M}}$ to remove a model and repeat from step 1.

For establishing that $\widehat{\mathcal{M}}^*_{1-\alpha}$ has an asymptotic coverage probability of $1-\alpha$, Hansen, Lunde, and Nason (2011, Assumption 1) state the following requirements. The equivalence test and elimination rule must be 'well behaved' in the sense that, asymptotically, (a) $H_{0,\mathcal{M}}$ is only rejected with probability less than or equal to α when it is true, (b) $H_{0,\mathcal{M}}$ is rejected with probability converging to one when it is false and (c) the probability of eliminating a superior model when $H_{0,\mathcal{M}}$ is false converges to zero. Assumptions

[2]Hansen, Lunde, and Nason (2011) specify the hypothesis slightly differently with $\Delta_{ij} = 0$ versus $\Delta_{ij} \neq 0$. If the differences are symmetric, this amounts to the same since $\Delta_{ij} \leq 0$ for all $i,j \in \mathcal{M}$ implies $\Delta_{ij} = 0$.

(a) and (b) are relatively standard for conventional statistical hypothesis tests. Assumption (c) needs to be confirmed for any elimination rule that will be considered. Theorem 1 in Hansen, Lunde, and Nason (2011) shows that the coverage property of the MCS algorithm is asymptotically guaranteed as well as that the probability of the selected set to contain any inferior model approaches zero asymptotically. In other words, $\widehat{\mathcal{M}}^*_{1-\alpha}$ asymptotically includes all superior but no inferior models at the given confidence level.

Note that in this setting there is no allowance for the type I error to accumulate because asymptotically the sequential testing procedure ensures that the "first time a superior model is questioned by the elimination rule is when the equivalence test is applied to $\mathcal{M}^*$" (Hansen, Lunde, and Nason, 2011, p. 460). Thus, the family-wise error rate (FWER) is asymptotically controlled at α. In the special case when $\mathcal{M}^*$ contains only a single model, Corollary 1 in Hansen, Lunde, and Nason (2011) states that the probability that $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}^*$ converges to one.

Since all stated assumptions concern asymptotic behavior, the MCS procedure may well be oversized in finite samples. HLN devise a formal concept (HLN, Definition 3) which they call 'coherency' between the equivalence test and elimination rule. It requires that as long as there are inferior models in the set, the probability of removing a superior model must not be larger than in the case when there is no inferior model in the set. In practice the assumption restricts the space of possible $\delta_{\mathcal{M}}$, $e_{\mathcal{M}}$ combinations to those where a rejection implies enough evidence that a specific model is inferior and can be eliminated. While the coherency requirement cannot assure that the family-wise error rate is controlled at α in finite samples, it contains the probability of removing superior models to an acceptable degree.

Next we state the specific equivalence and elimination rules used in Section 1.4. They are based on estimating the expected difference in losses $\Delta_{ij} = E[d_{t,ij}]$ with $d_{t,ij} \equiv L_{t,i} - L_{t,j}$ underlying the set of superior models (1.7). Hansen, Lunde, and Nason (2011, Assumption 2) assume that the loss differences $d_{t,ij}$ are strictly stationary, α-mixing and fulfill some moment condition. This assumption will be met by our DGPs in our Monte Carlo study.

To estimate the expected loss Δ_{ij}, the original sample $(y_t, \boldsymbol{x}_t)$, $t = 1, 2, \ldots, n$, is split into an estimation sample, $t = 1, 2, \ldots, n_{est}$, and an evaluation sample, $t = n_{est} + 1, \ldots, n$. The former is used to obtain $\hat{\boldsymbol{\beta}}_i$ and the latter allows to estimate Δ_{ij} by the relative sample loss statistic $\bar{d}_{ij} \equiv \bar{L}_i - \bar{L}_j$ with $\bar{L}_i = (n - n_{est})^{-1} \sum_{t=n_{est}+1}^{n} L_{t,i}$ where we define $L_{t,i}$ by the quadratic loss of the one-step ahead prediction error $L_{t,i} = (y_t - \hat{y}_{t|t-1,i})^2$. In our study we apply the T-max and the T-min statistic as two alternatives for the equivalence test. Both are based on the relative sample loss statistic

$\bar{d}_{i\cdot} \equiv \bar{L}_i - \bar{L}_\cdot$ with $\bar{L}_\cdot \equiv m^{-1} \sum_{i \in \mathcal{M}} \bar{L}_i$ and the following t-statistic

$$t_{i\cdot} \equiv \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{var}(\bar{d}_{i\cdot})}}, \tag{1.21}$$

and are given by

$$T_{\max,\mathcal{M}} \equiv \max_{i \in \mathcal{M}} t_{i\cdot}, \tag{1.22}$$

$$T_{\min,\mathcal{M}} \equiv \min_{i \in \mathcal{M}} t_{i\cdot}. \tag{1.23}$$

The T-max statistic (1.22) fulfills the coherency rule and is recommended by Hansen, Lunde, and Nason (2011) for empirical work on the basis of their simulation results. However, by redoing their simulations, Aka (2014) found that their simulation results were actually based on the T-min statistic (1.23). While the latter has good power, it violates the coherency condition stated above.[3] For the former reason we include it in our simulation setup.

The corresponding elimination rules are given by

$$e_{\mathcal{M},T_{max}} = \arg\max_{i \in \mathcal{M}} t_{i\cdot}, \tag{1.24}$$

$$e_{\mathcal{M},T_{min}} = \arg\min_{i \in \mathcal{M}} t_{i\cdot} \tag{1.25}$$

Both test statistics exhibit nonstandard distributions under the null hypothesis $H_{0,\mathcal{M}}$ which HLN approximate using a circular block bootstrap which also allows to compute $\widehat{Var}(\bar{d}_{i\cdot})$. Details are given in the supplement of Hansen, Lunde, and Nason (2011). This completes the MCS procedure. In sum, the MCS procedure requires to choose a significance level α, an equivalence and elimination rule, the ratio $r = n_{est}/n$ of dividing the sample into the estimation and evaluation sample, the number of bootstrap replications and the block size for the bootstrap.

Once the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$ is estimated and contains more than one model, which is typically the case, one has to decide how to proceed. Since according to the null hypothesis $H_{0,\mathcal{M}}$ all models exhibit the identical lowest risk, one can argue to use all models contained in $\widehat{\mathcal{M}}^*_{1-\alpha}$ in an identical way. This suggests to do model averaging and compute $\hat{b}_{ma}(\widehat{\boldsymbol{w}}_{T_{mcs}})$, $T_{mcs} \in \{T_{max}, T_{min}\}$, using a weight vector $\widehat{\boldsymbol{w}}_{T_{mcs}}$ that assigns

[3] We thank Peter Hansen and Asger Lunde for providing us with the source code of their Ox package. When redoing the simulations with the maximum statistic, the power turned out to be worse than for the maximum range statistic $T_{R,\mathcal{M}}$ which was also suggested by Hansen, Lunde, and Nason (2011) but requires $m(m-1)/2$ instead of $m-1$ comparisons and is even more computationally intensive. Details can be found in Aka (2014, Section 4.1).

equal weights to models in $\widehat{\mathcal{M}}^*_{1-\alpha}$ and zero weights to all other models:

$$\widehat{w}_{T_{mcs},i} \equiv \begin{cases} |\widehat{\mathcal{M}}^*_{1-\alpha}|^{-1} & \text{if } i \in \widehat{\mathcal{M}}^*_{1-\alpha}, \\ 0 & \text{otherwise,} \end{cases} \tag{1.26}$$

$i = 1, 2, \ldots, m_0$, where $|\cdot|$ denotes the number of elements.

1.2.7 Shrinkage Methods

Shrinkage methods can also be called penalized estimation or estimation with a regularization term. The regularization limits the flexibility of the parameters and therefore allows to estimate models with a large number of parameters such as, for example, the largest models in $\mathcal{M}^0$ or $\mathcal{M}_{all}$. Specific shrinkage methods differ w.r.t. the basic estimator and the regularization term. Often, the degree of regularization is controlled by the regularization parameter λ which has to be estimated. In the following we consider regularization of the OLS estimator.

The ridge estimator $\hat{\boldsymbol{\beta}}_{i,ridge}$ is obtained by summing over squared parameter values (except for the constant possibly) and is available in matrix form as

$$\hat{\boldsymbol{\beta}}_{i,ridge}(\lambda) = \arg\min_{\boldsymbol{\beta}_i} \left(||\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\beta}_i||^2 + \lambda \sum_{j=1}^{k_i} \beta_{i,j}^2 \right) \tag{1.27}$$

$$= \left(\boldsymbol{X}_i'\boldsymbol{X}_i + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}_i'\boldsymbol{y}. \tag{1.28}$$

Here, $||\cdot||$ denotes the Euclidean norm. Increasing the regularization parameters λ implies a more restrictive estimator which may lead to larger bias and smaller variance. Since the estimated $\hat{\beta}_{m_0,j}$ are different from zero with probability one, independently of the value of the regularization parameter λ, no model selection is conducted and all parameters of model i are estimated. Therefore the ridge estimator may be viewed as a constrained estimator of the encompassing model (1.5) if $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{m_{all}}$.

The lasso estimator, in contrast, is defined by summing over absolute values of the parameter values

$$\hat{\boldsymbol{\beta}}_{i,lasso}(\lambda) = \arg\min_{\boldsymbol{\beta}_i} \left(\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\beta}_i||^2 + \lambda \sum_{j=1}^{k_i} |\beta_{i,j}| \right). \tag{1.29}$$

Using absolute values allows for the possibility that some elements of $\hat{\boldsymbol{\beta}}_{i,lasso}$ can be estimated to be exactly zero which implies model selection. How many and which parameters are set to zero depends on the regularization parameter λ, among other

things. The larger λ, the more zeros may occur. Note that the selected model may not be included in $\mathcal{M}^0$ except if $\mathcal{M}^0$ coincides with the complete set of possible models $\mathcal{M}^{all} = \{1, 2, \ldots, m_{all}\}$, see Section 1.2.1. Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_l, \ldots, \lambda_g)$ denote a $(g \times 1)$ vector of increasing λ-values. Then we denote the model selected by λ_l by $\hat{i}_{\lambda_l} \in \mathcal{M}^{all}$ and the corresponding lasso estimator by $\hat{\boldsymbol{\beta}}_{\hat{i},lasso}(\lambda_l)$ or $\hat{\boldsymbol{b}}_{\hat{i}_{\lambda_l},lasso}$. It may well happen that different λ_l select the same model. Thus the number of different models selected $\hat{i}_{\lambda_l}$, $l = 1, 2, \ldots, g$ may be smaller than g. We denote the set of different models implied by the vector of regularization vectors $\boldsymbol{\lambda}$ by

$$\widehat{\mathcal{M}}_{lasso}(\boldsymbol{\lambda}) \equiv \{i \in \mathcal{M}^{all} : i = \hat{i}_{\lambda_l}, l = 1, 2 \ldots, g\}. \tag{1.30}$$

In order to select λ from $\boldsymbol{\lambda}$ we follow Hansen (2016, Section 4.1) and use 5-fold cross-validation with the R package glmnet (Friedman, Hastie, and Tibshirani, 2010). See also Hastie, Tibshirani, and Friedman (2009, Section 7.10) for an introduction to K-fold cross-validation. We therefore obtain the lasso estimate exhibiting lowest estimated risk given $\boldsymbol{\lambda}$ by $\hat{\boldsymbol{\beta}}_{\hat{i}_{\hat{\lambda}},lasso}(\hat{\lambda})$ or $\hat{\boldsymbol{b}}_{\hat{i}_{\hat{\lambda}},lasso}$. The same procedure for estimating λ is used for the ridge estimator (1.27) which delivers $\hat{\boldsymbol{\beta}}_{i,ridge}(\hat{\lambda})$. If a constant is included, we will not shrink its coefficient using either lasso or ridge. In that case, (1.27) and (1.29) have to be appropriately modified, e.g. by adjusting the indexation.

We further follow Hansen (2016) and include the post-lasso estimator of Belloni and Chernozhukov (2013). This estimator simply re-estimates $\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},lasso}(\hat{\lambda})$ for the same model as identified by lasso but using OLS, and thus with less or no bias. We denote this parameter estimate by $\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},postlasso}$. Note that all estimators of this section estimate a single model but differ in the effect of the regularization.

1.2.8 Two New Suggestions

In this section we propose to apply jackknife model averaging described in Section 1.2.5 to specific subsets of either the initial model set $\mathcal{M}^0$ or the complete model set $\mathcal{M}^{all}$ in order to combine the advantages of jackknife model averaging with the methods delivering the subsets.

First, we suggest to apply JMA to model confidence sets described in Section 1.2.6. There, we proposed to use equal weights to all models in $\widehat{\mathcal{M}}^*_{1-\alpha}$ in order to reflect the idea of the underlying null hypothesis. In light of the property of the estimator of $\mathcal{M}^*$ to possibly include inferior models when there is insufficient information in the data, using equal weights may give inferior models too large a weight. As JMA may perform quite well also if inferior models are in the model set $\mathcal{M}^0$, we propose to apply it to all

models contained in $\widehat{\mathcal{M}}^*_{1-\alpha}$, which delivers the MCS-based model averaging estimator

$$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{jma,T_{mcs}}) \equiv \sum_{i \in \widehat{\mathcal{M}}^*_{1-\alpha}} \hat{\boldsymbol{b}}_i \widehat{w}_{i,jma,T_{mcs}}, \quad T_{mcs} \in \{T_{max}, T_{min}\} \tag{1.31}$$

and MCS-based forecast combinations

$$\hat{y}_{t+h|t,fc}(\widehat{\boldsymbol{w}}_{jma,T_{mcs}}) \equiv \sum_{i \in \widehat{\mathcal{M}}^*_{1-\alpha}} \hat{y}_{t+h|t,i} \widehat{w}_{i,jma,T_{mcs}}, \quad T_{mcs} \in \{T_{max}, T_{min}\}. \tag{1.32}$$

The same idea can be applied to the set of models $\widehat{\mathcal{M}}^0_{lasso}$ implied by a vector $\boldsymbol{\lambda}$ of regularization parameters when using lasso estimation. This estimated set defined by (1.30) typically contains models with differing bias-variance trade-offs. So instead of picking a single model using K-fold cross-validation as in Section 1.2.7, one can apply JMA to the estimated initial set $\widehat{\mathcal{M}}^0_{lasso}$ from which one obtains the lasso-based model averaging estimator

$$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{jma,lasso}) \equiv \sum_{i \in \widehat{\mathcal{M}}^0_{lasso}} \hat{\boldsymbol{b}}_i \hat{w}_{i,jma,lasso} \tag{1.33}$$

and lasso-based forecast combinations

$$\hat{y}_{t+h|t,fc}(\widehat{\boldsymbol{w}}_{jma,lasso}) \equiv \sum_{i \in \widehat{\mathcal{M}}^0_{lasso}} \hat{y}_{t+h|t,i} \hat{w}_{i,jma,lasso}, \tag{1.34}$$

where the indexation of all models is given by $i \in \mathcal{M}^{all}$ and therefore $\widehat{\mathcal{M}}^0_{lasso} \in \mathcal{M}^{all}$. Considering all these methods and the possible differences between model averaging and forecast combinations, Section 1.2 described and suggested 18 different possibilities to compute h-step ahead predictions and impulse responses. Their performance will be compared in Section 1.4 based on the setup described in the following section.

1.3 Design of Monte Carlo Simulation

1.3.1 Data Generation

This section describes the way we generate the artificial data which is used in the Monte Carlo simulation. For simplicity, we focus on univariate linear autoregressive

processes as the data generating process (DGP):

$$\alpha_0(L)y_t = \nu_0 + u_t, \quad u_t|y_{t-1}, y_{t-2}, \ldots \sim \mathrm{N}(0, \sigma_0^2), \quad t = 1, 2, \ldots, n, \tag{1.35}$$

$$\alpha_0(L) = \alpha_{j_1,0}L^{j_1} + \alpha_{j_2,0}L^{j_2} + \cdots \alpha_{p_0,0}L^{p_0}, \tag{1.36}$$

where $\alpha_0(L)$ denotes the p_0-order lag polynomial with the specific set of lags of the DGP. All models which are considered for fitting the data are finite-order autoregressive models with different lag polynomials $\alpha_i(L) = \alpha_{i_1,i}L^{i_1} + \alpha_{i_2,i}L^{i_2} + \cdots \alpha_{p_i,i}L^{p_i}$ and a constant ν_i.

One objective of our simulation exercise is to gauge the ability of the various methods to identify not just a maximum lag order p, but to recover a DGP with a strict subset of lags up to order p. To this end, we proceed as follows. We pick DGPs from the set of processes that have non-zero coefficients for lags one, six, and seven, and zero coefficients for all others. To obtain DGPs that may be comparable to those found in applied work, we take into account three properties of any given process: the signal-to-noise ratio, the roots of the autoregressive polynomial $\alpha(L)$, and the frequency properties of the process. Denote the variance of y_t as σ_y^2 and the variance of the error term as σ_u^2. The precision (i.e. the inverse of the variance) of the OLS estimator is increasing in σ_y^2/σ_u^2, which we call the signal-to-noise ratio (SNR). More customary is the usual goodness-of-fit measure $R^2 = 1 - \sigma_u^2/\sigma_y^2$, which we fix in our simulations to control the finite sample estimation precision. We constrain the roots of the autoregressive polynomial to be greater than some value $\eta > 1$ in absolute value to ensure stationarity. We further analyse the spectral density of any candidate process and disregard those that derive the majority of their variance from extremely high or low frequencies.

For finding specific processes that fulfill these properties, we use a constrained optimisation algorithm. The objective function is given by the absolute deviation between current R^2 and target R^2, while the roots provide an inequality constraint. Since this problem is over-determined, we further filter out processes that do not possess the desired frequency properties. From the remaining processes we pick one at random.

In the simulation, we let R^2 vary between $0.2, 0.5$ and 0.8 and set $\eta = 1.1$, $\sigma_u^2 = 1$ and $\nu_0 = 0$ throughout. We thereby obtain processes as shown in Table 1.1. Each row presents one autoregressive process. The columns indicate the corresponding R^2, the length of the smallest root of the autoregressive lag polynomial and the values of the autoregressive coefficients at lags one, six and seven. The last three columns show the proportion of the variance attributable to frequencies below or equal to $\frac{1}{4}\pi$, $\frac{2}{4}\pi$ and $\frac{3}{4}\pi$. As the table shows, the inequality constraint is never binding as all three processes

have their smallest root fairly close to, yet above, 1.1. The spectral densities indicate that as the R^2 increases there is also a shift in weight to lower frequencies.

These processes are decidedly simple to keep the focus on the effects of model uncertainty inherent even in the most basic settings. Further investigations could explore to what extent the relative merits of the compared methods change when more complex, and more realistic, features are allowed for, including seasonality, trends, regime shifts, or higher dimensional systems. Preliminary experiments by the authors using GARCH processes suggested, though, that including conditionally heteroskedastic AR processes is unlikely to change the conclusions drawn from the current setting. Since we focus on point predictions, changes in residual variance will have little effect on the relative merits of the considered methods.

Table 1.1: Characteristics of the three chosen data-generating processes.

R^2	Smallest Root	Coefficients			Spectral Density		
		α_1	α_6	α_7	$\frac{1}{4}\pi$	$\frac{2}{4}\pi$	$\frac{3}{4}\pi$
0.2	1.105	0.214	-0.265	0.380	0.33	0.55	0.78
0.5	1.138	0.671	-0.422	0.342	0.73	0.85	0.93
0.8	1.131	0.855	0.338	-0.310	0.89	0.96	0.99

Notes: R^2 denotes the probability limit of R^2 for the corresponding correct model. The last three columns show the proportion of the variance attributable to frequencies below or equal to $\frac{1}{4}\pi$, $\frac{2}{4}\pi$ and $\frac{3}{4}\pi$. Values are rounded to third or second decimal place.

1.3.2 Choice of Initial Model Set and Auxiliary Parameters

The encompassing model of the initial collection of models $\mathcal{M}^0$ is an AR(8) model with a constant. The initial model set $\mathcal{M}^0$ contains all subset AR models obtained by considering all possible combinations of zero restrictions on the lagged variables. This provides $m_0 = m_{all} = 2^8 = 256$ models in $\mathcal{M}^0$ (a constant is always included). We consider six small to medium-sized samples with $n = 40, 60, 80, 100, 250, 500$ each with 50 burn-in observations. While sample sizes below 100 are smaller than typical macroeconomic time series, a more relevant measure is the ratio of observations to parameter estimates. Since the largest model has eight slope parameters and a constant, the lower bound of this ratio ranges from about 4.5 to 55.6 for sample sizes between 40 and 500. A ratio of five observations per parameter is representative of many real-world VAR applications and smaller sample sizes of 40 to 100 observations may therefore be more informative about the effect of model uncertainty in medium or large-scale models. For

each model, h-step ahead forecasts are computed iteratively for $h = 1, 2, \ldots, 15$ and the impulse response coefficients for $h = 1, 2, \ldots, 20$. The number of replications is set to $R = 5000$.

Further auxiliary parameters are set as follows. For the MCS algorithm, the significance level is set to $\alpha = 0.2$. This may seem rather high but ensures decent power to eliminate inferior models in smaller samples. The default block length in the circular block bootstrap is set to $l = 20$ and then estimated by the function b.star in the R package np (Hayfield and Racine, 2008). The number of bootstrap replications is set to $B = 1000$. The length of the estimation sample is set to $n_{est} = \lceil n^{4/5} \rceil$, such that $\lfloor n - n^{4/5} \rfloor$ observations remain for evaluating the relative loss differences between models. The two functions $\lceil\rceil$ and $\lfloor\rfloor$ round to the next larger and smaller integer respectively. This formula reflects preliminary investigations by the authors that with increasing sample size a larger fraction of the data should be used for estimating out-of-sample losses. Finally, both the T-max (1.22) and the T-min statistic (1.23) are used for testing equivalence. For the MCS, the equal weights as well as the JMA based weights are used for model averaging and forecast combinations.

The choice of the grid of regularization parameters follows the default behavior of the glmnet package and is as follows. The grid spans 100 values between $\lambda = 0.001$ and the smallest λ that will lead to all coefficients being set to zero.[4] The 100 values are then evenly distributed on a log scale between these two extremes. The choice of λ from this grid is described in Section 1.2.7.

1.3.3 Summary Statistics

For evaluating and comparing all 18 methods we use the following summary measures. They are all based on estimating the MSEP for h-step ahead predictions of y_{n+h} and the root mean square error (RMSE) for impulse response functions ϕ_h by averaging across all $r = 1, 2, \ldots, R$ simulation runs

$$\widehat{MSEP}(y_{n+h}, h, s) = \sum_{r=1}^{R} \left(\hat{y}_{n+h|n,r,s} - y_{n+h,r}\right)^2, \tag{1.37}$$

$$\widehat{RMSE}(\phi_h, h, s) = \sqrt{\sum_{r=1}^{R} \left(\hat{\phi}_{h,r,s} - \phi_h\right)^2}, \tag{1.38}$$

where s denotes one of the methods listed in the first column of Table 1.2. To succinctly summarize the results, we average across all H horizons. To avoid scale effects, we

[4] For ridge the smallest λ would be infinity. A suitable proxy is therefore chosen. See the package documentation of Friedman, Hastie, and Tibshirani (2010) for details.

average over relative MSEPs by relating the MSEP for method s and horizon h to the corresponding MSEP when the DGP is known

$$\widehat{RelMSEP}(y_{t+h}, h, s) \equiv \frac{\widehat{MSEP}(y_{t+h}, h, s) - \widehat{MSEP}(y_{t+h}, h, DGP)}{\widehat{MSEP}(y_{t+h}, h, DGP)}, \tag{1.39}$$

$$\widehat{AvRelMSEP}(y_{t+h}, H, s) \equiv H^{-1} \sum_{h=1}^{H} \widehat{RelMSEP}(y_{t+h}, h, s). \tag{1.40}$$

Since there is no uncertainty about the impulse responses of the DGP, we use the RMSE of estimating the impulse response values based on the correct model i_{DGP}, which only includes the lags of the DGP, to obtain relative quantities

$$\widehat{RelRMSE}(\phi_h, h, s) \equiv \frac{\widehat{RMSE}(\phi_h, h, s) - \widehat{RMSE}(\phi_h, h, i_{DGP})}{\widehat{RMSE}(\phi_h, h, i_{DGP})}, \tag{1.41}$$

$$\widehat{AverRelRMSE}(\phi_h, H, s) \equiv H^{-1} \sum_{h=1}^{H} \widehat{RelRMSE}(\phi_h, h, s). \tag{1.42}$$

Table 1.2: Description of all methods used in simulation study.

Group	Method	Model Averaging	Forecast Combination	Descriptions
1	AIC	$\hat{\boldsymbol{\beta}}_{\hat{i}_{AIC}}$		Section 1.2.2
1	HQ	$\hat{\boldsymbol{\beta}}_{\hat{i}_{HQ}}$		Section 1.2.2
1	SIC/BIC	$\hat{\boldsymbol{\beta}}_{\hat{i}_{SC}}$		Section 1.2.2
1	lasso	$\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},lasso}(\hat{\lambda})$		Section 1.2.7
1	post-lasso	$\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},postlasso}$		Section 1.2.7
1	ridge	$\hat{\boldsymbol{\beta}}_{m_0,ridge}(\hat{\lambda})$		Section 1.2.7
2	JMA	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{jma})$		Eq. (1.9), (1.17)
2	lasso-JMA	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{JMA})$		Eq. (1.33)
2	MCS t.max	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{T_{max}})$		Eq. (1.9), (1.26)
2	MCS t.min	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{T_{min}})$		Eq. (1.9), (1.26)
2	MCS-JMA t.max	$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{T_{max},jma})$		Eq. (1.31)
2	MCS-JMA t.min	$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{T_{min},jma})$		Eq. (1.31)
3	JMA		$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{jma})$	Eq. (1.12), (1.17)
3	lasso-JMA		$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{JMA})$	Eq. (1.34)
3	MCS t.max		$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{T_{max}})$	Eq. (1.12), (1.26)
3	MCS t.min		$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{T_{min}})$	Eq. (1.12), (1.26)
3	MCS-JMA t.max		$\hat{y}_{t+h\|t,fc}(\widehat{\boldsymbol{w}}_{T_{max},jma})$	Eq. (1.32)
3	MCS-JMA t.min		$\hat{y}_{t+h\|t,fc}(\widehat{\boldsymbol{w}}_{T_{min},jma})$	Eq. (1.32)

For comparing all 18 methods we adopt a two-stage procedure. First, we categorize the methods into three groups and look at each group individually. At the second stage we pool together the best performing methods from each group and judge their overall merits. The first group consists of those methods that perform model selection by placing all weight on a single model. For this group model averaging and forecast combinations are identical. The members of this group are listed in the top third of Table 1.2. The other two groups consist of those methods that place weight on more than one model and apply either model averaging (second group) or forecast combinations (third group). At the second stage we choose two methods from each group and compare the final set of methods again.

The criteria for choosing the second-stage methods are as follows. A method should deliver the best results for either large or small sample sizes for at least two out of the three DGPs. A method can be among the worst for a particular sample size or DPG as long as it shows merits for other constellations. However, if the performance of a method is dramatically worse, we will eliminate that method as a suitable contender, even if it shows merits in other regards. We also reserve some discretion in selecting methods for further comparison by considering the respective type of method. If three methods present themselves as good candidates, but two are of a similar type and with similar performance—both are shrinkage estimators for example—then only one of them will be chosen for the next round. This is to guarantee that the final comparison is informative on as broad a range of selection schemes as possible.

1.4 Results

In this section we summarize the main findings from our Monte Carlo simulation study using the setup described in Section 1.3. Having suitable measures and a structured procedure for comparison at hand, we look at the results for h-step-ahead predictions and for impulse response analysis in turn.

1.4.1 Forecasts

We illustrate the results by plotting the averaged relative MSEPs (1.40) for all sample sizes and DGPs. Figures 1.1 to 1.3 contain the first-stage results for each of the three groups. The three panels in each graph correspond to one of the three DGPs. Our performance measure, the averaged relative MSEPs, is plotted on the y-axis against the sample size on the x-axis. In all three graphs we see a clear trend towards zero as the sample size increases. This is as expected, because estimation and model uncertainty

is reduced as $n \to \infty$. The relative differences between the estimated forecasts and the DGP forecasts therefore start to vanish.

Looking at Figure 1.1 for the single-model methods, a clear pattern emerges. Lasso and ridge have an advantage at small sample sizes and perform between 5 and 15 percentage points better than post-lasso or ordinary information criteria. For larger sample sizes, the differences generally diminish, as noted above. However, the Schwarz criterion now has a slight edge over the other methods at large sample sizes and ridge is slightly inferior to lasso. Thus, for this group we pick lasso and Schwarz as the winners.

For the model-averaging group, Figure 1.2 reveals a similarly clear pattern. For smaller sample sizes and across DGPs, MCS with the T-max statistic has an advantage. Yet, with increasing R^2 and/or sample size, JMA starts to dominate MCS and the other methods. When the sample size is large and the signal is low, MCS based on equal weighting and the T-max statistic suffers from low power and is not able to remove or downweight inferior models as accurately as the other methods are capable of. This in turn leads to inferior forecasting performance. For small sample sizes and low R^2, the estimated MCS using the T-max statistic is almost equivalent to averaging across the initial model set $\mathcal{M}^0$. Apparently, it is better to average across almost all models in those situations than to apply one of the other averaging techniques.

Besides the MCS using equal weights, all other methods result from applying first a model selection tool and subsequently weighting with JMA the remaining (smaller) group of models. For forecasting purposes, forming these combinations does not yield an advantage over simply applying JMA to $\mathcal{M}^0$ in the present context. The combinations are either inferior or equivalent to JMA itself but never superior. Thus, for the second group we pick MCS based on the T-max-statistic and JMA as winners.

For the DGPs that we have chosen, model averaging and forecast combinations show only very slight differences with respect to forecasting, as can be seen when comparing Figures 1.2 and 1.3. The ordering of the third group is therefore the same as in the second group and we pick again MCS based on the T-max-statistic and JMA for the next stage.

Figure 1.4 depicts the key results, where all six winners of the first stage compete against each other. Now, some of the methods appear as their forecast combination (fc) version and as their model averaging (ma) version at the same time. However, as noted before, the respective versions do not make a large difference for the chosen DGPs.

Particularly noteworthy is the large difference between selection and averaging schemes. The two best model selection procedures, SIC and lasso, are the two worst performing methods when only few observations per parameter are available. Even at

$n = 100$ SIC is still among the worst methods across all three DGPs. When more observations become available, however, SIC begins to prevail. Another noteworthy feature is that model averaging has a slight advantage over forecast combinations, at least when evaluated over all $H = 15$ horizons.

Finally, JMA shows a promising performance, being among the best methods for $n > 80$, and only being beaten by simple MCS averaging when very little information is available (few observations per parameter, low/medium signal-to-noise). Thus, the overall results for forecasting indicate that using simple MCS averaging is useful for low information content, and otherwise relying on JMA will on average lead to generally good results. Surprisingly, the increasingly popular lasso and standard criteria such as SIC cannot be recommended from this analysis.

1.4.2 Impulse Response Analysis

The results for impulse response analysis are not as clear-cut as for forecasting. The ranking of methods is more heterogeneous, depending to a greater extent on sample size and signal-to-noise. The bottom line is, however, that for small samples MCS with the T-max statistic and using JMA weights performs fairly well. With sufficiently many observations, on the other hand, the Schwarz criterion is the best choice across DGPs. Another prominent feature is that several measures increase with sample size, in contrast to the previous section. The denominator of the current measure (1.41) approaches zero as $n \to \infty$ while the denominator of the previous measure (1.39) converges to a constant. Depending on the speed of convergence of the numerator and denominator of (1.41), this ratio may therefore behave more erratically.

The group of single-model methods in Figure 1.5 indicates a more dispersed yet still similar performance pattern as in the prediction exercise. In small samples, lasso and ridge do well. In larger samples, Schwarz outperforms all others. This is consistent across DGPs. AIC, HQC, and post-lasso perform between 20 and 200 percentage points worse in larger samples than Schwarz does, and are therefore markedly inferior by our measure. The RMSE of ridge seems to converge at a completely different rate than that of the true model. We therefore rank it lower than lasso and choose lasso and Schwarz as winners.

When it comes to model averaging (Figure 1.6), every method is at some point best performing and there is therefore no clear ordering. MCS based on the T-max statistic performs well in small samples for $R^2 = 0.2$ and 0.5 but shows very poor performance for $R^2 = 0.8$. Lasso-JMA and simple JMA do well for high signal-to-noise ratios, but not for $R = 0.2$. Methods without large performance losses, on the other hand, are given by the MCS using the T-min statistic and by combining MCS and

JMA with either T-max or T-min, even though they do not sizeably outperform other methods. These three methods are therefore fairly robust tools, which we pick for later comparison.

Comparing model averaging and impulse response combinations yields some difference, especially for smaller sample sizes. The overall strengths and weaknesses, however, are again the same. That means every method is again at some point the best. The following points stand out. For $R^2 = 0.8$ one has a substantially worse performance of those methods that include the most models, such as JMA or MCSJMA with the T-max-statistic. This can be attributed to the fact that we cannot weight models optimally for combining impulse responses, in contrast to out-of-sample predictions. Nonetheless, the ranking between the methods remains similar. We have, therefore, heterogeneous results across DGPs and sample sizes for all methods except the combination of MCS and JMA, which is again fairly robust across DGPs and we therefore pick both versions for the next round.

Overall, we see in Figure 1.8 that when the sample size is relatively large, using the Schwarz criterion for selecting models in the context of impulse response analysis dominates all other methods. Especially when faced with a low-signal stochastic process and when the observation-to-parameter ratio is above 25 ($n \geq 250$) does Schwarz outperform other approaches by 20 to 200 percentage points. In contrast, when the observation-to-parameter ratio is small, so between 4 and 12, then Schwarz is dominated by essentially all other methods. Our results indicate that relying on model averaging via MCSJMA using the T-max-statistic is the most robust method across DGPs. This method also performs slightly better than simple JMA for $R = 0.8$ and many observations per parameter. Very relevant for applied work is also the weak performance of lasso, since it is increasingly popular with practitioners. This weak performance is not remedied by relying on unbiased or less biased parameter estimates via post-lasso.

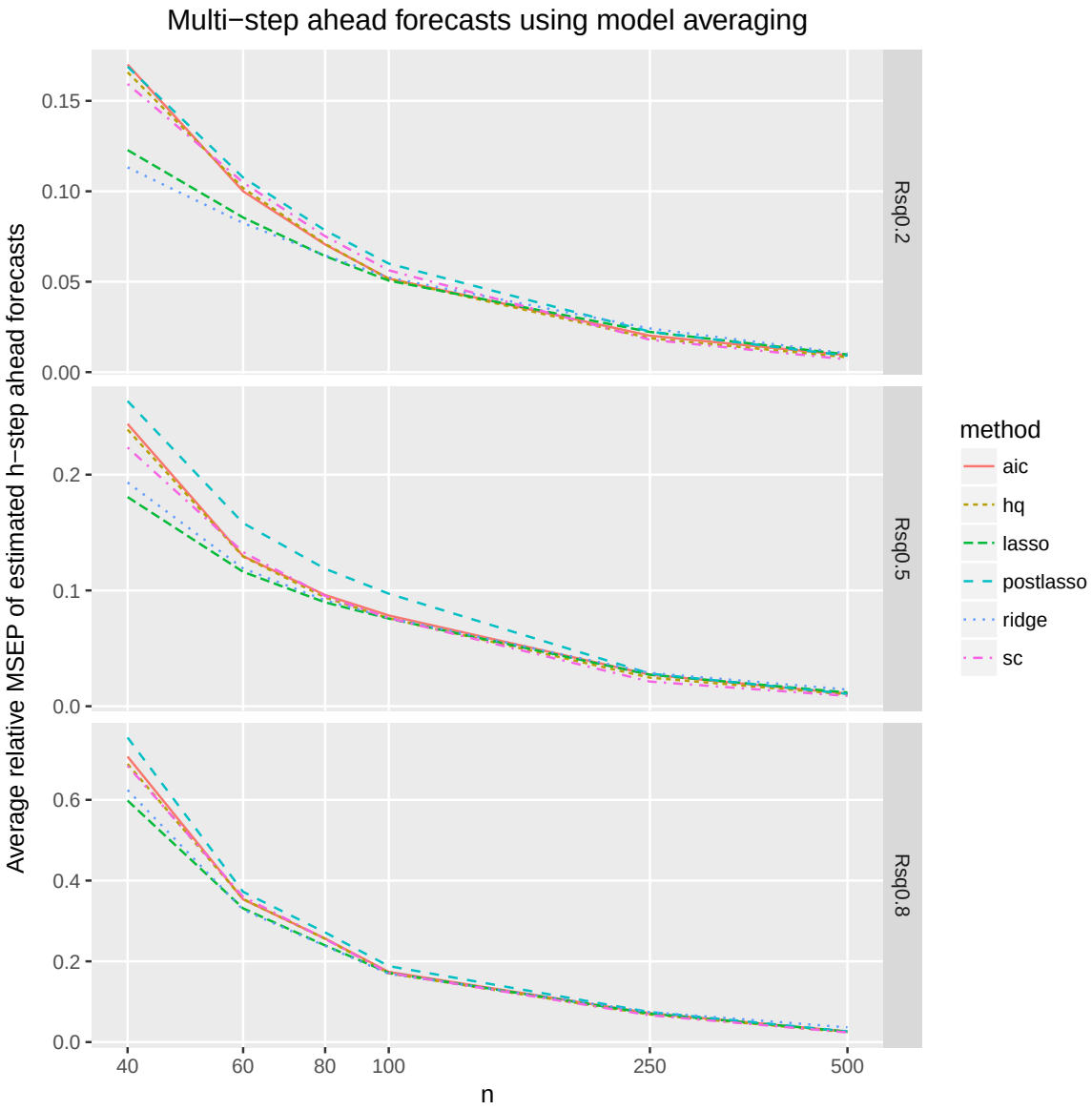

Figure 1.1: Averaged relative MSEPs for h-step ahead predictions for methods selecting and estimating a single model.

Notes: Each line shows the average relative mean squared error of prediction (1.40) and is based on $R = 5000$ replications. The DGPs are AR(8) processes with zero restrictions and differ w.r.t. their signal-to-noise ratio. Their specification is given in Table 1.1. References for the methods used are given in Table 1.2. The various auxiliary parameters chosen for some methods are described in Section 1.3.2. All simulations are carried out in the R programming language.

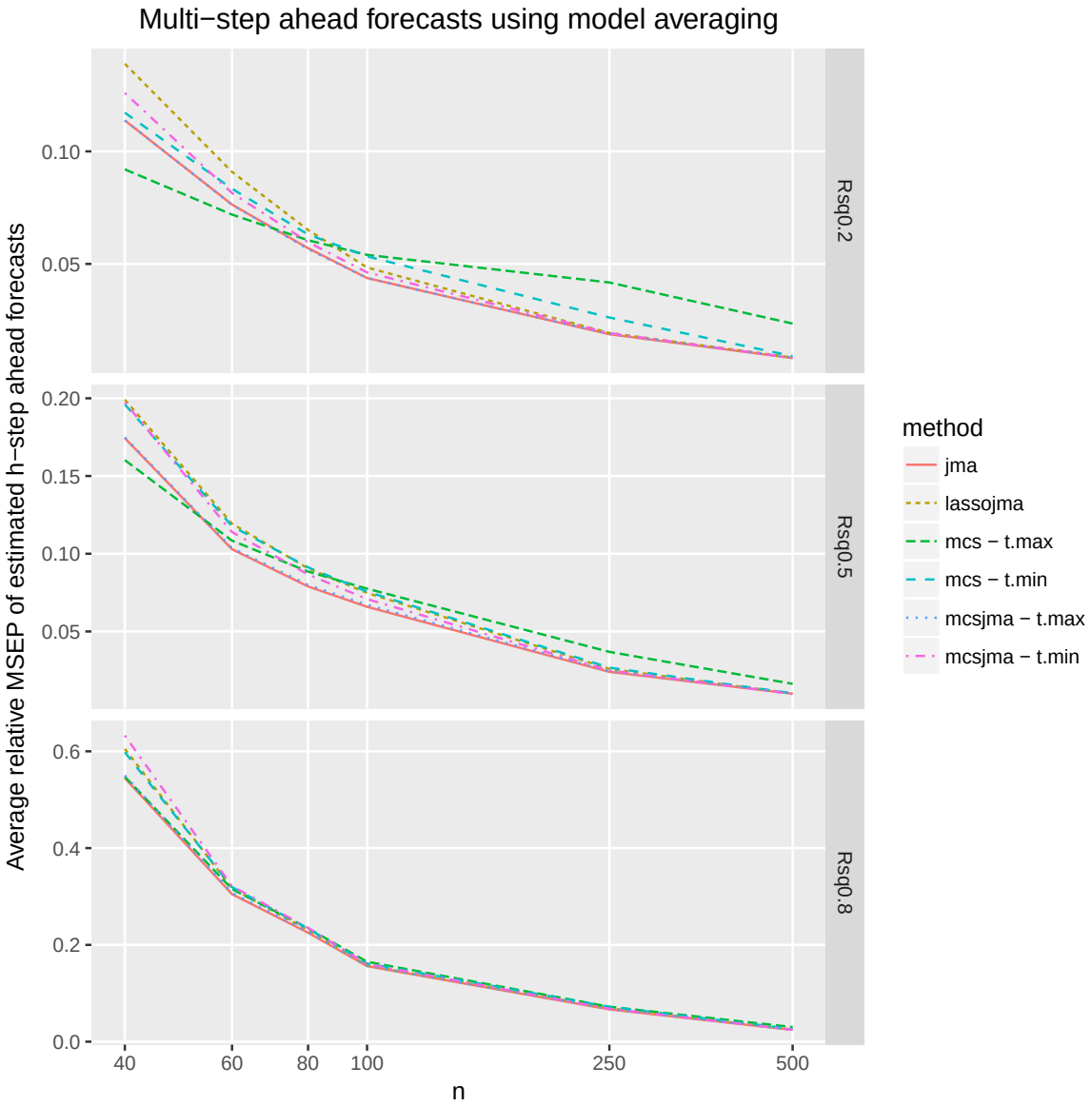

Figure 1.2: Averaged relative MSEPs for h-step ahead predictions for methods using model averaging.

Notes: see Figure 1.1.

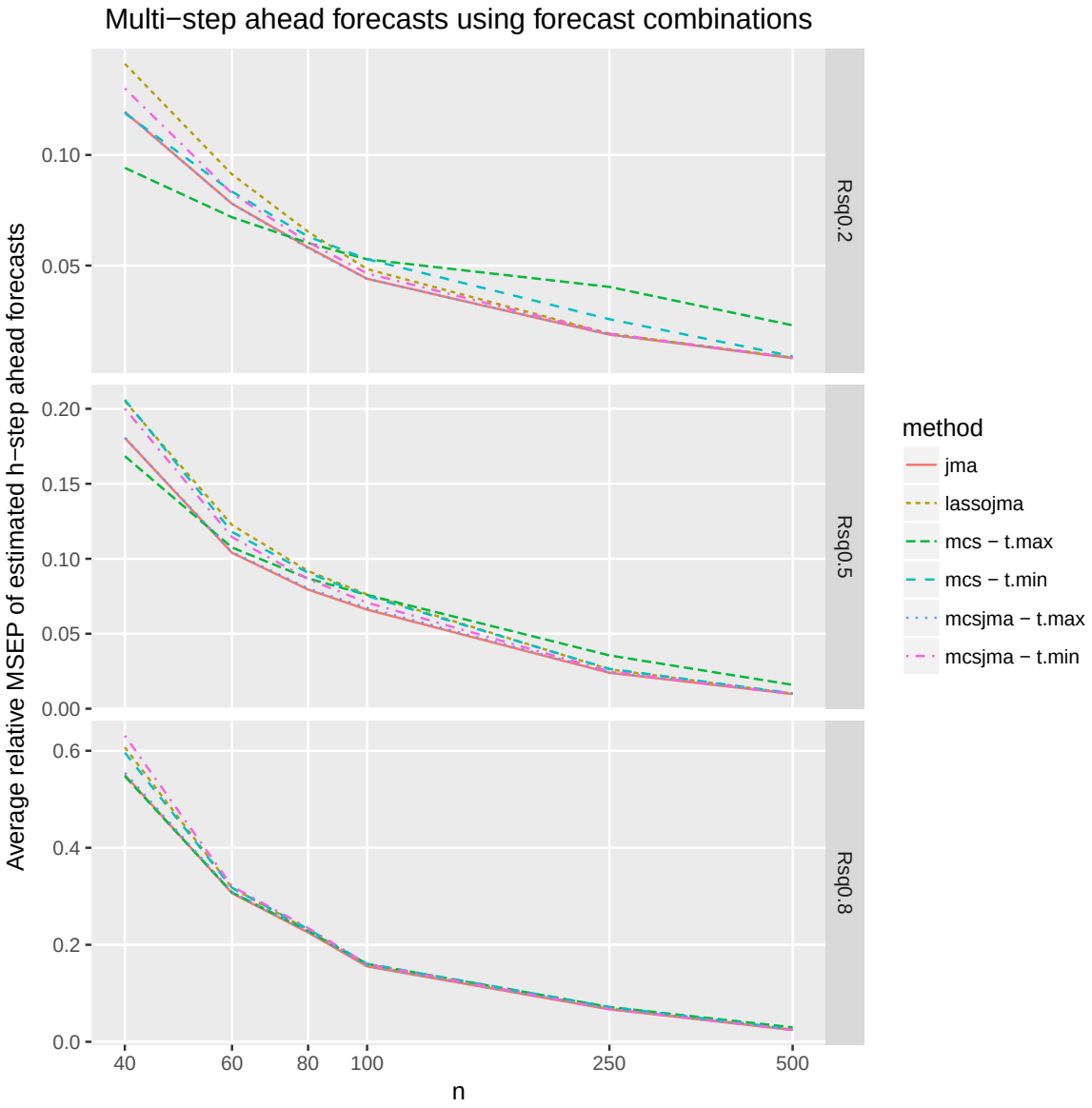

Figure 1.3: Averaged relative MSEPs for h-step ahead predictions for methods using forecast combinations.

Notes: see Figure 1.1.

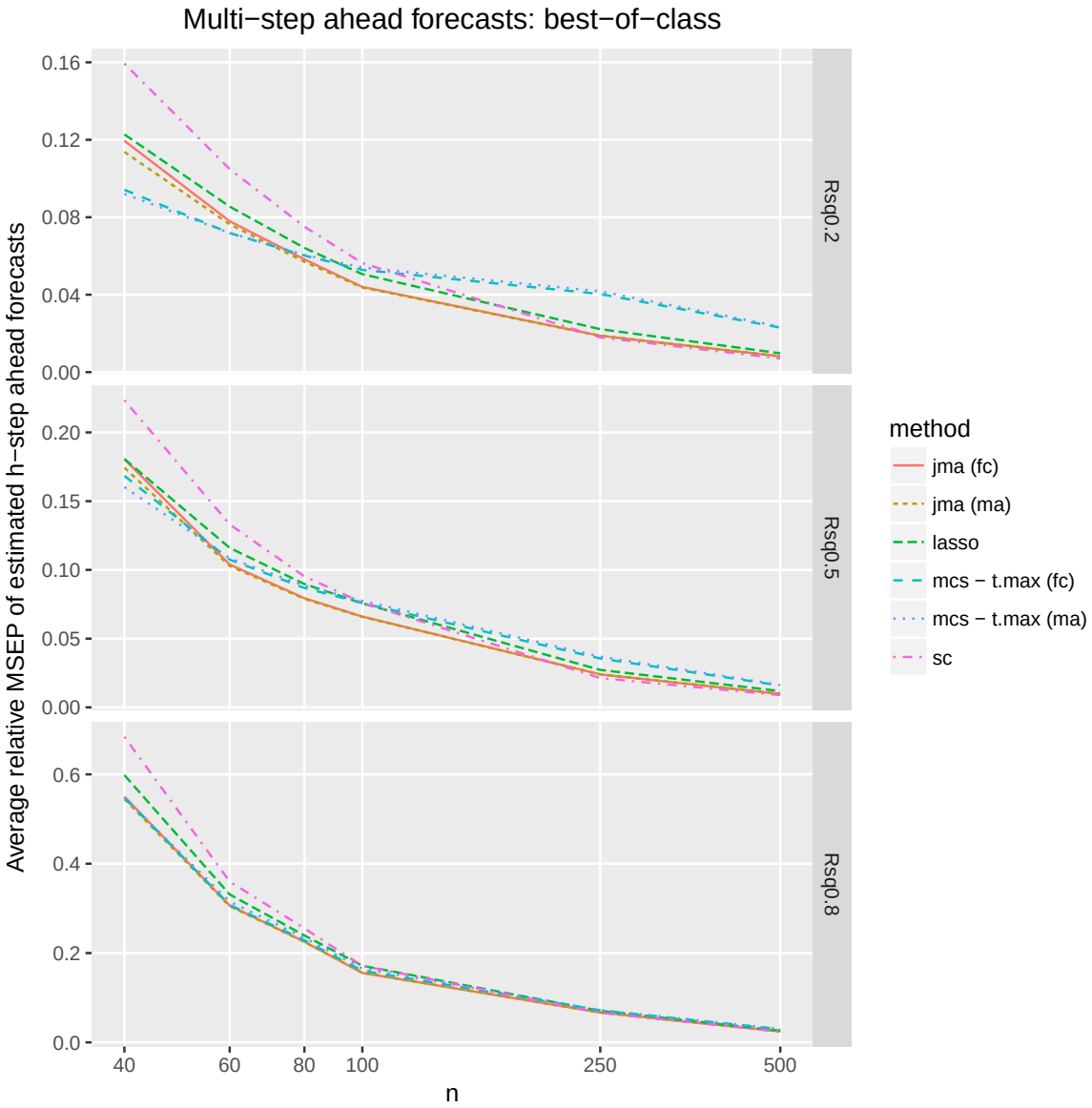

Figure 1.4: Averaged relative MSEPs for h-step ahead predictions for the best performing methods.

Notes: see Figure 1.1.

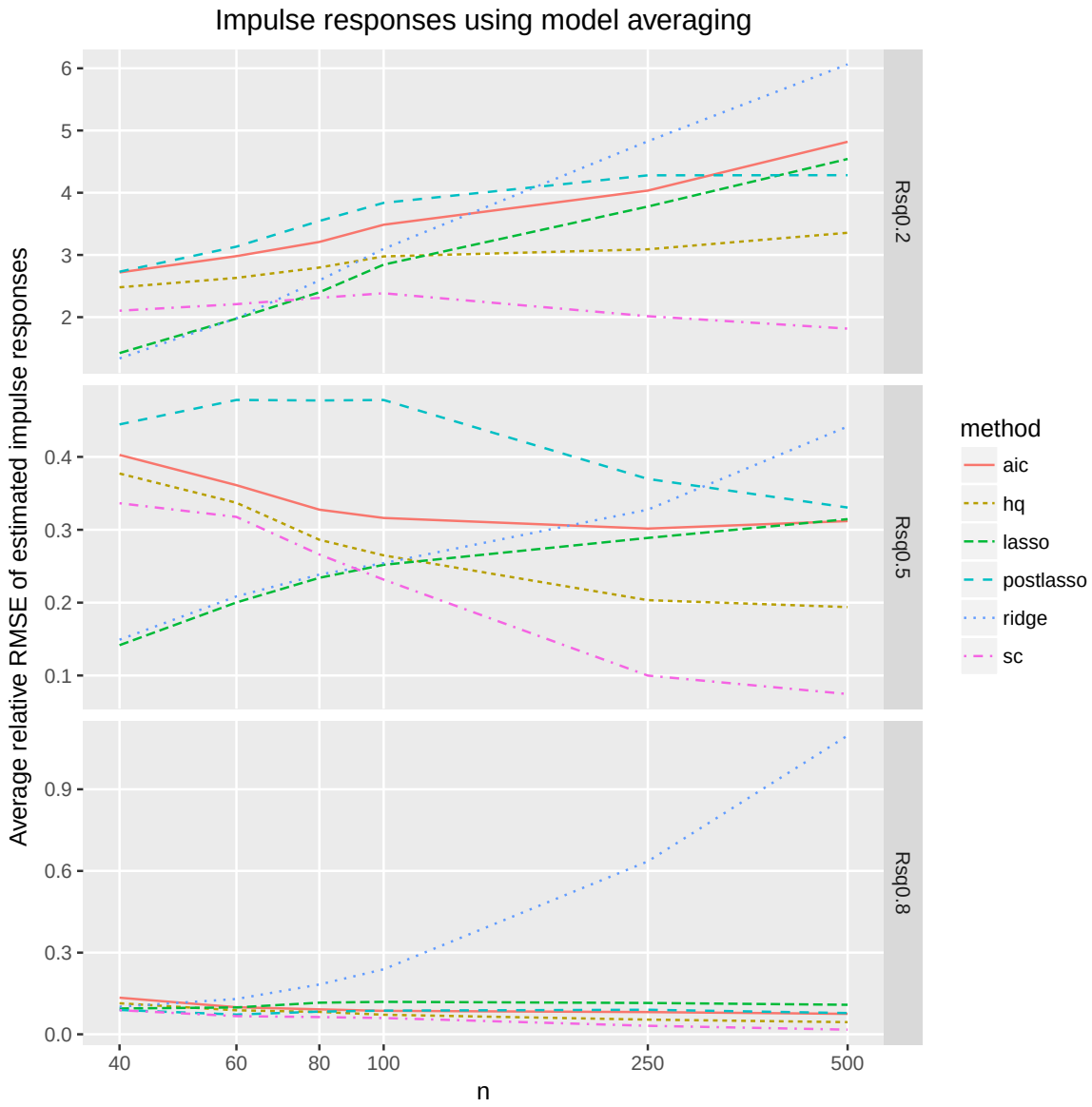

Figure 1.5: Relative RMSEs averaged over $H = 20$ impulse response estimates for methods selecting and estimating a single model.

Notes: Each line shows the average relative root mean squared error (1.42) and is based on $R = 5000$ replications. The DGPs are AR(8) processes with zero restrictions and differ w.r.t. their signal-to-noise ratio. Their specification is given in Table 1.1. References for the methods used are given in Table 1.2. The various auxiliary parameters chosen for some methods are contained in Section 1.3.2. All simulations are carried out in the R programming language.

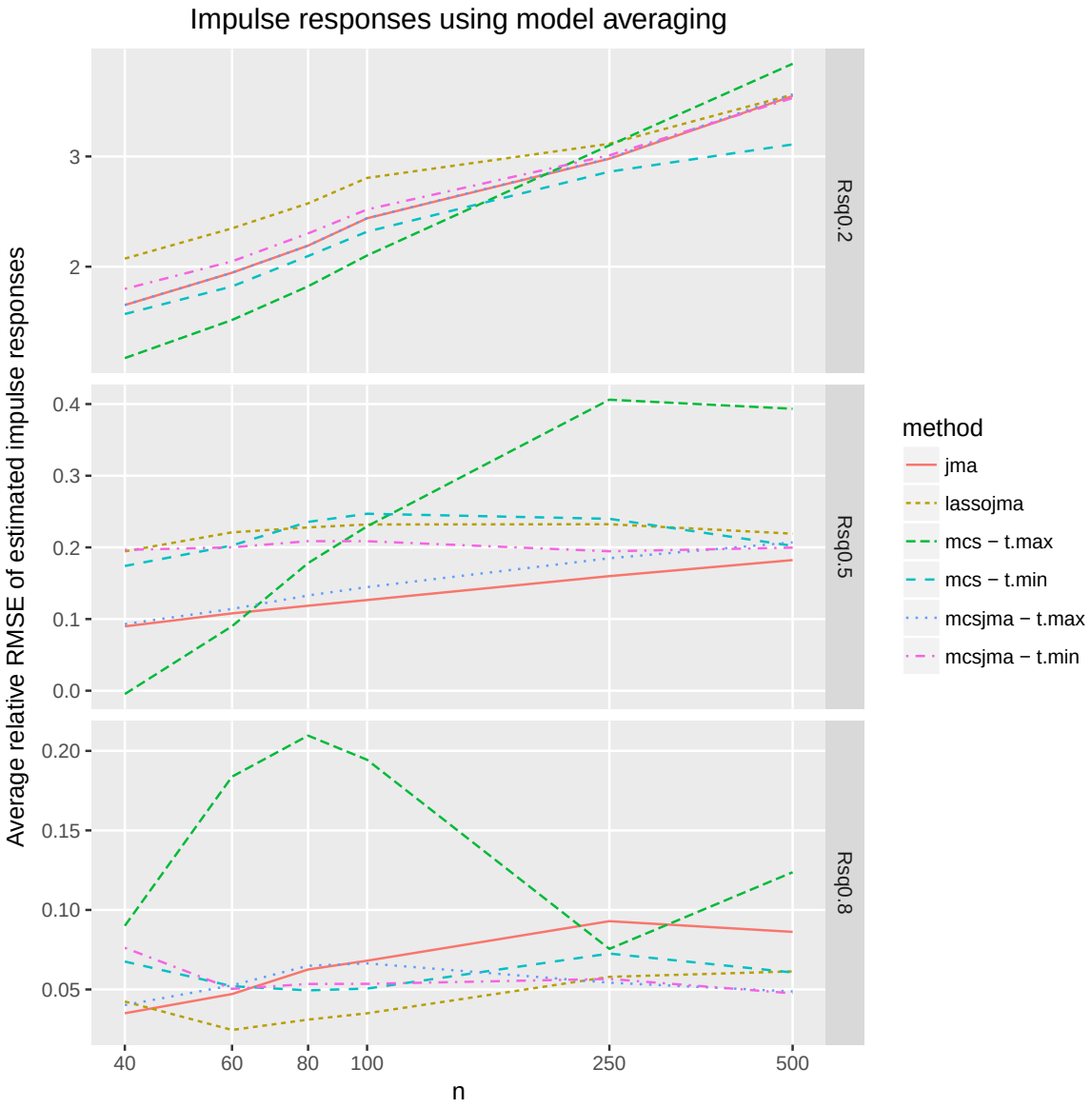

Figure 1.6: Relative RMSEs averaged over $H = 20$ impulse response estimates for methods using model averaging.

Notes: see Figure 1.5.

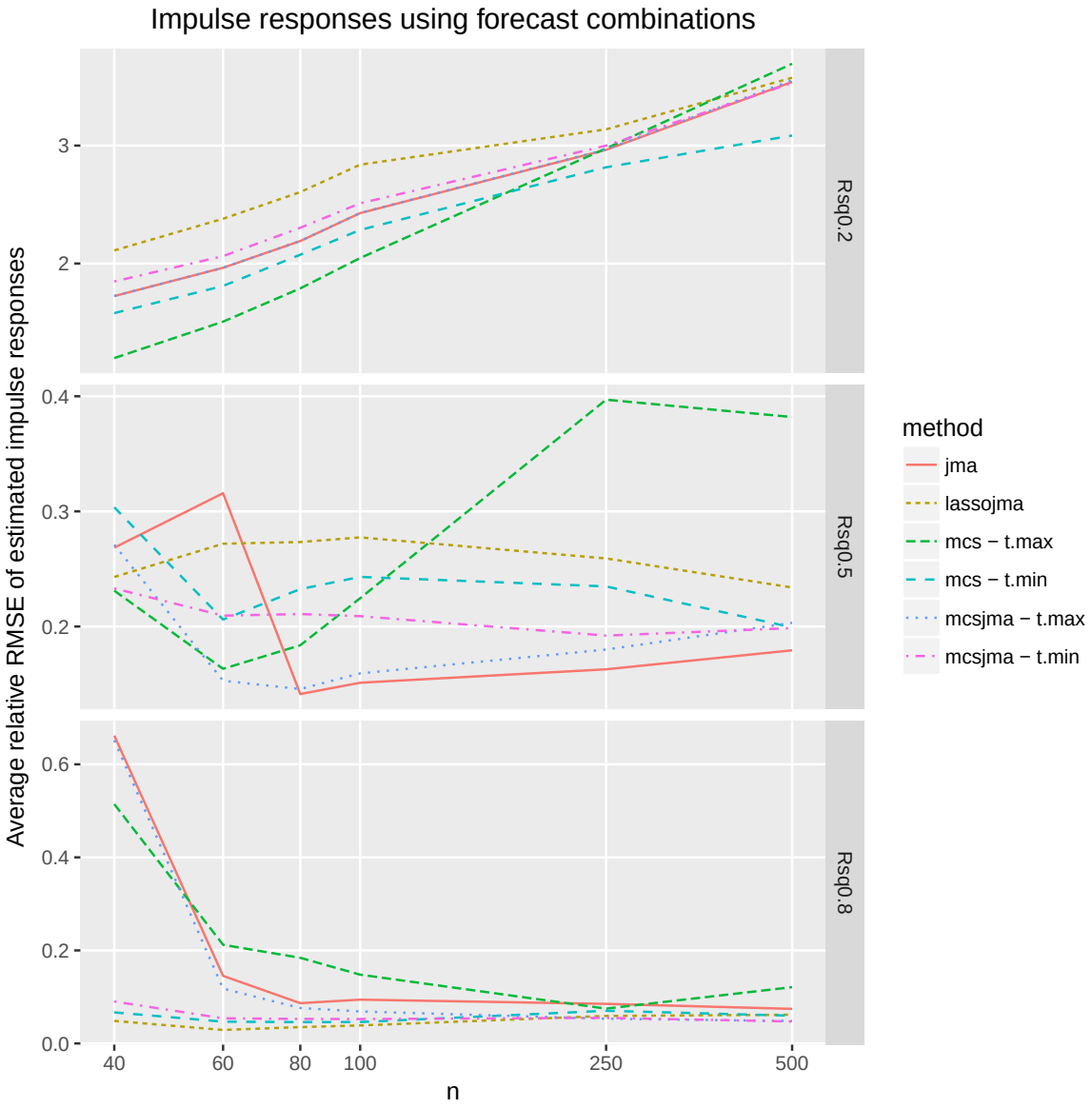

Figure 1.7: Relative RMSEs averaged over $H = 20$ impulse response estimates for methods using forecast combinations.

Notes: see Figure 1.5.

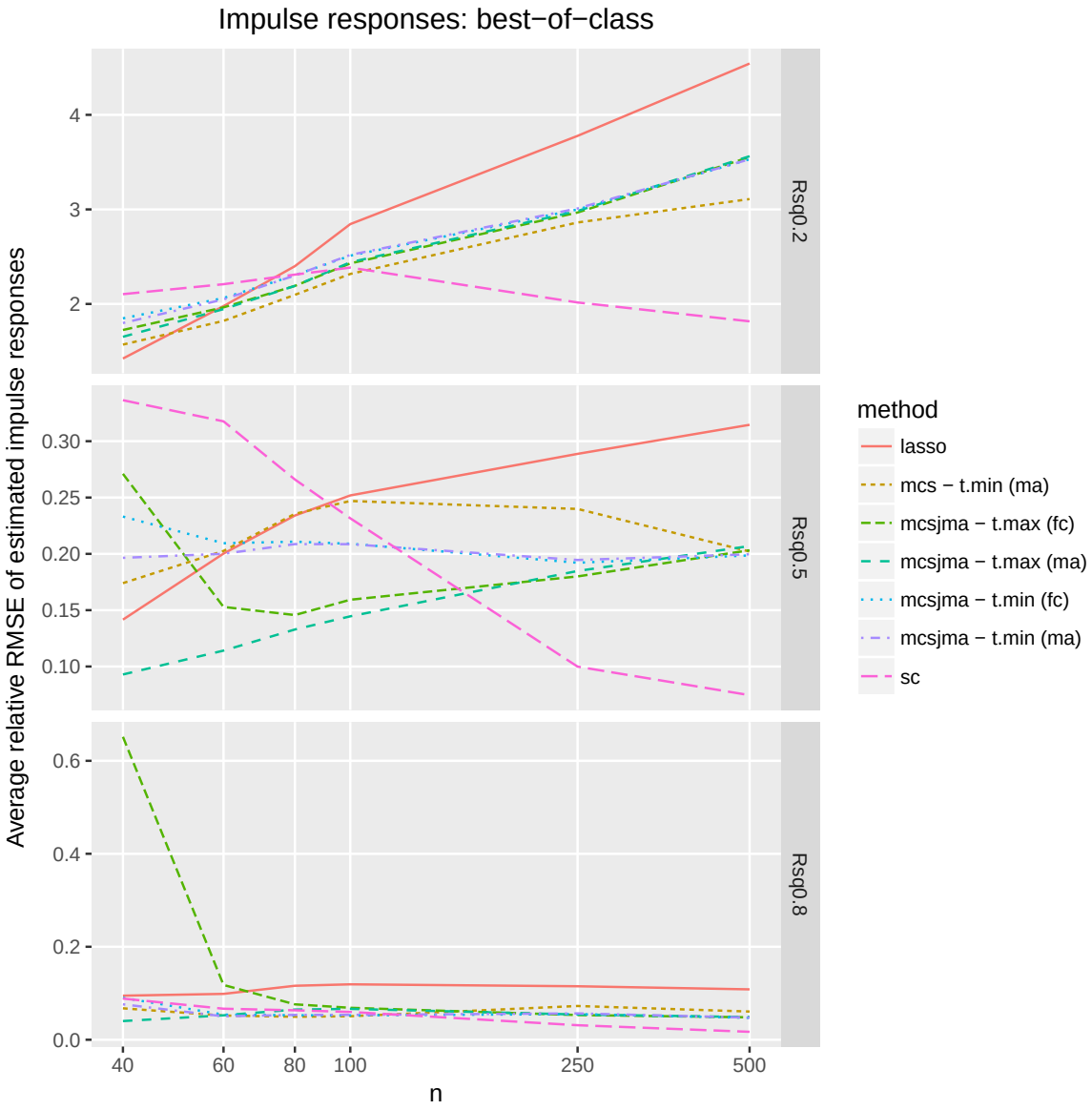

Figure 1.8: Relative RMSEs averaged over $H = 20$ impulse response estimates for the best performing methods.

Notes: see Figure 1.5.

1.5 Conclusion

In a Monte Carlo simulation, we have studied the effects of model selection on point forecasts and impulse response estimates in the context of simple autoregressions. More specifically, we have looked at the effectiveness of different selection and weighting schemes when accounting for model uncertainty in subset regressions. Effectiveness is measured in our study by reductions in mean square error (relative to the data-generating process and averaged over multiple horizons). These reductions can be considerable, amounting up to 40% for forecasting and up to 70% for impulse response estimates.

The methods that we compare include standard information criteria, the model confidence set (MCS), jackknife model averaging (JMA), and penalized regression via lasso and ridge. We have paid particular attention to model confidence sets, since they specifically address the issue of multiple testing in the context of model selection and may therefore be beneficial. For forecasting and impulse response analysis it can, however, be unclear how to proceed with the analysis once a (potentially large) set of models has been estimated via the MCS.

One possibility to deal with many models is to average across all of them. We study equally weighted MCS averages as well as JMA-weighted MCS averages. We further contrast model averaging, in which parameters are averaged, and forecast combinations, in which non-linear functions of parameters are averaged.

We draw important lessons for applied work. When evaluating a large number of models, 256 in our case, standard information criteria only perform well when model uncertainty is fairly low. In that case they are among the best performing methods for both forecasting and impulse response analysis. In case uncertainty is high, there are meaningful advantages in adopting averaging schemes, either based on simple MCS averages or JMA-weighted averages. Importantly, standard shrinkage methods such as lasso and ridge underperform in most situations. One may therefore speculate that they excel in high-dimensional settings, e.g. when there are more variables than observations, but not in our case.

According to our Monte Carlo results, the Schwarz criterion is particularly effective in selecting a good model when sufficiently many observations are available (per parameter) for both forecasting and impulse response analysis. For forecasting, weighting all models via JMA works almost as well in large samples. With few observations it is much better than Schwarz. JMA is therefore a robust choice for forecasting. This is in contrast to impulse response estimates where JMA fares relatively poorly and may be improved by first removing inferior models via the MCS. We further find that averaging parameters (model averaging) is more effective than averaging functions of

parameters (forecast combinations), with larger gains in impulse response analysis than in forecasting.

Subsequent research could investigate the potential of the proposed methods for lag selection in multivariate time series models and compare them to other relevant tools such as stepwise subset selection.

CHAPTER 2

Connecting the Dots: Structural VARs and Causal Graphs

2.1 Introduction

Measuring and uncovering causal relations is key for many scientific endeavours, yet often fraught with conceptual and practical ambiguity. Over the last two decades, graphs have become an increasingly useful tool for operationalising causal concepts. They facilitate reasoning about causal relations and, given a set of specific assumptions, they also facilitate identification of causal relations between several variables. In macroeconomics, these causal graphs have also been applied as a means to identify SVAR models.

Causal graphs and structural VAR modelling are two distinct fields with different, but related, concepts. While some authors in macroeconomics have elaborated on graphical concepts in passing, the precise relationship between graphs and VARs has not always been made quite clear. The current study addresses some of these shortcomings by explicating the relationship more thoroughly. What kind of VAR models, for example, can be identified by exploratory causal graph procedures? Can we tell from a graph whether a VAR is identified? Answering these questions, and others, is the main contribution of this paper.

Since causal graphs may be unfamiliar to most macroeconomic researchers, I will first take a step back and establish context in Section 2.2. In particular, I contribute by reviewing the literature on causal graphs in macroeconomic analysis in more depth. I then discuss the main method underlying graph-causal search procedures and highlight some pitfalls that researchers should be aware of.

Merging graphs into macroeconomic analysis may be worthwhile for the following reasons. Graphs have a precise definition in terms of sets, being constructed from vertices and edges that connect these vertices. They thus lend themselves to formalisa-

tion, while preserving a great deal of intuition through ease of visualisation and some abstraction. SVAR analysis investigates economically meaningful structures, usually involving interpretations of cause and effect. Merging causal graphs into the existing SVAR toolset may thus allow another kind of formalisation and to facilitate interpretation. This kind of graphical formalisation further opens the door to exploratory analysis of causal relations. On the upside, such an exploratory approach based on graphs could formalise existing, but informal, practices of exploratory data analysis. On the downside, it is easy to grow overconfident with the results provided by such data-driven causal insights.

Data-driven causal search procedures, while potentially informative, should be applied with caution since they may lead researchers astray, as I highlight below. As with other search procedures, they inductively infer properties about the data generating process (DGP) from statistical test decisions. With causal search, fairly strong claims about the DGP may result from a negative test decision, a failure to reject, without any quantification of uncertainty surrounding this decision. Investigators usually confine themselves to deduction with good reason, relying only on positive test outcomes—especially on matters as delicate as cause and effect. Faced with uncertain circumstances, how does one verify whether an effect is indeed absent or whether it is simply too elusive due to the sample being too short or the data too noisy? There is an unquantified potential for severe misclassifications. Thus, resting on exploratory, inductive causal data analysis warrants great caution. A small simulation exercise will illustrate this point.

Nonetheless, causal graphs can prove insightful, in particular as a formal language to reason about identifying assumptions. The identification of economically meaningful shocks is a core concern in the SVAR literature and a multitude of methods exist to achieve identification. While the necessary and sufficient conditions for identification via zero restrictions are well known in their algebraic form, causal graph theory may aid practitioners by translating these algebraic conditions into more easily digestible visual conditions. Graphs thus offer a communication device that may further pedagogical or presentational purposes. In this context, this study makes properties of graphs explicit that have so far only been implicitly acknowledged, if at all. A more detailed account of causal graphs, causality and macroeconomics is also provided while avoiding too many technicalities. The study thus also fills the gap of an introductory review of existing causal graph methods and studies in the context of SVAR models.

In sum, while pure data exploration is most likely ill-suited to inform the researcher on the true causal structure that generated the data, the insights provided by causal

graph theory may be used to improve presentation, to provide graphical criteria for ascertaining identification, and, possibly, to carefully explore properties of the data.

Section 2.2 reviews the literature on causality and causal graphs with a focus on macroeconomics. Section 2.3 describes the main exploratory approach that has been used to identify structural VARs and illuminates some of the required assumptions. Section 2.4 illustrates how causal search can go wrong. Finally, Section 2.5 clarifies the properties of causal graphs in the context of structural VAR analysis. Section 2.6 will conclude and point to questions for further research. Appendix 2.A.1 summarises key concepts in probabilistic graph and causal graph theory.

2.2 Literature Review

This section reviews the nexus of causality, macroeconometrics, and causal graphs. Concepts of causality are implicit in most applied economic and statistical research but are only occasionally made precise and may vary depending on context and method. Nevertheless, there is a rich history in econometrics as to what constitutes a causal effect and how to measure it. I will start by shortly reviewing these past discussions in Section 2.2.1. Next, Section 2.2.2 will discuss the literature on causal graphs. The section highlights two distinct ways in which causal graphs have been applied, one exploratory, the other confirmatory. Finally, Section 2.2.3 will focus on how causal graphs have been integrated into macroeconomic research, concluding that the dominant approach is exploratory, not confirmatory.

2.2.1 A Few Notes on Causality

A universal definition of causality is not within the scope of this paper. It is a topic of similar breadth and complexity as answering 'what is truth?' Nonetheless, I will briefly mention some features that are frequently attributed to causality (at least as it pertains to economic systems) in order to have some idea of what we are talking about. First and foremost, causality describes a stable and autonomous relation between two entities across space or time. The insight of such an abstract characterisation is as follows. Stability suggests that we can identify a reliable mechanism that connects both entities. A mechanism that can be exploited for predicting the state of one entity after observing or changing the state of the other. Autonomy guarantees that this mechanism is unaffected by outside alterations. These two features together are very useful because they make a causal relation susceptible to controlled manipulation and independent of context. A third feature that would be convenient is asymmetry. With asymmetry we would be in a better position to preclude some relations from being causal. If they

are not causal, we know that the relations would not necessarily hold up if transferred to another context. If asymmetry is a given, then we could, for example, expect that any cause happens before its effect materialises or that directionality prevents an effect from simultaneously causing its cause. Causal feedback loops would therefore not be present. However, whether asymmetry is indeed a feature of all causal relations is a more debated issue. For in-depth discussions of causality, including its historical and various philosophical aspects, the Oxford Handbook of Causation (Beebee, Hitchcock, and Menzies, 2009) is a plentiful source.

Economists and econometricians are concerned with measurement. So when it comes to causality we need to be a good deal more pragmatic if we want to identify and estimate such an abstract notion as a causal effect. But finding the right balance between pragmatism and rigour has proven difficult. Statisticians, for example, rigorously advise to distinguish between causality and other purely associational concepts relating two or more random variables to each other, such as correlation. While they have been confident in their handling of these associational measures, there has usually been some weariness, discomfort or even outright rejection of causal analyses. Pearl (2009, Epilogue) gives a detailed, while slightly one-sided account of this attitude in the statistics profession. The weariness towards causality is also exemplified by Cox and Wermuth (1996) in their multivariate statistics textbook. Even though they concede that "statisticians concerned with the interpretation of their analyses have implicitly always been interested in causality", they state that they "have not in this book used the words causal or causality" since firm conclusions about causality would be rare, at least from a single study (p. 285). While the latter statement is certainly a helpful warning, it is also seldom spelt out under which circumstances and to what degree a study could support causal conclusions, firm or otherwise.

In response to such neglect, Angrist and Pischke (2017, p. 126) lament that "newer and widely-used tools for causal analysis [...] get cursory textbook treatment if they're mentioned at all." They further document that causal effects, including the causal interpretation of regression estimates and possible pitfalls, cover around 3 % of (hand-picked) contemporary econometric textbooks, up from 0.7 % in the 1970s. They claim that this low coverage is at odds with actual research practice. In fact, econometricians do have a range of tools at hand to study causal effects in well-defined frameworks.

In macroeconometrics, matters of causality take centre stage in structural equation modelling. In this kind of modelling a set of equations is arranged such that each equation reflects a specific, self-contained component of the economy. There is a clear understanding that each equation contains a variable, usually on the left-hand side, whose value is determined, up to some random noise, by an economic process that

accepts all right-hand side variables as its input. This kind of modelling and its theoretical foundation originates from the econometrics literature of the 1950s and 60s (Wold, 1954; Strotz and Wold, 1960; Basmann, 1963, 1965) and especially from work associated with the Cowles Comission (Simon, 1954; Koopmans, 1949). The notion that influence runs in a certain direction, is self-contained and stable is decidedly a causal notion. Coming up with causal relations between the variables under scrutiny is therefore a necessary requirement for structural analysis. A major question is how to come up with them. Can we deduce causal relations from higher-order principles or do we need to generalise from properties of the data at hand? Whether both or none of these approaches are legitimate touches upon age-old debates about the nature of scientific inquiry and knowledge. See Keuzenkamp (2004) for a perspective on these philosophical aspects in the context of econometric modelling. Meanwhile, the predominant approach in the econometrics profession seems to be that causal relations are in large part informed by theoretical considerations and open to falsification by testing whether theoretical expectations hold in the data. Sims (1977), for example, suggests to "test whether a system is structural by using it to predict the effects of an intervention" and notes that "we may prove the system is *not* structural, but there can be no guarantee [that it's structural]" (p. 28). A further point of debate is whether a structural system is always recursive, which would preclude instantaneous feedback loops between variables. Of course, if one subscribes to the view that causal relations are always asymmetric, then recursiveness is a natural way of modelling the economy (Strotz and Wold, 1960).

A particular class of structural models is the focus of this paper. In structural vector autoregressive (SVAR) models (Sims, 1980), structure is imposed in such a way that the causal effect of a very particular kind of intervention can be studied, while keeping a priori restrictions to a minimum. This intervention takes the form of an innovation to each of the equations of the system. An innovation that has a specific economic interpretation and whose dynamic, causal effect can be traced through the system and over time. A review of SVAR models in the context of causality and structural modelling is given by Kilian and Lütkepohl (2017, chap. 7).

The discussion on causality in econometrics has many more facets. Others who have contributed to this discussion besides those already cited include Granger (1969), Sims (1972), and Lucas (1976). Zellner (1979, 1988) and Leamer (1983, 1985) offer interesting perspectives on previous developments. More recently, White and Lu (2010) and White and Pettenuzzo (2014) develop a unifying framework connecting previous notions of causality in structural and reduced-form equation models. For recent reviews

of philosophical aspects regarding causality in macroeconomics, see Henschen (2018) and Maziarz and Mróz (2019).

Microeconometricians, too, have a well-defined framework at their disposal to study and measure causality. Their framework is born from an experimental mindset, probably since controlled manipulation is far easier and more frequent than in macroeconomic systems. Holland (1986) sets out three constituent marks of this framework. First, any causal analysis should investigate the effect of causes instead of the causes of effects. We can, for example, investigate the effect of smoking on lung health and at least hope to gain a definitive answer as the question sets out an action and its consequence, a beginning and an end. Finding out the cause of lung cancer, on the other hand, is open-ended and subject to revisions of the current state of research. This may seem trivial, but in everyday conversations just as in academic discussions it is not uncommon to ask "what caused the great recession?" Second, a causal effect is always a relative measure. There needs to be a baseline scenario, for which the cause is absent, for comparison to when the cause is present. This naturally gives rise to creating a treatment and a control group. Holland (1986) emphasises that it must be possible to expose every unit of study *in principle* to both treatment and non-treatment. Thus, the third constituent mark is that not everything can be a cause in this framework. Gender, for example, generally cannot be controlled by an experimenter. Therefore, studying the *causal* effect of being male or female on labour income is a question that is ill-posed for empiricists since, if it cannot be changed at random, it can also not be identified by empirical means (although one may vary it on paper, for example on written application forms). This attitude is sometimes summarised under the heading of analysing 'potential outcomes' as opposed to more generic counterfactuals. A similar slogan is 'no causation without manipulation.'

These three characteristics give rise to a well-defined causal effect (or 'treatment effect') which centres on the average difference between treatment outcome and control outcome. Estimating these averages consistently is, of course, a whole other matter and dealt with in a large body of literature. Imbens and Rubin (2015) offer a classic introduction while Hernán and Robins (2020) merge recent developments including graphical models; the contributions in Morgan (2013) provide a critical assessment of the potential outcome framework. As these books demonstrate, the concept of the average treatment effect (ATE) and related measures are nowadays well established in the microeconomic literature. Nonetheless, the estimation of these statistics is still developing, especially with new methods from the machine learning literature arriving to econometrics. See Athey and Imbens (2019) for a recent review.

A recent comparison of the potential outcome framework and the graphical approach to causality is given by Imbens (2019). The author suggests that the two approaches may complement each other. He also finds that the graphical approach merits more attention than it has received in the econometrics literature so far. In particular, he praises graphical tools for their pedagogic value and for their ability to answer complex causal queries in a systematic fashion. Though he also highlights drawbacks. One of them is the lack of convincing graphical applications which would be essential for showcasing the value added: insights that are less intuitive, difficult or impossible to gain within the predominant potential outcome framework. Without examples that go beyond simple toy models there is little reason for researchers to be convinced of the graphical toolset's practical merits. A further blind spot pointed out by Imbens (2019) is the lack of attention that topics besides identification receive. Confirmatory causal graph analysis is mainly concerned with deciding whether certain causal effects are identifiable or not, given the joint probability distribution of the variables under investigation and some structural assumptions. Causal graph analysis therefore only rarely touches upon topics such as data collection, modelling, estimation, or inference.

2.2.2 Causal Graphs

The literature on causal graphs broadly falls into two categories. One branch develops and applies methods that search for causal relations in any given data set, the other seeks to efficiently verify whether assumptions made for a specific causal analysis hold in the data. The first branch is quite aptly called causal search or causal discovery, the other I will call causal validation. I will discuss both in turn.

The causal search endeavour is very much embedded in the machine learning literature where attempts are made at finding (causal) relations in the data without the need for subject-specific background knowledge (see Peters, Janzing, and Schölkopf, 2017). Retrieving causal relations from data by means of graph theory was in particular popularised by Spirtes, Glymour, and Scheines (2000) and Pearl (2009). Spirtes, Glymour, and Scheines (2000) develop and collect algorithms that automate the search for causal effects without the need for domain specific background knowledge or experimental setups involving counterfactuals such as randomised control trials. They proceed as follows. Initially, they assume the stochastic dependence relations between variables known. Given certain axioms that relate graphical models to the underlying probability distribution of the data generating process, they show how graphs can guide the identification of causal effects. Of particular importance are patterns of conditional and unconditional independence which preclude the existence of certain causal relations and which can be directly inferred from the graph. In practice, these stochastic inde-

pendence relations are not observed, of course, and what researchers are left with are empirical moments of the data. Some of the criticism that graphical search algorithms have drawn originates here. In the end, claiming that causality can be inferred from purely observational data hinges to a great extent on how well independence between variables can be inferred from the data.

Robins and Wasserman (1999), for example, criticise the claim of Spirtes, Glymour, and Scheines (2000) that identification of causality is possible without subject-specific background knowledge. They argue, in a Bayesian setting, that Spirtes, Glymour, and Scheines (2000) implicitly assume a bound on the growth of the number of unmeasured confounders in their asymptotics. If the bound is exceeded, then the probability of assigning causality between two variables approaches 1, independent of whether such a causal link exists. The authors therefore caution against the use of such search algorithms when confronted with purely observational data. In such instances, they maintain, the number of unmeasured confounders will be difficult to control and arguably very large. Humphreys and Freedman (1996) heavily criticise the agenda of Spirtes, Glymour, and Scheines (2000), too. They note that the authors do not define what they mean by causation and simply introduce an assumption that "arrows represent causation" (p. 114). Furthermore, Spirtes, Glymour, and Scheines (2000) do not account for any statistical uncertainty faced by their algorithms and make unreasonable assumptions about their data, such as i.i.d. normal or multinomial observations, which cannot be taken for granted when attempting to *automate* the search for causality. They therefore conclude that "the whole development is only tangentially related to long-standing philosophical questions about the meaning of causation, or to real problems of statistical inference from imperfect data" (pp. 113-114). Besides these two shortcomings, they also point out a lack of empirical evidence to support the claims of Spirtes, Glymour, and Scheines (2000) about the efficacy of their methods.

Further criticism is centred on a reversed burden of proof. In practice, causal links will only be established when tests are sufficiently powerful and the data sufficiently informative. If there is insufficient information, then causality will not be assigned, which is the opposite of what most (social science) researchers would default to if in doubt. Instead, it is common to start with the assumption that most variables are causally intertwined and to reverse this judgement only if there is indeed clear evidence indicating the opposite. While taking a single insignificant correlation as indicative of non-causality is not too far fetched, problems emerge when this insignificant result serves as the basis for further causal reasoning. For example, it could happen that the direction of causality between two variables A and B is found by way of an insignificant

correlation between variables C and D. Thus, a rather strong claim is potentially supported by rather weak evidence.

A final, statistical criticism relates to the use of repeated hypothesis testing to uncover dependency patterns. When rejecting the null hypothesis of no correlation, the overall probability of making an erroneous decision is not bounded by the nominal significance level of the individual test. As is usual in multiple testing situations, the type I errors can accumulate and the overall probability of making at least one false rejection is not controlled for. The use of repeated testing is one reason why Kilian and Lütkepohl (2017, p. 235) argue that in the context of VAR analysis "this data-driven identification approach is not well suited for uncovering economically meaningful structures." A second reason for their scepticism is that structural errors need not correspond to specific variables in a VAR setting. They therefore view it as problematic to relate zero restrictions in the mapping between structural and reduced-form errors to causal links between variables. The methods have nonetheless been applied in empirical work, most prominently by Swanson and Granger (1997); Demiralp and Hoover (2003); Moneta (2008).

The other branch of the literature that seeks to validate or falsify an assumed causal structure proceeds as follows. An assumed structure gives rise to a graph and is validated by deriving a set of testable implications from that graph. It can then be verified whether these implications hold in the data. If they do not hold, the assumptions were successfully falsified. Similarly to causal discovery, the testable implications are provided by conditional dependence and independence relations that should hold in the data if the assumed causal structure is true. Causal graphs are useful in this regard because an investigator can easily inspect a graph visually and read the dependence relations off the graph using certain criteria (Pearl, 2009). Thus, the value added by causal validation is twofold. First, graphs lend themselves to easy visual interpretation and can summarise and simplify communication of key identifying assumptions. Second, they ease deduction of conditional dependence relations implied by those identifying assumptions and thus guide empirical verification of those assumptions.

Using causal graphs in such a confirmatory manner or as a guiding tool for the identification of causal effects is increasingly popular in the social, behavioural, and health sciences. Steiner et al. (2017) provide an overview of causal graphs in the context of treatment effect analysis and experimental and quasi-experimental research designs. These include randomised control trials (RCTs), regression discontinuity designs (RDDs), instrumental variables (IVs), and propensity score matching (PSM). They emphasise how practitioners can use graphical identification criteria to falsify

unbiased estimation of causal effects when adopting a particular research design, for example one of the four designs just mentioned.

Elwert (2013) applauds graphical causal models for their ability to translate statistical frameworks into more accessible graphs. As also mentioned above, researchers can more easily falsify identifiability of causal effects and derive testable implications of the hypothesised causal structure. The author offers an overview of graphical causal methods with a focus on identification of causal effects in quasi-experimental setups. He covers three different relationships between variables that give rise to an association between variables. These three relationships correspond to three basic graphical components that are the building blocks of causal graphs. Importantly, only one of these relationships is causal, while the other two imply spurious correlations. One important insight of this perspective is that "conditioning on variables [...] can induce as well as remove bias" (p. 245), which is not necessarily as often stressed in standard regression frameworks. Elwert (2013) further elaborates how a concept called d-separation formalises the analysis of whether a sequence of variables, called a path, transmits association (see Appendix 2.A.1). If a path transmits association, it is called 'unblocked', otherwise it is 'blocked.' Finally, graphical identification criteria exploit the insights of d-separation by establishing visual rules for determining identification of causal effects. Such criteria can also be applied to SVAR models, which is the focus of Section 2.5 and for which the next section will lay more groundwork.

2.2.3 Causal Graphs and Structural Vector Autoregressions

This section will discuss causal reasoning for the identification of structural VAR models. All papers in this section focus on SVAR models that are identified via short-run zero restrictions. These restrictions are always imposed in the context of an A-type model

$$\boldsymbol{A}_0\boldsymbol{y}_t = \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{A}_P\boldsymbol{y}_{t-P} + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

where $\boldsymbol{A}_0$ captures the contemporaneous effects between the K endogenous variables contained in the vector $\boldsymbol{y}_t$ and the error term $\boldsymbol{\varepsilon}_t$ is serially and contemporaneously uncorrelated. System (2.1) is underdetermined and the matrix $\boldsymbol{A}_0$ therefore not identified unless $K(K+1)/2$ restrictions are placed on its elements.[1] These restrictions are usually motivated by economic theory (for an overview see Kilian and Lütkepohl, 2017), but there have also been attempts to motivate short-run zero restrictions empirically through causal reasoning. Among the first to do so were Swanson and Granger (1997) who aim at reducing the "subjective nature of error orthogonalization in the

[1] This condition is necessary, but not sufficient. See Lütkepohl (2005) for details.

Figure 2.1: An example of a Swanson and Granger (1997) chain graph where each variable has a contemporaneous causal effect on at most one other variable without feedback loops.

VAR methodology" (p. 364). Their ad hoc search procedure for viable restrictions relies on the fact that imposing a particular set of over-identifying restrictions implies specific conditional correlation patterns among the reduced-form residuals of a VAR. Most importantly, some of these conditional correlations will be zero. Making use of these patterns, they suggest forming a baseline model which will exhibit "at least some features which are in accord with the data" (p. 362). As a second step, the over-identifying restrictions are tested via t-tests. The baseline model they propose is rather restrictive. It only considers chains of variables where one variable is able to affect at most one other variable contemporaneously (see Figure 2.1). Furthermore, only recursive models without feedback loops between variables are allowed for. The upside is that, under these conditions, Swanson and Granger (1997) prove the consistency of the least squares estimator and the asymptotic validity of t-tests used in the search procedure.

The logic implied by patterns of partial correlations can be extended, however, to models beyond simple unidirectional chains. For models where each variable may affect multiple other variables contemporaneously, the patterns of zero partial correlations become more involved. It turns out that graphs are a convenient tool for finding out which zero partial correlation patterns are implied by a given SVAR identification scheme. Under the assumption of normally distributed errors or linearity, zero correlation is sufficient for independence and the results of the causal graph literature apply. That is, one may test which partial correlations are indeed absent in the data and whether they would thus corroborate the assumed structure on $\boldsymbol{A}_0$. Alternatively, one may reverse this line of argument and ask what kind of identification schemes would be in agreement with an observed correlation pattern among the reduced-form residuals, with the caveats mentioned above. The latter would constitute a graphical search procedure. These graph-theoretical search procedures are part of the literature on machine learning and artificial intelligence and are described in Spirtes, Glymour, and Scheines (2000), Pearl (2009), Koller and Friedman (2009), and Peters, Janzing, and Schölkopf (2017). The algorithm applied by most of the cited literature below uses

the PC-1 algorithm outlined in Spirtes, Glymour, and Scheines (2000, pp. 84–85)[2], or variants thereof.

In the context of SVAR identification, Demiralp and Hoover (2003) assess the efficacy of the PC-1 graphical search procedure through evidence from Monte Carlo simulations. They find that the success of the algorithm in recovering a model that nests the true structure depends heavily on the quality of the information contained in the data. As might be expected, for low signal-to-noise ratios, the algorithm will impose false restrictions quite frequently. With sufficient signal strength, on the other hand, the algorithm testing for patterns of partial correlations will recover the true (over-identified and recursive) SVAR model in up to 80 % of cases and a model nesting the true model in up to 97 % of cases.

Thus, while the graph-theoretical approach can recover the correct structural model successfully in specific circumstances, the downside to this methodology is that it fails to indicate when these circumstances apply. It may, for example, suffer from low power and impose a fairly arbitrary structure simply out of the inability to reject a certain set of null hypotheses. Furthermore, there is no account of the uncertainty that a set of hypotheses was wrongly accepted. In practice, it is therefore difficult to conclude whether a graph-theoretical algorithm successfully recovered the correct structure or whether the results are driven by uninformative data.

Demiralp, Hoover, and Perez (2008) address this issue of uncertainty by devising a bootstrap method that may indicate the reliability of the algorithm's outcome. The method could thus serve as an "effective tool for assessing our confidence in causal orders identified by graph-theoretic search algorithms" (p. 509). The bootstrap is based on first resampling reduced-form residuals, feeding these residuals to the original model to create new time series data, re-estimating based on the new bootstrap data and finally applying the graphical search procedure to recover another set of $\boldsymbol{A}_0$ matrices. Demiralp, Hoover, and Perez (2008) evaluate how often certain features of the original structure emerge during the bootstrap. They do not provide any formal justification for why the original structure would be expected to re-emerge as part of the bootstrap, but instead speculate that if the data is just noisy, the structure would change rather frequently, whereas very informative data would show a stable pattern across bootstrap samples. Demiralp, Hoover, and Perez (2008) compare the summary statistics of the bootstrap method to Monte Carlo simulations and take similar rejection frequencies as

[2]The name 'PC' apparently derives from the initials of the authors **P**eter Spirtes and **C**lark Glymour. This naming convention is not standard throughout the literature, however. For sake of simplicity, I will follow this partially adopted convention. In addition, since there are various incarnations of the algorithm and to avoid confusion with **p**rincipal **c**omponent analysis, I adopt the convention to call the specific algorithm outlined by Spirtes, Glymour, and Scheines (2000, pp. 84–85) the PC-1 algorithm.

evidence that the bootstrap procedure mimics the Monte Carlo sufficiently well that it can "provide useful guidance on the reliability of inference" (p. 528). However, the authors note a slight misalignment between either the type I error rate or the type II error rate and choose to match the type I error rate by using different significance levels during the Monte Carlo simulation and the bootstrap procedure.

Two things remain unclear. First, since the procedure is purely heuristic, the degree to which the results generalise to other settings remains uncertain. Second, it is unclear whether the p-values of the individual hypothesis tests would already have signalled the reliability of inference sufficiently well. Arguably, the test decisions that will shift frequently will be those with p-values close to the nominal size of the test. On the other hand, the fact that the residuals are estimated quantities might be better captured by the bootstrap. In addition, an erroneous test decision can have ramifications throughout the graph, beyond the single decision at hand. This effect is neglected when just considering p-values.

There is an important difference between the work of Swanson and Granger (1997) and Demiralp and Hoover (2003). The former exclusively consider t-tests that are asymptotically valid if the contemporaneous structure they assume is true. Demiralp and Hoover (2003) do not form an a priori assumption. Hence, during the graph-theoretic search procedure all kinds of hypothesis tests are performed, some of which will suffer from endogeneity bias inherent in the search. In finite samples this may lead to problematic conclusions and will be further discussed in Section 2.4. Notwithstanding the fact that some hypothesis tests are affected by bias, the graphical search procedure has been applied by a number of researchers in macroeconomics and finance. It is not part, however, of the standard macroeconomic tool set.

Kwon and Bessler (2011) review the fundamentals of the graph-theoretical approach to causal inference in a macroeconometric context. They also review some of the applied literature and stress the assumptions—foremost the Markov and stability condition—that are required to infer a causal structure from empirical regularities. They establish context as regards previous approaches to causality in macroeconomics such as Granger causality and structural equation modelling.

Noteworthy studies that have applied graphical modelling include Hoover, Demiralp, and Perez (2009), Demiralp, Hoover, and Perez (2014), Fragetta and Melina (2011, 2013), Moneta (2008) and Jinjarak and Sheffrin (2011). Hoover, Demiralp, and Perez (2009) study the role of monetary aggregates in the transmission of monetary policy shocks to output and inflation. Their motivation lies with the Federal Reserve judging conditions for aggregate demand through measures of liquidity. However, the authors claim that little is known about the interplay of monetary aggregates with

interest rates, inflation, output and stock markets, apart from the quantity theory of money that the Fed relies on. Hence, they apply the PC-1 algorithm to a VAR model of 11 monthly variables, including core inflation, industrial production, the federal funds rate, liquid deposits and various measures of money market and stock market conditions. The algorithm returns a set of overidentified SVAR models, which the authors evaluate further via the bootstrap of Demiralp, Hoover, and Perez (2008). Having settled on a contemporaneous structure, they account for non-stationarity of the data by transforming the model to a vector error correction model (VECM) and further restrict the specification by applying a general-to-specific search on its lag structure.

From this rather data-driven specification search they conclude that the quantity-theoretic approach of the Fed is rejected by the data. While liquidity deposits have a delayed effect on industrial production and core inflation, its role in the transmission of monetary policy to these two variables is "almost immeasurably small." I will add that the model is also at odds with conventional wisdom on monetary policy: the federal funds rate does not react contemporaneously to any other variable in the system. Its response to inflation is only indirect through other variables and delayed by at least one month. Furthermore, there is no direct or indirect contemporaneous effect at all between the federal funds rate, industrial production and core inflation. Indeed, core inflation does not interact with any other variable contemporaneously. While the data do not reject the final model specification, it would be worthwhile to study the robustness of their final conclusions by contrasting their findings with more conventional and less restrictive specifications.

Demiralp, Hoover, and Perez (2014) perform a similar study. Focusing on the price puzzle in monetary policy, they empirically identify a SVAR model via the PC-1 algorithm. The data covers 12 monthly US time series ranging from February 1959 to June 2007 and include consumer and commodity prices, industrial production, a measure for the output gap, and various interest rates and monetary aggregates. The empirical strategy is analogous to Hoover, Demiralp, and Perez (2009), evaluating the outcome of the algorithm via the bootstrap and cutting down on the number of lag coefficients via a general-to-specific testing strategy. The significance level used in the PC-1 algorithm is again set at 10 % and the final model specification cannot be rejected by a likelihood ratio test at the 10 % significance level either. With 53 out of the 66 possible parameters set to zero, the $\boldsymbol{A}_0$ matrix is highly restricted.

With this specification, the authors evaluate whether the inclusion of certain variables mitigates the price puzzle. They split the sample in a pre- and post-Volcker period and find that results differ between the two samples. Noticeably, for the later period starting in 1990, the study concludes that *excluding* commodity prices or output

gap measures mitigates the appearance of a price puzzle, yet overall the puzzle remains unresolved. Here again, it would have been interesting to assess which features of the model drive the results and how conventional specifications compare to the empirically motivated model.

The idea of comparing competing models through the lens of partial correlations is picked up by Fragetta and Melina (2011) and Fragetta and Melina (2013). Both papers rely on conditional independence graphs (CIGs) instead of directed acyclical graphs (DAGs) and therefore on a slightly different variant of a causal search algorithm, but the underlying principles remain the same. Fragetta and Melina (2013) study different recursive identification schemes in the context of monetary policy SVAR models. Authors such as Christiano, Eichenbaum, and Evans (2005) argue that the inertia of macroeconomic developments warrants the restriction of zero immediate impact of monetary policy on output, wages, productivity and the general price level. Monetary aggregates and financial variables, on the other hand, are free to react on impact to a monetary impulse. Christiano, Eichenbaum, and Evans (2005) also assume that the central bank observes and considers contemporaneous macroeconomic developments, except for monetary and financial variables. Sims and Zha (2006) doubt the credibility of the last assumption. Instead, they propose that financial markets are simultaneously observed by the central bank and free to immediately respond to policy. Current values of prices and output, on the other hand, are usually only known with a delay and are therefore not part of the central bank's information set when deciding on policy. Fragetta and Melina (2013) pitch these two competing assumptions regarding the central bank's information set against each other. They find that only high-frequency data, such as commodity prices, enter the central bank's information set in a VAR model identified by empirical means. It should be noted, though, that the study's data set covers just four variables—output, the federal funds rate, general prices, and commodity prices—whereas the data used in the original studies cover broader sets of eight to nine variables. It is unclear whether the conclusions also hold in larger sized VAR models. What is more, Sims and Zha (2006) explicitly allow for simultaneous effects between financial markets and the monetary authority within the quarter. This cannot be reflected in the empirically identified model since the search algorithm precludes the existence of feedback loops.

Fragetta and Melina (2011) carry out a similar exercise for fiscal policy. Here, discussions revolve around the effect of a discretionary, deficit-financed fiscal spending rise on investment, consumption, hours worked and wages. Classical models predict that agents foresee and offset future tax rises by decreasing current consumption and increasing labour supply, which is associated with a fall in wages. From a Keynesian

perspective, habit as well as rigidities in wage and price setting lead to an increase in consumption and therefore a more pronounced response of output to a fiscal spending shock. Fragetta and Melina (2011) identify a fiscal policy VAR empirically and find that a rise in spending affects output, consumption, hours worked and real wages positively, while investment falls.

Moneta (2008) revisits the debate on the origins of business cycle fluctuations. The study finds that fluctuations in output, consumption and investment are driven by disturbances to monetary and not just real variables in an empirically identified VAR model. The paper makes two further contributions. First, the methodology accounts for the fact that the residuals used for estimating partial correlations are estimated quantities. Second, the author argues that the results of the PC-1 algorithm will be more robust against type II errors if a more complete set of partial correlations is tested. These modifications may lead to fewer orientated edges in the graph, necessitating a greater reliance on theory.

Jinjarak and Sheffrin (2011) employ the PC-1 algorithm in the context of the balance of payments and real estate markets. They find, through the lens of an empirically identified VAR, that capital account surpluses in the US can impact real estate prices indirectly by lowering mortgage interest rates. They do not find evidence for a pull effect (bullish real estate markets driving up consumption and drawing in capital) and only weak evidence for a direct push effect (capital inflows push up real estate prices).

Bryant, Bessler, and Haigh (2006) exercise a form of causal validation by comparing competing hypothesis debated in the literature on future markets. They draw on work published as Bryant, Bessler, and Haigh (2009), where they demonstrate how causal hypotheses can be disproved through inspection of specific patterns of partial correlations. They argue that focusing on a specific causal hypothesis (instead of learning the complete causal structure among multiple variables) requires less stringent assumptions. They do not, however, avoid the problem of a reversed burden of proof. A rejection of a causal hypothesis still requires acceptance of some null hypotheses and may therefore be adversely affected by low power. But they do differentiate between a weak and strong basis of rejection, arguing that more complex patterns are unlikely to emerge by chance and therefore represent a stronger basis than simple patterns. Bryant, Bessler, and Haigh (2006) use this framework for testing whether hedgers pay a risk premium on commodity and currency future markets and whether price volatility is (in part) driven by uninformed traders. They reject both of these hypotheses for most of the markets they study.

Finally, Moneta et al. (2011) relax assumptions of normally distributed errors and system linearity. For non-normality, they discuss the use of independent component

analysis. For non-linearity, they rely on nonparametric kernel estimators. With these they are able to test for (conditional) independence by checking whether the equality between joint densities and the product of marginal densities holds. The authors conduct a small simulation study where they compare the use of standard Fisher z-statistics (which assume linearity) to nonparametric tests. In case of non-linear DGPs, a particular test based on the Euclidean distance measure for densities seems to perform well with respect to size and power. It is also reliable and comparable to the z-statistic in the linear case. However, nonparametrics suffers from the curse of dimensionality and allows only a small conditioning set. Some properties of the DGP are also not quite clear, for example its signal-to-noise ratio.

2.3 Methods for Causal Discovery

Probabilistic graph theory equates vertices in a graph with random variables. The relation between two vertices is represented by an edge. In causal graph theory, the focus lies on directed graphs for which the relations between vertices are interpreted as being causal. The edges in a directed graph possess arrows and therefore a direction. The vertices being pointed at from a particular vertex, say vertex A, are called the children of A. Analogously, the vertices pointing towards A are called the parents of A. These sets can be extended to descendants and ancestors to include the children of children and parents of parents and so forth. While in principal the causal direction between two vertices can run both ways, the methods discussed in this section are restricted to directed acyclical graphs (DAGs), which have the defining characteristic that vertices cannot be their own ancestors.[3] This reflects an assumption that causation is asymmetric and necessarily one-directional. It implies that edges may only point in one direction, otherwise two vertices would be parent and child to each other at the same time. In other words, the system is restricted to be recursive.

The goal of the methods discussed in this section is the discovery of causal relations among contemporaneous variables purely from the data itself. Such an undertaking is fraught with difficulties. The very fact that the system is endogenous in the first place implies the possibility of an endogeneity bias in any ad hoc specification that is not guided by theory. Nonetheless, the methods presented here suggest that it may be possible to discover clues about specific, local aspects of a system under a suitable set of assumptions. These assumptions will be discussed below. Retrieving causal relations from observational data could be advantageous whenever (quasi) controlled experiments are not available, too costly or unethical. In macroeconomics that is

[3]Sometimes, ancestry is defined to be reflexive. In that case every vertex is its own ancestor and descendant, but still no loops are allowed in the graph.

usually the case and the methods for causal discovery may therefore be especially helpful.

In the literature on causal graphs, there are three particular types of methods for inferring (aspects of) a graph from data. The first type is the 'constraint-based' approach, the second the 'score-based', and the third a Bayesian approach. The constraint-based approach exploits the fact that specific types of subgraphs, and therefore specific causal constellations, correspond to stochastic independence relations between variables. Methods with this kind of approach systematically test and impose local independence constraints, and thereby reach conclusions about the graph (see Verma and Pearl, 1990; Spirtes, Glymour, and Scheines, 2000). The second class of methods explores the space of graphs and assigns a score to each structure it encounters, such that the highest scoring graph will serve as an estimate of the true structure (e.g. Chickering, 2002). Standard information criteria such as AIC or BIC can serve as a scoring function. Finally, a Bayesian approach with a prior over the space of DAGs can equally explore graphs and assign posterior weights (see Madigan and Raftery, 1994). An important feature of all three approaches is the fact that they identify equivalence classes of observationally equivalent graphs. This reflects the fact that, even with the appropriate assumptions on the data, oftentimes it is only possible to uniquely identify part of the graph structure. Graphs are especially useful in this respect, because they can easily be characterised as part of a certain equivalence class, and thus guide search procedures in effectively exploring the space of observationally distinct structures, which leads to a reduction of the search space. In the more recent literature there are also a number of hybrid approaches and further extensions of these three approaches.

These search endeavours, in particular the constraint-based approach, are close in spirit to attempts at automating model specification in econometrics (e.g. Hendry and Krolzig, 2004, 2005) and share some of the same difficulties encountered there. Both these approaches conduct sequences of hypothesis tests and need to trade off size vs. power, especially since the *acceptance* of some null hypotheses has a major effect on the outcome. Furthermore, the procedures are generally only point-wise consistent (Robins et al., 2003; Leeb and Pötscher, 2005) and it is therefore hard to impossible to consistently estimate moments of the post-selection distribution of parameters as well as the overall sampling uncertainty of the estimated structural model.

We will now discuss the assumptions underlying causal discovery methods. An important assumption in working with causal graphs is the causal Markov condition. It states that, conditional on its parents, a variable is statistically independent of any other variable except of its descendants and parents. In many applications, this seems

a reasonable assumption to make. Furthermore, it allows to construct convenient factorisations of joint probability distributions by simply inspecting a graph. Take vertices A, B, C. If there are no edges connecting these vertices in the true graph, then they are quite evidently independent and $P(A, B, C) = P(A)P(B)P(C)$. But suppose the graph consists of a single path $A \to B \to C$ such that A is ancestor of B which is a parent of C, then by the Markov condition one valid factorisation is $P(A, B, C) = P(A)P(B|A)P(C|B)$. However, another valid factorisation would be $P(A, B, C) = P(C)P(B|C)P(A|B)$, which already hints at the fact that some graphs are observationally equivalent to each other.

The Markov condition is key for deducing testable implications from a given graph, as well as for the implementation of algorithms that inductively infer graphs from observational data. However, the causal Markov condition only establishes a mapping from graphs to classes of probability distributions. It implies that 'no link between vertices' is a sufficient condition for conditional independence of two variables. What is needed in order to have any hope of uncovering properties of a causal graph from the data is to make 'no link' a necessary condition as well. If it were not a necessary condition, we might observe that two variables seem completely unrelated, yet we could not conclude that these two variables do not cause each other either directly or indirectly. To preclude these cases, the 'faithfulness' or 'stability' condition is regularly assumed. It states that no other conditional independence relations hold other than those implied by the causal Markov condition. Amongst other things, this axiom effectively excludes probability distributions in which variables depend on each other in such a way that the overall effect between two dependent variables cancels out, and they thus appear independent even in population.

The faithfulness assumption may seem innocuous at first. A number of commentators, for example, have observed that the distributions which contain cancelling effects, and are therefore unfaithful, are of Lebesgue measure zero for continuously distributed variables (e.g Robins and Wasserman, 1999). They may therefore seem little more than a pathology. However, those distributions for which independence constraints hold are equally of Lebesgue measure zero. Furthermore, in finite samples, the share of distributions for which effects appear to 'cancel out' simply due to sampling uncertainty will very likely be a non-trivial proportion. This phenomenon is illustrated in more depth in Section 2.4. These observations have led to the definition of 'strong faithfulness,' which requires that the true probability distribution is sufficiently far away from 'pathological', unfaithful cases. Under such an assumption it is even possible to prove uniform consistency of causal discovery methods (Zhang and Spirtes, 2003). See Uhler et al. (2013) for a more elaborate discussion of strong faithfulness.

Another frequently invoked assumption is 'causal sufficiency', which is equally often assumed in the econometric literature under the term 'unconfoundedness' or 'conditional independence assumption.' It states that all variables affecting more than one of the variables under investigation are observed; i.e. there are no unobserved confounders. This assumption is sometimes relaxed by explicitly allowing for latent variables when searching for valid causal graphs. For SVAR analysis, the absence of unobserved confounders may facilitate identification, but it is not a necessary condition. B-type SVAR models, for example, explicitly account for a set of latent variables affecting more than one endogenous variable.

In the structural VAR literature, almost all applications using causal graph procedures have focused on one particular constraint-based method which will now be elaborated.

PC-1 Algorithm

The PC-1 algorithm conducts a data-dependent sequence of independence tests. With Gaussian data or linear systems, these tests are typically partial correlation tests. Depending on the correlation pattern found, it adds edges between variables and orientates their direction. To provide an illustration of how correlation patterns are informative about the causal structure, suppose again there are three variables A, B, and C. If $\text{corr}(A, B) \neq 0$, $\text{corr}(B, C) \neq 0$ and $\text{corr}(A, C) \neq 0$, but $\text{corr}(A, C|B) = 0$, then edges are added between variables A and B, and B and C. There will not be an edge between A and C because of the causal Markov and faithfulness conditions: Taking B into account seems to eliminate the effect of A on C (or vice versa) and therefore suggests that neither A nor C is parent to the other. The direction of causality remains unresolved, however. This first example is shown on the left of Figure 2.2. On the right there is a second example that shows the case when $\text{corr}(A, B) \neq 0$, $\text{corr}(C, B) \neq 0$, $\text{corr}(A, C|B) \neq 0$, but $\text{corr}(A, C) = 0$. The only way to rationalise this finding is to have causation flow from A to B and from C to B since otherwise we would observe $\text{corr}(A, C) \neq 0$ as well. Here, the faithfulness condition is again relevant to preclude cases in which the correlation between A and C cancels out due to other—possibly unmeasured—effects.

Constraint-based discovery algorithms, of which the PC-1 algorithm is a prime example, operate by the above logic. Important differences between algorithms depend on whether the true graph is assumed to be acyclic and whether unconfoundedness is assumed to hold. The PC-1 algorithm assumes both and proceeds as follows. First, it explores all unconditional independencies, followed by conditioning on more and more variables to see under which circumstances an effect vanishes. In case an effect does

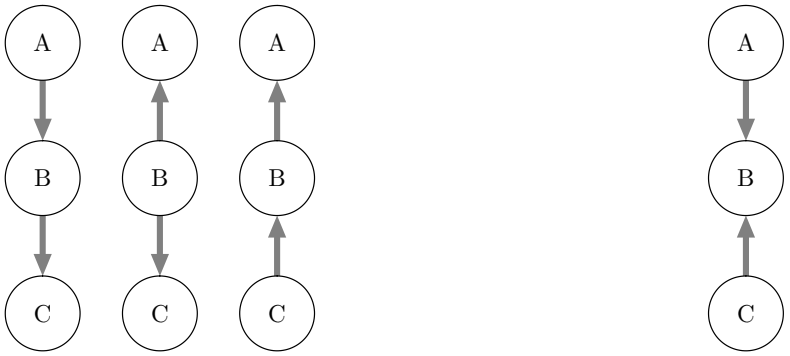

Figure 2.2: The left three causal graphs all imply the same correlation patterns. The right graph corresponds to a unique pattern.

become statistically insignificant once a certain conditioning set is formed, the corresponding edge is recorded as absent. It further relies on the fact that the independence pattern leading to a (sub)graph as on the right of Figure 2.2 is unique. Therefore, if such a pattern is found and if A and C share no edge, then the edges can be orientated as shown in that figure.

The output of the algorithm at this stage will in general be a graph with some directed as well as undirected edges. Such a graph is called a partially directed acyclic graph (PDAG). Since acyclicity was one of the main assumptions from the outset, the PDAG actually reflects an equivalence class of DAGs where each member fulfils those independence relations implied by the PDAG. It is often possible to identify edges that are orientated the same way in every single member of the equivalence class. In that case the PC-1 algorithm will orientate those edges in the PDAG as well. The outcome of this last step is a so-called completed partially directed acyclic graph (CPDAG) and the final output of the algorithm.

Further important differences between implementations of this kind of algorithm exist. For one, differences emerge depending on how the conditioning set in testing for conditional independence is formed. When testing the independence of A and B, it would in theory be sufficient to consider only those variables in the conditioning set which are adjacent to either A or B. However, in practice, due to sampling uncertainty or violation of one of the assumptions, the results could differ when a broader set of conditioning variables is allowed for. In addition, there could be contradictory patterns that cannot be resolved within the framework of causal DAGs. Thus, how conditioning sets are formed and how contradictory results are dealt with play an important role in determining the outcome, as well as more general properties of the algorithm such as speed, error rate, and order dependence.

More detailed expositions of the PC-1 algorithm and discussions of its properties can be found in Spirtes, Glymour, and Scheines (2000); Koller and Friedman (2009);

Colombo and Maathuis (2014), amongst others. Due to the ability of constraint-based algorithms to infer structure by examining only local properties, it is well suited for being employed in sparse, high-dimensional settings (Kalisch and Bühlmann, 2007) and has gained some popularity there.

A common theme among algorithms that construct causal graphs from observational data is that they expect independence relations as input. In other words, they are silent about how to assert independence in the first place. These test decisions are crucial, however, for the success of causal graph algorithms in finding the truth. There are several options for testing independence. The simplest assume a parametric framework. They are easy to implement and come at a low computational cost. Distributional assumptions have to be made, though, usually assuming normally distributed variables. Non-parametric tests of independence, on the other hand, either estimate the joint and marginal densities of the variables involved or rely on resampling schemes such as permutation tests (e.g. DiCiccio and Romano, 2017). These tests are therefore less restrictive as regards distributional assumptions. The downside is the usual curse of dimensionality, leading to less precise estimation in larger systems or with few observations.

Among parametric tests for conditional independence, Fisher's z is frequently used. Also, most structural VAR applications such as Demiralp and Hoover (2003), Hoover, Demiralp, and Perez (2009) rely on Fisher's z to test whether sample correlation coefficients are significantly different from zero. This test statistic is also implemented in popular statistical software packages, like *pcalg* for the R language (Kalisch et al., 2012). Fisher's z uses the inverse hyperbolic tangent transformation to map correlation coefficients to the real number line. Let $\hat{\rho}$ be a sample partial correlation coefficient between A and B conditional on a collection of random variables $\mathcal{Z}$. Then Fisher's z is given by

$$z = \frac{1}{2} \log \frac{1 + \hat{\rho}}{1 - \hat{\rho}} \,. \tag{2.2}$$

Let ζ be the population counterpart to z. By Theorem 4.2.5 of Anderson (2003), we have $\sqrt{T}(z - \zeta) \overset{a}{\sim} \mathrm{N}(0, 1)$, where T is the sample size. Suitable hypothesis tests of $\mathrm{H}_0 : \rho = 0$ can be constructed using this asymptotic approximation. Alternatively, and asymptotically equivalently, one may also test $\rho = 0$ by standard t-tests in a regression framework where A is regressed on B and the conditioning variables have been added as additional regressors.

A concern with any testing procedure is the likelihood of actually recovering, or at least nesting, the true causal structure. Expressing this likelihood is unfortunately rather difficult. The previously mentioned difficulties with regards to post-model-selection inference is one aspect. A related aspect is the mix of rejected and non-

rejected hypotheses that give rise to a particular graph estimate. Both type I errors and type II errors will affect the estimate unfavourably and may reverberate through the graph beyond the locally tested property. This adds considerable complexity to evaluating the effect of a false decision. A third issue is the use of multiple hypothesis testing. This issue may be the easiest to remedy by relying on Bonferroni bounds or techniques that address the family-wise error rate or false discovery ratio.

The next section will illustrate how graph estimates can be adversely affected by the above issues if 'strong faithfulness' does not hold.

2.4 Pitfalls in Causal Discovery

This section illustrates a particular weakness of graphical causal algorithms that has not always been made explicit in the literature so far. One feature of the algorithms discussed in this paper is their reliance on a sequence of hypothesis tests. This sequence implies, amongst other things, a multiple testing problem that has been addressed elsewhere. Apart from multiple testing, the sequential approach becomes problematic when combined with another feature of these algorithms: placing a lot of structure under the null hypothesis in order to inductively infer causal relations. That is, they routinely test for effects $\beta_{AC|Z}$ between any two variables A and C conditional on the set of variables Z by setting up the pair of hypotheses

$$H_0 : \beta_{AC|Z} = 0, \tag{2.3}$$

$$H_1 : \beta_{AC|Z} \neq 0. \tag{2.4}$$

If the null is not rejected, the algorithms will *accept* the fact that $\beta_{AC|Z} = 0$ and decisions based on tests further down the sequence of tests will be conditional on this acceptance. But strictly speaking, a non-rejection is simply a failure to falsify the null hypothesis and not an acceptance of the null. The test may have just suffered from low power due to the nature of the test or due to insufficiently informative data. Take, for example, the (rather typical) case where the data is noisy and a researcher wishes to investigate causal relations. Without placing much structure on the data, the researcher observes that most variables are significantly correlated with each other, while two variables A and C are not. Even though this does not directly provide evidence that the two variables are unrelated, he or she concludes that any meaningful direct causation between those two variables is unlikely. While this conclusion may suggest itself quite naturally, it has further ramifications on the causal relations between other variables than those two. For instance, if it were the case that the correlation between A and C is no longer insignificant when conditioning on the set Z, logical

conclusions can be drawn about relations between members of Z and A or C. However, what is no longer considered at this point by the usual graphical search algorithms is the likelihood that the very first conclusion, the absent effect between A and C, was false. Ideally, one would have to account for the probability of type II errors in previous test decisions. Unfortunately that probability is unknown in practice. Accepting the null as given is exacerbated by the numerous hypothesis testing conducted by the algorithm. Thus, the researcher is not only confronted with an accumulation of type I errors, but also of type II errors.

A related criticism is that in some disciplines, and especially in macroeconomics, the default presumption is that all observables are contemporaneously causally related. Philosophically one may doubt whether two processes can causally impact each other truly instantaneously, but with macroeconomic quantities typically observed at monthly or quarterly frequency and aggregated over multiple agents, such contemporaneous effects are more plausible. Thus, if the data is noisy or there is insufficient evidence to identify a significant effect, investigators may want to exercise caution before proclaiming the possibly very far-reaching conclusion that two variables are not causing one another.

A related practice is the use of hypothesis testing for variable selection or the interpretation of regression results. Here, a non-rejection is often taken as an acceptance and some researchers act on the premise that the null hypothesis is in fact true, for example by excluding variables which have no significant effect. However, it is also recognised that pruning models in this way will invalidate standard inference on the parameters that remain (Leeb and Pötscher, 2005). A common feature with causal search algorithms is to condition each decision on previous results, thus failing to account for the full degree of uncertainty about the final result. Again, the criticism here is not so much about the acceptance of a single null hypothesis, but the repeated acceptance of null hypotheses and their joint ramifications for the search algorithm's final output without any quantification of uncertainty arising from false negatives.

A simple example will illustrate the kind of ramifications a false acceptance of a null hypothesis may have. Take a static stochastic system governed by three structural equations

$$A = \varepsilon_a \tag{2.5}$$

$$B = \beta_1 A + \varepsilon_b \tag{2.6}$$

$$C = \beta_2 B + (\phi - \beta_1\beta_2)A + \varepsilon_c, \tag{2.7}$$

where $\varepsilon_k \sim \mathrm{N}(0, \sigma_\varepsilon)$, for $k = a, b, c$, are three independently and normally distributed unobserved innovations. The three random variables A, B, C are observable. Let any set of observations be denoted as $\{a_t, b_t, c_t\}_{t=1,\dots,T}$ with T being the sample size. These equations are structural because they describe the very process of how values of A, B, and C are determined. If one were to experiment and intervene, say, on variable B in (2.7) by manipulating its value, one could change the expectation of C without affecting the remaining system. In other words, the innovations ε_k are strictly exogenous and one can interpret the coefficients as causal effects.

In this setting, the causal ordering is clear. Variable A is a cause of B, and A and B are in turn causes of C. In practice, this ordering will not be known ex-ante, such that a causal search algorithm may be applied. Of central importance for the success of the search algorithm is the parameter ϕ. This parameter will be manipulated in the simulation exercise to two effects. First, as any parameter, it will influence the share of explainable variation in C and thus the prediction accuracy when forming an expectation of C conditional on A and B. In a regression setting, ϕ will influence the estimation precision for any given sample size and the power of standard Wald tests to reject a wrong null hypothesis. Second, keeping β_1 and β_2 constant, ϕ controls the magnitude of the omitted variable bias in the regression

$$C = \phi A + \beta_2 \varepsilon_b + \varepsilon_c \tag{2.8}$$

$$= \beta_3 A + \tilde{\varepsilon}_c. \tag{2.9}$$

Equation (2.8) is easily obtained by substituting (2.6) into (2.7). Estimating regression (2.9) with least squares will yield $\mathrm{E}[\hat{\beta}_3|A] = \phi$ as there is an omitted variable bias of size $\beta_1\beta_2$. Parameter ϕ may thus be used to control the estimation precision in regression (2.9) and thus the likelihood of rejecting the hypothesis $H_0\colon \beta_3 = 0$. For values of ϕ close to zero, the two variables C and A will seem unconditionally unrelated. Conditioning on B, however, a regression following (2.7) may be more likely to indicate a relation of C to A and B, depending on the values of β_1 and β_2. Note that ϕ must not be equal to 0 as this would violate the faithfulness assumption necessary for the application of causal graph algorithms. Faithfulness states that the graph can in principle be recovered. It is thus a kind of identifiability assumption.

Certain parametrisations of ϕ, β_1, and β_2 will fool standard causal graph algorithms into orientating edges in reverse order with high probability. This phenomenon is illustrated in the following Monte Carlo simulation. In the simulation, data is created according to the process outlined in (2.5) to (2.7) and fed to the PC-1 algorithm. The DGPs are specified as $\beta_1 = \beta_2 = 2, \sigma_\varepsilon = 1, \phi = 0.1, 0.25, 0.5$ and

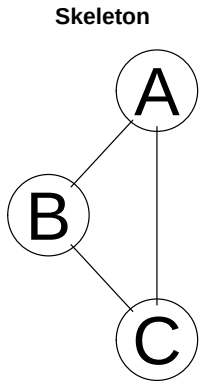

Figure 2.3

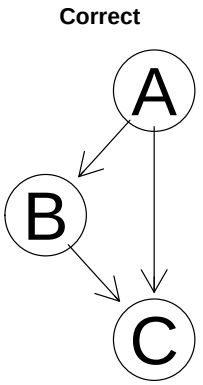

Figure 2.4

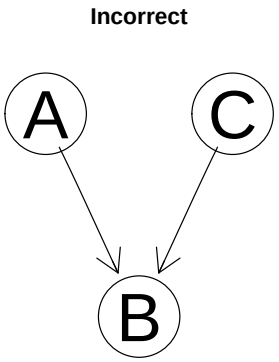

Figure 2.5

$T = 100, 250, 500, 1000, 2000$. Each DGP is repeated $R = 10000$ times such that R outcomes of the PC-1 algorithm are available per parameter constellation.

The results focus on two particular outcomes shown in Figure 2.3 and 2.5. Figure 2.3 displays the skeleton of the structure (2.5)–(2.7). The skeleton contains precisely those edges present in the true graph, but all edges lack orientation. This orientation is displayed in the correct graph in Figure 2.4. Note that the correct graph in Figure 2.4 is nested in the skeleton correct graph in Figure 2.3. The latter is therefore not wrong but overspecified in the sense that the skeleton does not impose erroneous restrictions on (2.5)–(2.7), but it is not the most parsimonious specification either. Figure 2.5, on the other hand, displays an incorrect graph revealing a reversed edge between node B and C and an omitted edge between node A and C. The two graphs in Figures 2.3 and 2.5 are the two dominant outcomes obtained from applying the PC-1 causal search procedure to the simulated data. The graph in Figure 2.4 cannot be recovered by the algorithm because it does not imply any (un)conditional independencies. Thus, in the best case, the algorithm will set-identify the correct equivalence class of graphs. In this case, the skeleton in Figure 2.3 is the only correct outcome of the algorithm as it represents the CPDAG associated with the class of Markov equivalent graphs containing the correct graph.

Table 2.1 presents the results of the simulation exercise, with the relative number of occurrences of correct and incorrect outcomes being tabulated. As just discussed, the correct outcome of the PC-1 algorithm is the skeleton correct graph. The frequencies at which the algorithm estimates this graph are tabulated in the upper panel of Table 2.1. The three rows correspond to different values of ϕ, as indicated in the first column. The remaining columns list the frequencies for different sample sizes. The probability of recovering a graph which nests the true graph increases both with ϕ and the sample size. For $\phi = 0.5$ and 2000 observations, this probability is approximately 1. The

Table 2.1: Simulation results in relative frequencies.

	Sample size T				
ϕ	100	250	500	1000	2000
Skeleton Correct					
0.10	0.25	0.31	0.39	0.55	0.76
0.25	0.45	0.68	0.89	0.99	1.00
0.50	0.82	0.99	1.00	1.00	1.00
Incorrect					
0.10	0.75	0.69	0.61	0.45	0.24
0.25	0.55	0.32	0.11	0.01	0.00
0.50	0.18	0.01	0.00	0.00	0.00

The first column indicates the value of ϕ, all other columns contain the simulation results for different sample sizes. The upper panel indicates the relative frequency at which the estimated graph is equal to the skeleton correct graph in Figure 2.3. The lower panel indicates the frequency at which the incorrect graph in Figure 2.5 is estimated.

picture changes, however, for lower values of ϕ. When $\phi = 0.1$ and when there are 500 observations, the probability of recovering the correct outcome is roughly half. As shown in the lower panel, the other 54 % of the time, the incorrect graph in Figure 2.5 is estimated. The results in the lower panel exactly mirror those of the upper panel. Thus, the skeleton correct graph in Figure 2.3 and the incorrect graph in Figure 2.5 are the exclusive outcomes of the PC-1 algorithm in this simulation exercise.

I will add three remarks. First, the frequencies tabulated in Table 2.1 may be interpreted as lower bound probability estimates which are valid for certain intervals of ϕ and T. Based on symmetry and monotonicity, the probability of choosing the incorrect graph is at least 32 % if ϕ lies in the interval $[-0.25, 0.25]$ and the sample size is 250 or less. Second, it is natural to expect that the algorithm will fail to correctly identify very weak effects for a certain fraction of cases. After all, if ϕ is close to zero and observations are scarce, precise estimation and inference on whether the parameter is different from zero will be difficult for any procedure. But what is more problematic is the complete lack of an uncertainty measure for stating that variable C causes variable B, which is the wrong conclusion. It is wrong because the correct causal direction is from B to C, as noted above, and the correct outcome of the algorithm would be to leave the edge undirected. As we have seen here, a wrong orientation can happen a fair amount of the time if there are opposing forces at work that weaken each other to some extent but do not cancel each other out completely (which would violate the faithfulness assumption). In this application, it is the direct effect of A on C that

negates its indirect effect via B. As a side-effect, the direction of causation between B and C is reversed. To summarise, those causal search procedures that have formalised the logical conclusions of conditional independences, like the PC-1 algorithm does, may reach rather strong ontological claims by conditioning on a negative, a failure to reject a null hypothesis. This non-rejection may be false, the probability of it happening is not controlled for in standard correlation tests, and it has ramifications throughout the graph, like wrongly orientating other edges. Third, this phenomenon will likely be even more problematic in higher dimensional systems. With more interaction terms, there is a greater potential for such counteracting effects as exemplified here. With more variables, the potential for wrongly orientated edges due to false negatives is also greater.

To address these concerns, some authors have introduced the notion of 'strong faithfulness.' In addition to usual faithfulness, this assumption also excludes effect sizes that, while not precisely negating each other, are still 'too close' in the sense that they are unlikely to be distinguishable with typical sample sizes (see Section 2.3 and Zhang and Spirtes (2003); Uhler et al. (2013)). For the above example, the parameter constellations would have to be adjusted for strong faithfulness to hold, and results would likely be less adverse. In practice, it may be just as difficult to assess whether the strong variant holds as it is for simple faithfulness.

Incidentally, this paper is not the first to take note of these effects. Spirtes, Glymour, and Scheines (2000, p. 113–121) assess the reliability of the PC-1 algorithm by conducting simulation exercises. They compare different incarnations of the algorithm and conclude that "none of the procedures are reliable on all dimensions when the graphs are not sparse" (p. 115). This conclusion is reached by considering theoretical aspects and simulation evidence. From a theoretical standpoint, the authors argue that single mistakes during the elimination stage will propagate less easily to other parts of the estimated graph with the so-called SGS algorithm than with the PC-1 algorithm (p. 83–88). The former is therefore judged more stable, primarily because it considers more independence relations. However, both algorithms will be misled during the orientation phase by mistakes made during the elimination phase.

In practice, independence needs to be tested for. Therefore, the authors run Monte Carlo simulations, albeit of limited scope (p. 113–121). The simulations assess the probability of making certain mistakes when recovering a graph from observational data. These mistakes include the erroneous removal or addition of an edge and the erroneous omission or addition of an arrow. Finally, one may ask whether the true model is nested in the estimated set of models or whether some alternative model is nested in the estimated set. However, the simulation is of limited scope since each permu-

tation is repeated only ten times, with different parametrisation each time, rendering the effective repetitions one per DGP. These are therefore rather imprecise probability estimates without regard for any remaining simulation uncertainty. However, it still offers the authors a chance to examine typical patterns across different parameter constellations.

They find that "at high average degree and low sample sizes the output of each of the procedures tends to omit over 50 % of the edges in the true graph" and "with high average degree the percentage of edges omitted even at large sample sizes is significant" (p. 116). Addressing these issues, they suggest that using "higher significance levels [...] may improve performance at small sample sizes" (p. 116). However, they ignore the fact that higher significance levels may result in more erroneous edge additions. Moreover, they fail to emphasise the risks involved in accepting the estimated graphs at face value or the lack of uncertainty measures, even though they recognise the procedure's fallibility.

Yet, elsewhere they document the detrimental impact of counteracting variables discussed above, while leaving its precise origins unclear. On pages 203–207, Spirtes, Glymour, and Scheines (2000) observe that the PC-1 algorithm is susceptible to committing frequent type II errors in practice. They even cite an empirical example involving a medical experiment on nineteen rats which seems to suffer from a similar constellation as outlined above. An omitted variable bias (the authors do not identify it as such) eradicates pair-wise correlations, yet when conditioned via multiple regression, the partial correlations are significant. It seems this is driven by highly negatively correlated and thus counteracting variables. The PC-1 algorithm will be misled at the first elimination or second orientation stage, because of these unconditionally uncorrelated variables.

The authors point out on page 203 that "[e]rror probabilities for search procedures are nearly impossible to obtain analytically" and instead rely on Monte Carlo simulations. They generate data based on a graph estimated by the PC-1 algorithm from the rat data. The original sample size was only 19, yet in the simulation they can adjust the number of observations. Creating 100 simulation samples, they find that "[e]ven at sample size 1,000 the search makes an error of type 2 [...] in 55 % of the cases" (p. 206). They note that with highly correlated regressors, the model is "nearly unfaithful" (p. 205). They conclude that "[i]n the absence of very strong prior causal knowledge, multiple regression should not be used to select the variables that influence an outcome [...] in data from uncontrolled studies" (p. 207). What they leave unmentioned is that almost all implementations of the PC-1 algorithm, and other variants like the SGS

algorithm, rely on a form of multiple regression for estimating partial correlations, and thus for inferring stochastic dependence.

2.5 Properties of Causal Graphs in SVAR Analysis

We will now investigate the relation between causal graph theory and structural VAR modelling in more depth. This section will highlight some of the properties inherent to causal graph algorithms when applied to structural VAR models. We will first formalise the relation between graphs and VARs in order to then characterise the outcome space of causal graph algorithms in terms of SVAR models more precisely. We will also discuss under what conditions a graph represents a (partially) identified SVAR model and shortly touch upon on how the most frequently applied algorithm relates to standard likelihood ratio tests.

To formalise the discussion of structural VAR models and graphs, let us specify their relation more closely. The basis for our discussion is an A-type SVAR model

$$\boldsymbol{A}_0 \boldsymbol{y}_t = \sum_{p=1}^{P} \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{2.10}$$

where the contemporaneous effects between the K endogenous variables $\boldsymbol{y}_t$ are given by matrix $\boldsymbol{A}_0$, normalised to have a unit diagonal, and the reduced-form lag coefficients are given by $\boldsymbol{A}_p^* = \boldsymbol{A}_0^{-1}\boldsymbol{A}_p$. The error term is serially uncorrelated with $\mathrm{E}(\boldsymbol{\varepsilon}_t) = 0$ and diagonal covariance matrix $\mathrm{var}(\boldsymbol{\varepsilon}_t) = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. The defining characteristic making this model *structural* is that one or more elements of $\boldsymbol{\varepsilon}_t$ admit an economic interpretation. The graph associated with this model is defined as follows.

Definition 2.5.1. The directed graph $G_{\boldsymbol{A}}$ associated with model (2.10) is defined by $\boldsymbol{\Xi} := \boldsymbol{I}_K - \boldsymbol{A}_0'$, where $\boldsymbol{\Xi}$ is the adjacency matrix of $G_{\boldsymbol{A}}$.

The adjacency matrix of a directed graph usually encodes a single arrow $y_i \to y_j$ in G as $\boldsymbol{\Xi}_{ij} = 1$, but we will simply require that there is such an arrow whenever $\boldsymbol{\Xi}_{ij} \neq 0$.[4] As such, whenever zero restrictions are placed on $\boldsymbol{A}_0$, $G_{\boldsymbol{A}}$ reflects the critical assumptions placed on (2.10). Since (2.10) will generally be underdetermined, a necessary condition dictates that at least $K(K-1)/2$ restrictions are to be placed on $\boldsymbol{A}_0$ for (2.10) to become identified (assuming only restrictions on $\boldsymbol{A}_0$ are considered). Most economically credible restrictions on $\boldsymbol{A}_0$ are in the form of zero restrictions and may thus be conveniently visualised using $G_{\boldsymbol{A}}$. We will employ the stylised case of a

[4] Alternatively, for keeping with convention, the adjacency matrix may also be defined by $\boldsymbol{\Xi}_{ij} = 1_{\boldsymbol{I}_{ij} \neq \boldsymbol{A}'_{ij,0}}$, where 1_c is the indicator function evaluating to 1 whenever c is true and 0 otherwise.

three-equation economy involving prices, output, and the interest rate to illustrate the concept. Suppose our economy is governed by a supply, demand and monetary policy equation such that the following representation holds

$$\begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix} \begin{bmatrix} \pi_t \\ y_t \\ i_t \end{bmatrix} = f(\mathcal{Z}_{t-1}) + \begin{bmatrix} e_{s,t} \\ e_{d,t} \\ e_{mp,t} \end{bmatrix}, \tag{2.11}$$

where π_t is a measure of inflation in period t, y_t is a measure of output growth, i_t is the nominal interest rate, and $\mathcal{Z}_{t-1}$ is a set collecting information on these three variables up to period $t-1$. The function f is linear such that the system conforms to a VAR. The three shocks $e_{s,t}$, $e_{d,t}$, and $e_{mp,t}$ represent a supply, demand, and monetary policy shock. Applications with this simplified model at their core are widespread across the literature, even though such applications carry a number of limitations. See Christiano, Eichenbaum, and Evans (1999) and Kilian and Lütkepohl (2017, Ch. 8) for details. Here, the focus is on how SVAR models of this type relate to graphs.

The adjacency matrix in this case is given by

$$\boldsymbol{\Xi}' = \begin{bmatrix} 0 & 0 & 0 \\ -a_{21} & 0 & 0 \\ -a_{31} & -a_{32} & 0 \end{bmatrix}, \tag{2.12}$$

such that there are arrows going from the first node to the second and third and from the second node to the third. Figure 2.6 displays the graph $G_{\boldsymbol{A}}$ for the system of equations (2.11). The graph summarises the (hypothesised) contemporaneous relations between the endogenous variables. Inflation impacts both output and the interest rate within period t, but not vice versa. Hence there are arrow heads pointing from π_t to y_t and i_t, but not vice versa. Any missing arrow heads thus correspond to zero restrictions placed on $\boldsymbol{A}_0$. Note that the temporal dynamics as well as the error terms have been abstracted away in the graph. That is justified as, under the assumptions made above, they are inconsequential for the identification of the SVAR model via short-run exclusion restrictions on $\boldsymbol{A}_0$. This will be different, however, if restrictions are also placed either on lagged components or in the form of a B-type model or long-run restrictions.

The main takeaway from the previous example is that identifying restrictions for an A-type SVAR model can be suitably summarised in the form of directed graphs. While this may improve communication of important assumptions, we will now turn to other features of directed graphs in the context of VAR models. In essence, graphs as in Figure 2.6 may also aid in assessing identification schemes or, alternatively, in

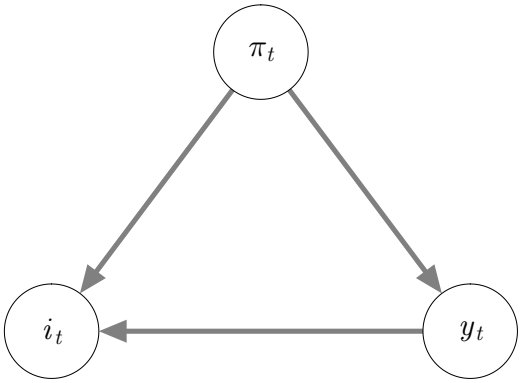

Figure 2.6: The graph $G_{\boldsymbol{A}}$ representing a highly stylised macroeconomic model.

discovering schemes supported by the data. For the latter, a number of algorithms exist, as outlined in previous sections, that test for conditional independence relations in the data and infer which kind of directed graphs adequately represent the DGP. That is, under two key assumptions, they are able to infer structural, *asymmetrical* relations between random variables.

Even though these algorithms have been applied for the discovery of structural VAR models in the literature, their properties have not always been made explicit in the context of SVAR models. The first set of properties are fairly straightforward and reflect the example discussed above. But to reiterate, while these results are often implicitly recognised, they are not always explicitly mentioned. A starting point is the following.

Lemma 2.5.1 (Recursive identification)**.** *For an A-type SVAR model, a recursive identification scheme is equivalent to $G_{\boldsymbol{A}}$ being a DAG.*

Proof. By recursiveness (Wold, 1954) the variables in $\boldsymbol{y}_t$ may be reordered such that $\boldsymbol{A}_0$ in (2.10) is upper triangular. More specifically, there always exists a permutation matrix $\boldsymbol{Q}$ of dimension $K \times K$ such that $\boldsymbol{A}_0^\dagger = \boldsymbol{Q}\boldsymbol{A}_0\boldsymbol{Q}'$ is upper triangular and (2.10) may be rewritten as $\boldsymbol{A}_0^\dagger \boldsymbol{y}_t^\dagger = \boldsymbol{Q}\boldsymbol{A}_0\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{y}_t = \sum_{p=1}^{P} \boldsymbol{Q}\boldsymbol{A}_p\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{y}_t + \boldsymbol{Q}\boldsymbol{\varepsilon}_t$. The adjacency matrix of $G_{\boldsymbol{A}^\dagger}$ is lower triangular since $\boldsymbol{\Xi}^\dagger = \boldsymbol{I} - \boldsymbol{A}_0^{\dagger\prime} = \boldsymbol{Q}\boldsymbol{\Xi}\boldsymbol{Q}'$. Hence, $G_{\boldsymbol{A}^\dagger}$ is a DAG and isomorphic to $G_{\boldsymbol{A}}$ due to permutation similarity of the two adjacency matrices $\boldsymbol{\Xi}^\dagger$ and $\boldsymbol{\Xi}$, i.e. the orientation of edges and adjacencies of nodes are preserved. Therefore, $G_{\boldsymbol{A}}$ is also a DAG. □

The latter result may aid in assessing whether a specific identification scheme is recursive, independent of the variable ordering in $\boldsymbol{y}_t$. Simply draw the associated graph, which looks just the same across variable permutations, and check whether it is acyclic. This perspective is so intuitive that Wold (1954) already employed a version of

a DAG to communicate his ideas about recursiveness; though his arrangement differed slightly, as it illustrates a repeating pattern of the dynamic development of variables over time.

Sometimes the focus lies solely on the reaction of variables to one particular economic shock $e_{p,t}$. The remaining $K-1$ shocks, and the dynamic response to them, is of less importance. If the assumption of recursiveness is still viable, then in such a case the set of variables may be split into three parts $\boldsymbol{y}_{1,t}$, $y_{2,t}$, and $\boldsymbol{y}_{3,t}$. The split is such that $y_{2,t}$ and $\boldsymbol{y}_{3,t}$ may react contemporaneously to changes in $\boldsymbol{y}_{1,t}$ and $e_{p,t}$, but $y_{2,t}$ will react to changes in $\boldsymbol{y}_{3,t}$ only with a delay. A common case is monetary policy, where some variables are inside the central bank's information set Ω_t when forming a decision on the policy stance at time t, while others are not immediately observed or considered by the bank. Here, $e_{p,t}$ is a monetary policy shock, $\boldsymbol{y}_{1,t}$ are the variables inside Ω_t, $\boldsymbol{y}_{3,t}$ are outside of Ω_t, and $y_{2,t}$ is the policy instrument. Importantly, there must not be any instantaneous feedback from $\boldsymbol{y}_{3,t}$ to $\boldsymbol{y}_{1,t}$; otherwise $y_{2,t}$ would also implicitly react to $\boldsymbol{y}_{3,t}$ via $\boldsymbol{y}_{1,t}$.

We can express such a setting quite naturally in graphical terms with the help of a partially directed acyclic graph (PDAG). That's a DAG where some edges remain undirected, but overall edge directions are still constrained in such a way that no cycles are possible. A path v_k — v_l — v_m— v_k is a legitimate path of a PDAG, but cannot be directed as $v_k \to v_l \to v_m \to v_k$. The focus on one particular shock is reflected in Figure 2.7. The left panel summarises a recursive system of seven variables taken from Christiano, Eichenbaum, and Evans (1999), where P_t is the price level, Y_t is output, $PCOM_t$ is a commodity price index, FF_t is the federal funds rate, NBR_t denotes non-borrowed reserves, TR_t denotes total reserves, and M_t represents a monetary aggregate in period t.

Note how the graph can be divided into three components such that edges within each component are not directed. Furthermore, edges between any two components are always orientated in the same direction. This is characteristic of a PDAG. In the left panel of Figure 2.7, the component to the left $\boldsymbol{y}_{1,t} = (P_t, Y_t, PCOM_t)'$ represents those variables which the central bank observes and takes into account when formulating policy. The second component is $y_{2,t} = FF_t$, the policy instrument. The third component $\boldsymbol{y}_{3,t} = (NBR_t, TR_t, M_t)'$ comprises the remaining variables, which are assumed to be affected by the policy change within period t, but do not impact variables in the first or second components contemporaneously. Christiano, Eichenbaum, and Evans (1999) do not take a stance on how the variables within each component affect each other; the edges within components are therefore undirected. The system can equally be repre-

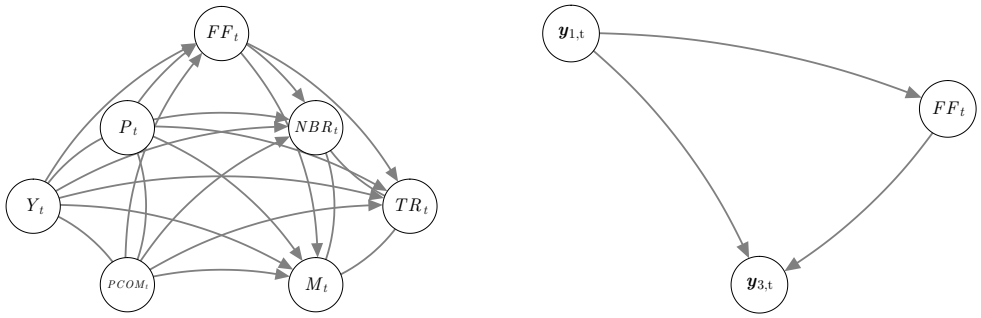

Figure 2.7: The left graph displays the semi-structural, seven variable system of Christiano, Eichenbaum, and Evans (2005) as a PDAG in which a shock to the federal funds rate, interpreted as a money supply shock, is identified. The right graph summarises the block-recursive structure as a DAG.

sented by a DAG, such as on the right of Figure 2.7, where each component—instead of each variable—is now represented by a single node.

For the identification of shocks in PDAGs we have the following sufficient condition, which is a variation of Proposition 4.1 in Christiano, Eichenbaum, and Evans (1999). The main difference is that the proof explicitly takes into account overidentified settings, too. The concepts presented in Appendix 2.A.1 may be helpful in following the proof.

Proposition 2.5.2. *Let $\boldsymbol{y}_t$ be a set of variables following a joint VAR process as in (2.10) with $\boldsymbol{A}_0$ non-singular. Suppose the structural relations among $\boldsymbol{y}_t$ can be summarised as a PDAG G over $\boldsymbol{y}_t$, then a shock ε_t^s to variable $y_{k,t}$ is identified if every edge incident to $y_{k,t}$ is orientated, regardless of orientation.*

Proof. We will show that there always exists a transformation that we can implicitly apply through the Cholesky decomposition and that preserves the response to ε_t^s. To that end, let $\boldsymbol{z}_{1,t} = \mathrm{PA}(y_{k,t}, G)$, $z_{2,t} = y_{k,t}$ and $\boldsymbol{z}_{3,t} = \mathrm{CH}(y_{k,t}, G)$. All remaining nodes can be divided into two further sets: those that are strict descendants of $\boldsymbol{z}_{3,t}$, denoted as $\boldsymbol{z}_{4,t} = \mathrm{DES}(\boldsymbol{z}_{3,t}, G) \setminus \boldsymbol{z}_{3,t}$, and those that are not, $\boldsymbol{z}_{5,t} = V(G) \setminus \{\boldsymbol{z}_{1,t}, z_{2,t}, \boldsymbol{z}_{3,t}, \boldsymbol{z}_{4,t}\}$. Let the number of elements in $\boldsymbol{z}_{i,t}$ be K_i, and $K_1 + \cdots + K_5 = K$. Arrange $\boldsymbol{z}_t = (\boldsymbol{z}_{1,t}', \ldots, \boldsymbol{z}_{5,t}')'$ such that we can write the system as $\boldsymbol{A}_0 \boldsymbol{z}_t = \sum_{p=1}^{P} \boldsymbol{A}_p \boldsymbol{z}_{t-p} + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t = (\boldsymbol{\varepsilon}_t^1, \varepsilon_t^s, \boldsymbol{\varepsilon}_t^3, \boldsymbol{\varepsilon}_t^4, \boldsymbol{\varepsilon}_t^5)'$, and it is assumed that $\mathrm{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'] = \boldsymbol{I}$. Observe that the matrix

$\boldsymbol{A}_0$ can be partitioned as

$$\boldsymbol{A}_0 = \begin{pmatrix} \boldsymbol{A}_{11,0} & \mathbf{0} & \mathbf{O} & \mathbf{O} & \boldsymbol{A}_{15,0} \\ \boldsymbol{a}'_{21,0} & a_{22,0} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' \\ \boldsymbol{A}_{31,0} & \boldsymbol{a}_{32,0} & \boldsymbol{A}_{33,0} & \boldsymbol{A}_{34,0} & \boldsymbol{A}_{35,0} \\ \boldsymbol{A}_{41,0} & \mathbf{0} & \boldsymbol{A}_{43,0} & \boldsymbol{A}_{44,0} & \boldsymbol{A}_{45,0} \\ \boldsymbol{A}_{51,0} & \mathbf{0} & \mathbf{O} & \mathbf{O} & \boldsymbol{A}_{55,0} \end{pmatrix}, \tag{2.13}$$

where $\boldsymbol{A}_{ij,0}$ is of dimension $K_i \times K_j$. Note that $\boldsymbol{a}_{42,0} = \boldsymbol{a}_{52,0} = \boldsymbol{a}_{24,0} = \boldsymbol{a}_{25,0} = \mathbf{0}$ because $z_{2,t}$ is by construction neither parent nor child to any node $y_{i,t} \in \{\boldsymbol{z}_{4,t}, \boldsymbol{z}_{5,t}\}$; $\boldsymbol{A}_{53,0} = \boldsymbol{A}_{54,0} = \mathbf{O}$ because $\boldsymbol{z}_{5,t}$ is by construction not descendant of $\boldsymbol{z}_{3,t}$ or $\boldsymbol{z}_{4,t}$.[5] The remaining zero blocks follow from the fact that G is acyclic.

Suppose none of the $\boldsymbol{z}_{i,t}$ are empty and $\boldsymbol{A}_0$ is therefore non-singular. In that case construct a matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma}\boldsymbol{A}_0$ is a lower block-triangular matrix. Let $\boldsymbol{\Gamma}$ be partitioned with block dimensions equivalent to those of $\boldsymbol{A}_0$, let $\boldsymbol{\Gamma}_{15} = -\boldsymbol{A}_{15,0}\boldsymbol{A}_{55,0}^{-1}$, and all remaining blocks $\boldsymbol{\Gamma}_{ij} = \boldsymbol{I}_{K_i}$ if $i = j$ and $\boldsymbol{\Gamma}_{ij} = \mathbf{O}$ otherwise, where $\boldsymbol{I}_{K_i}$ is the identity matrix of dimension K_i. The matrix $\boldsymbol{\Gamma}\boldsymbol{A}_0$ can be further shaped in lower triangular form by proceeding in the same fashion as Christiano, Eichenbaum, and Evans (1999). That is, compute the QR decomposition for the upper left $K_1 \times K_1$ block $(\boldsymbol{\Gamma}\boldsymbol{A}_0)_{11} = \boldsymbol{Q}_{11}\boldsymbol{R}_{11}$ and for the bottom right $(K_3 + K_4 + K_5) \times (K_3 + K_4 + K_5)$ block $(\boldsymbol{\Gamma}\boldsymbol{A}_0)_{33} = \boldsymbol{Q}_{33}\boldsymbol{R}_{33}$, where non-singularity of both $\boldsymbol{Q}_{11}$ and $\boldsymbol{Q}_{33}$ is implied by non-singularity of $\boldsymbol{A}_0$, and $\boldsymbol{R}_{11}$ and $\boldsymbol{R}_{33}$ are lower triangular matrices with positive diagonal elements.[6] Form the matrix $\boldsymbol{Q} = \operatorname{diag}(\boldsymbol{Q}'_{11}, 1, \boldsymbol{Q}'_{33})$ and observe that $\boldsymbol{Q}\boldsymbol{\Gamma}\boldsymbol{A}_0$ is lower triangular with strictly positive diagonal elements.

Let $\boldsymbol{u}_t$ be a vector of reduced-form residuals. Apply the above transformation to

$$\begin{aligned} \boldsymbol{u}_t =& \boldsymbol{A}_0^{-1}\boldsymbol{\varepsilon}_t = \boldsymbol{A}_0^{-1}\boldsymbol{\Gamma}^{-1}\boldsymbol{Q}'(\boldsymbol{Q}\boldsymbol{\Gamma}\boldsymbol{\varepsilon}_t) \\ =& \tilde{\boldsymbol{A}}_0^{-1}\tilde{\boldsymbol{\varepsilon}}_t, \end{aligned}$$

and note that the (K_1+1)th column of $\tilde{\boldsymbol{A}}_0^{-1}$, $\boldsymbol{a}_{\cdot(K_1+1)}$, and the respective row of $\tilde{\boldsymbol{\varepsilon}}_t$ are unaffected by the transformation. The fact that the vector $\boldsymbol{a}_{\cdot(K_1+1)}$ is uniquely pinned down follows from the uniqueness of the Cholesky decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \mathrm{E}[\boldsymbol{u}_t\boldsymbol{u}'_t]$. Further note that the ordering of the first K_1 variables and the last $K_3 + K_4 + K_5$ variables do not matter. Let $\boldsymbol{P}_{11}$ and $\boldsymbol{P}_{33}$ be permutation matrices of order K_1 and

[5] I beg for pardon for the slight abuse of notation. The objects $\boldsymbol{a}_{ij,0}$ and $\boldsymbol{A}_{ij,0}$ may have different dimensions, of course. In that case strict equality would not hold and the dimension of the zero elements would need to be appropriately adjusted.

[6] To obtain a lower triangular matrix from the QR decomposition of $\boldsymbol{A}$, simply start with the last column instead of the first in the Gram-Schmidt process.

$K_3 + K_4 + K_5$, form $\boldsymbol{P} = \text{diag}(\boldsymbol{P}_{11}, 1, \boldsymbol{P}_{33})$, and observe that the $(K_1 + 1)$th column of $\boldsymbol{\Gamma}\boldsymbol{A}_0$ is appropriately reordered in

$$\boldsymbol{P}\boldsymbol{\Gamma}\boldsymbol{A}_0\boldsymbol{P}'\boldsymbol{P}\boldsymbol{u}_t = \bar{\boldsymbol{A}}_0\boldsymbol{P}\boldsymbol{u}_t,$$

and $\bar{\boldsymbol{A}}_0$ is still lower block-triangular. Apply the same transformation as above with another two QR decompositions, take the inverse and note that the (K_1+1)th column is again unaffected.

If any of $\boldsymbol{z}_{1,t}, \boldsymbol{z}_{3,t}, \boldsymbol{z}_{4,t}, \boldsymbol{z}_{5,t}$ are empty, the construction still goes through by omitting the relevant entries in $\boldsymbol{z}_t$ and $\boldsymbol{A}_0$, and adjusting $\boldsymbol{\Gamma}$, $\boldsymbol{Q}$, and $\boldsymbol{P}$ accordingly. □

The fact that partial identification is achieved under the conditions of Proposition 2.5.2 is quite intuitive, since all variables affecting $y_{k,t}$ are known and predetermined with respect to ε_t^s. Since the PC-1 algorithm returns a PDAG, it may partially identify certain shocks in a statistical sense. Whether they possess a meaningful economic interpretation is another question. In this regard it should be kept in mind that the recursiveness assumption in applying the PC-1 algorithm is by default, not by consideration of circumstance.

For non-recursive systems, graphs can be helpful in deciding whether certain causal effects can be identified using instrumental variables. The intuition is simple. Suppose $x_i \leftrightarrow y_j$ is an edge in graph G which we wish to identify and where both x_i and y_j are observable; further, let ε^{y_j} be the shock to y_j and suppose there are no latent variables present in G other than shocks. One way to identify the effect $x_i \to y_j$ is to use an instrument z_i for which $\text{E}[z_i\varepsilon^{y_j}] = 0$ and $\text{E}[z_i x_i] \neq 0$ holds. In G, these two conditions translate into z_i not being a descendant of y_j, though an ancestor of x_i.

Proposition 2.5.3. *Let $G = (V, E)$ be a directed graph associated with an A-type SVAR model. A node $y_j \in V$ is identified if every parent $x_i \in \text{PA}(y_j)$ is either identified itself or if there is an ancestor $z_i \in \text{ANC}(x_j)$ such that $z_i \notin \text{DES}(y_j)$ and $z_i \neq z_l$ for any other unidentified parent x_l of y_j. The corresponding VAR is fully identified if every node is identified.*

A proof may be sketched as follows. For the jth equation of an A-type SVAR model, regressors fall into two categories: those whose data-generating equation has previously been identified and those that have not. For each of the already identified variables, construct an auxiliary variable by netting out any effect of y_j. For each member of the unidentified group, we know by Proposition 2.5.3 that there is a unique and valid instrument. Every structural parameter can now be consistently estimated by replacing identified endogenous variables with auxiliary variables and employing IV estimation for the resulting set of regressors.

These considerations may be helpful for, first, ascertaining whether IV is a suitable method for estimating causal effects and, second, to construct appropriate instrument sets from a given graph. It should be noted, however, that the criterion above is sufficient but not necessary, as are many others. For example, under some circumstances it may be possible to construct a suitable instrument z_i by conditioning on another set of variables W that block effects of ε^{y_j} on z_i. In that case z_i can be a descendant of y_j and such cases would therefore not be covered by Proposition 2.5.3. In other cases the VAR may be identifiable, but no instruments are available, such as in a three variable system with a graph $y_1 \to y_2 \to y_3 \to y_1$.

Deciding whether a graph admits IV estimation and building algorithms that construct sets of valid instruments is not a trivial task. Finding criteria and building efficient algorithms for instrumental variables in directed graphs or directed acyclical graphs with latent confounders is subject to an active field of research in artificial intelligence (Kumor, Chen, and Bareinboim, 2019; van der Zander, Textor, and Liśkiewicz, 2015; Weihs et al., 2018).

For structural VAR analysis, these methods for constructing instrumental variable sets support efficiency and ease of computation. In contrast to maximum likelihood, IV estimation always offers a closed-form solution if the system is just-identified. In addition, constructing the IV set can be automated for generic applications once the corresponding graph is known. That is useful for implementing the sign identification scheme of Ouliaris and Pagan (2016) in an automated fashion, for example. Details of that scheme are deferred to Chapter 3.

In a similar fashion, Arefiev (2014, 2016a,b) ventures towards a fusion of causal graphs and structural VARs by unifying results from both strands of literature. The two later working papers are essentially split-ups of the earlier one. In particular Arefiev (2016a) translates algebraic rank conditions necessary for identification of an SVAR model to graphical conditions and obtains results similar to Proposition 2.5.3. A major difference is that Arefiev (2016a) considers dynamic simultaneous equation models (SEMs) in which identifying restrictions are also placed on coefficients of lagged endogenous or exogenous variables. The graphs in those papers also take these additional variables into account.

To shed light on the properties of SVAR systems identified through graph-causal means, we have another helpful equivalence. This one involves exact identification schemes.

Lemma 2.5.4 (Exact identification)**.** *For an A-type SVAR model with zero restrictions and K variables, an exact identification scheme is equivalent to $G_{\boldsymbol{A}}$ being connected and having $K(K-1)/2$ arrow heads.*

Proof. Let $\boldsymbol{C}$ be a selection matrix consisting of zeros and ones. The maximum number of arrow heads in a directed polygraph of dimension K is K^2. Thus, if precisely $\frac{K(K-1)}{2}$ arrow heads are allowed, we have $\boldsymbol{C}\,\mathrm{vec}(\boldsymbol{\Xi}') = \boldsymbol{0}$ with $\mathrm{rk}(\boldsymbol{C}) = \frac{K(K+1)}{2}$. Since $\boldsymbol{\Xi}' = \boldsymbol{I} - \boldsymbol{A}_0$, we also have

$$\boldsymbol{C}\,\mathrm{vec}(\boldsymbol{A}_0) = \boldsymbol{C}\,\mathrm{vec}(\boldsymbol{I}) \equiv \boldsymbol{c} \tag{2.14}$$

and the necessary rank condition for identification is therefore just fulfilled.

Now suppose $G_{\boldsymbol{A}}$ is not connected. Then without loss of generality, the K vertices can be split into two disconnected groups $\{1, \ldots, K_1\}$ and $\{K_1{+}1, \ldots, K_2\}$, with $K_1{+}K_2 = K$. Each group forms a subsystem that itself needs $K_j(K_j - 1)/2$ zero restrictions, $j = 1, 2$, in order to be just-identified. Together they imply $K(K-1)/2 - K_1K_2$ zero restrictions, while the disconnect implies $2K_1K_2$ zero restrictions. In total we have $K(K-1)/2 + K_1K_2$ restrictions altogether and therefore either $K_1 = 0$ or $K_2 = 0$.

□

A corollary to Lemma 2.5.4 is that a complete directed graph, complete in the sense that there is exactly one edge between any two vertices, represents an exact identification scheme. That is because a simple counting exercise reveals that a complete graph of dimension K has precisely $K(K-1)/2$ edges. If $K = 2$, there is a single edge in a complete graph. Adding a node to a complete graph of size k will increase the number of edges by k. Thus the number of edges in a complete graph with K nodes is $\sum_{k=1}^{K-1} k = K(K-1)/2$.

But note that the converse to the above assertion is not necessarily true: an exact identification scheme need not form a complete graph. Take Figure 2.8 as a case in point. The figure shows a multigraph, a graph where more than one edge between two vertices is allowed. There are 3 restrictions imposed on the graph, and if the unrestricted elements of the coefficient matrix $\boldsymbol{A}_0$ are drawn, say, from a normal distribution, the system will be identified. Yet, variables U and W are not adjacent and the graph is therefore not complete. In this instance, as in others, the value of $\boldsymbol{A}_0$ is crucial for identification. Suppose the contemporaneous effect of U on V is actually zero, then the SVAR model implied by Figure 2.8 will no longer be identified. Thus, not only is it crucial that certain effects are absent, but also that other effects are present (or do not counteract each other). As previously mentioned, in the probabilistic graph literature this kind of mapping between distributions and graphs is called 'faithfulness' or 'stability'. It is an assumption stating that only those conditional independence relations implied by the graph hold in the joint probability distribution of the data and no other. For example, if U and V were in fact independent even though

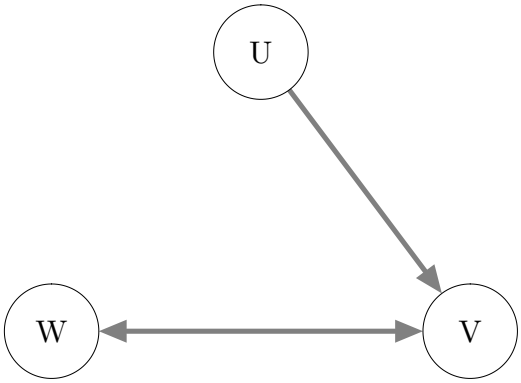

Figure 2.8: Illustration of an incomplete graph that coincides with an exact identification scheme.

the corresponding graph contains an edge between these two variables, the faithfulness condition would be violated since that independence relation would not be implied by the graph.

Another observation inspired by Lemma 2.5.4 and Figure 2.8 is that $G_{\boldsymbol{A}}$ need not be acyclic for (2.10) to be identified. However, Lemma 2.5.1 and 2.5.4 together imply that if the identifying restrictions can be summarised as a complete directed acyclic graph, a necessary condition for identification is fulfilled. In practice, a complete DAG will be very likely to also suffice for identification and one may conjecture that together with the faithfulness assumption it will even be guaranteed. Nonetheless, the graph does not state numeric values and the necessary and sufficient rank condition should thus be checked once parameter estimates are available (Lütkepohl, 2005, p. 361). The bottom line is that a complete DAG is nothing else than a visual view of the usual recursive Cholesky approach to short-run identification.

This equivalence of complete DAGs to the Cholesky approach has implications for the outcome space of causal search algorithms. In particular, a Cholesky ordering, or indeed any other exact identification scheme, will never be the outcome of algorithms that test for patterns of conditional independence in the way the PC-1 algorithm does.

Lemma 2.5.5. *All SVAR models discovered by the PC-1 algorithm will be recursive.*

Proof. By Koller and Friedman (2009, Proposition 3.5) the output of the PC-1 algorithm will be acyclic. In particular, the output will be a set of DAGs that are equivalent up to the conditional independence relations implied by them. By Lemma 2.5.1 if $G_{\boldsymbol{A}}$ is a DAG, then the SVAR model is recursive. Hence, with every element returned by the algorithm being a DAG, all of the implied SVAR models will be recursive. □

The algorithm is specifically designed to recover a true graph G^* that is assumed to be a DAG. It is therefore hardly surprising that the output is acyclic, which maps to

recursiveness in terms of SVAR terminology. For exact identification schemes, this has the following implication.

Proposition 2.5.6. *None of the SVAR models discovered by the PC-1 algorithm will be exactly identified.*

Proof. Suppose $K > 2$ and there is a unique SVAR model that is returned by the PC-1 algorithm and that is just-identified. By Lemma 2.5.4 the associated graph $G_{\boldsymbol{A}}$ is connected and has $K(K-1)/2$ arrow heads. For the algorithm to impose any of the necessary $K(K-1)/2$ zero restrictions, some vertices need to be d-separated. Say vertices i and j are d-separated by a set of vertices S, possibly an empty set. Since the graph is connected, there is a path from i to j, but i and j cannot be adjacent. Since the two vertices are not adjacent, yet we have $K(K-1)/2$ arrow heads, one arrowhead needs to be placed elsewhere than between i and j. Iterating through the graph, there are at least two vertices k and l such that $k \to l$ and $k \leftarrow l$. This contradicts Lemma 2.5.5.

If $K = 2$, then i and j are either adjacent or not. If they are adjacent, they cannot be d-separated. If they are not adjacent, the system is over-identified. In either case the system will not be just-identified. □

A unique just-identified SVAR system cannot be discovered by the PC-1 algorithm. In fact, if the true system is just-identified and correctly estimated, not a single edge will be directed, such that the output is the set of all just-identified systems. This mirrors the well known fact that just-identified systems are observationally equivalent. However, if the system is over-identified, the algorithm may recover the complete graph, or partially identify nodes if the conditions of Proposition 2.5.2 are fulfilled. Moreover, if an exact non-recursive system were the ground truth, then the PC-1 algorithm may, even asymptotically, either force a loop like $k \leftrightarrow l$ as $k \to l$, thus imposing at least one more restriction than necessary and over-identify the system, or it may leave the complete graph undirected. The former case is clearly undesirable as it misspecifies the system.

If the PC-1 algorithm is only able to detect over-identified systems, a natural question is how this relates to standard likelihood ratio tests of overidentifying restrictions that are frequently used in structural VAR analysis. Can a system estimated via the PC-1 algorithm be rejected by a likelihood ratio test? Will the two coincide asymptotically? Standard implementations of the PC-1 algorithm conduct sequences of partial correlation tests that are, at least asymptotically, Wald tests. As is well known (Engle, 1984), Wald and likelihood ratio tests are asymptotically equivalent. Thus, if the size of both testing procedures is appropriately corrected to take multiple testing into account,

one may conjecture that asymptotically a system recovered by the PC-1 algorithm will not be rejected by a likelihood ratio test. In finite samples, however, all bets are off. The set of null hypotheses that were not rejected by the sequence of tests conducted by the PC-1 algorithm may jointly be rejected by a likelihood ratio test. Equivalently, a likelihood ratio test may not reject a joint hypothesis whereas the sequence of Wald tests would.

Overall, even though the two testing regimes do not fully coincide, one can view the PC-1 algorithm as being informative about what kind of over-identifying restrictions are likely not to be rejected by the data. This feature of the algorithm may be used to carefully explore data characteristics.

2.6 Conclusion

This paper has explored the nexus between structural vector autoregressive (SVAR) models and causal graphs. There is a straightforward mapping between SVAR models identified via short-run restrictions and graphs which is convenient for visualising, communicating and checking core identification assumptions with the help of a graph. Furthermore, the outcome space of the PC-1 algorithm for discovering graph structures has been discussed in the context of SVAR models. The outcome of the algorithm is always acyclic such that only recursive models can be discovered. Mirroring the well known observational equivalence of just-identified models, the algorithm will either leave the complete graph undirected or impose some over-identifying restrictions. These restrictions may partially identify shocks, fully identify the model or fail to identify any shocks, depending on context. One may conjecture that these restrictions will asymptotically not be rejected by a likelihood ratio test of comparable size.

There are three weaknesses of the algorithm. One is particular to SVAR models, where non-recursive identification schemes are usually not excluded a priori. Thus, if the true structure is non-recursive, the algorithm will either not impose restrictions at all or force through an over-identified, recursive scheme. The latter is clearly undesirable as it may heavily misspecify the model. A second weakness is related to uncertainty in finite samples. An important assumption made by the algorithm is that the joint distribution and the graph are *strongly faithful* to each other. This means that the (conditional) independence relations between random variables implied by the graph can reliably be recovered, and no other independencies will be found. But in finite samples there is always a range of outcomes where independence relations are not rejected due to insufficiently informative data. Thus, the mapping of independence relations between distribution and graph is noisy. What is more, there is no measure of uncertainty that could reflect how reliable the outcome of the algorithm is. Once an

independence relation has been attested, it is assumed as certain in all further steps. The algorithm may therefore suffer from an accumulation of type II errors that would distort the outcome. If there were some measure to indicate how severe this problem is, this particular weakness may be alleviated. For the third weakness, that is the case: The reliance on multiple hypothesis testing means that type I errors are also accumulating. That may be addressed with standard procedures controlling the FWER or false discovery rate (FDR). As it stands, causal discovery methods like the PC-1 algorithm may serve as exploratory tools for assessing which over-identifying restrictions could be supported by the data. But researchers should be cautious in accepting the algorithm's outcome as given.

The field of causal discovery is growing. It particularly thrives on insights from the machine learning literature. Thus, future research may address other forms of causal discovery, for example by relying on other tests for conditional independence or on methods that assess the direction of causality by other means than through conditional independence tests. These could include score-based or Bayesian learning procedures which have so far found little application to structural VARs with the noticeable exception of Ahelegbey, Billio, and Casarin (2016). In the context of the PC-1 algorithm a straightforward extension would be to control the FWER or FDR. For SVAR analysis, representing identification restrictions other than for A-type models may also be a worthwhile endeavour. B-type models, for example, would rely on graphs and methods that are able to represent latent variables. Finally, graphical discovery algorithms might supplement existing identification schemes by carefully exploring properties of the data and thereby improve modelling of data characteristics.

2.A Appendix

2.A.1 Probabilistic Graph Theory

This appendix reviews concepts from probabilistic graph theory in a nutshell. As graph theory is not a mainstream econometric tool, I believe reviewing the material here is beneficial for understanding results and discussions more thoroughly. More detailed accounts of probabilistic graphs are available in textbooks by Pearl (2009), Koller and Friedman (2009), and Spirtes, Glymour, and Scheines (2000). Among the first to write on the subject of probabilistic graphs in textbook form is Pearl (1988). Summaries with a similar econometric perspective are given by the papers of Demiralp and Hoover (2003) and Kwon and Bessler (2011). For details on graph theory itself, see the introductory textbooks by West (2001), Hartsfield and Ringel (2003) or the graduate-level text by Diestel (2017).

A graph usually consists of two sets, a set of vertices and a set of edges. Vertices, which are sometimes also called nodes, are the primitives of a graph and may represent any kind of object: letters, numbers, points, variables, economic concepts, or traffic junctions. An edge is always based on two vertices, not necessarily distinct, and is viewed as a connection between the two vertices. These in turn are called the endpoints of the edge. In addition, there is an endmark specifying the nature of each endpoint.

Definition 2.A.1 (Vertex)**.** Let $v_1, \dots, v_K$ be a finite number of distinct objects of interest. Call each one a vertex and all together a set of vertices V.

Definition 2.A.2 (Edge)**.** Let m be the set of possible endmarks and let V be the set of available vertices. Define $e_l = (V_l, m_l)$ as an edge based on $V_l = (v_{l_1}, v_{l_2}) \in \mathfrak{V} \subseteq V \times V$ and $m_l = (m_{l_1}, m_{l_2}) \in \mathfrak{m} \subseteq m \times m$. Finally, denote $E = \{e_1, \dots, e_L\}$ as a set of edges.

Definition 2.A.3 (Graph)**.** Let V be a set of vertices, m a set of marks, and E a set of edges based on V and m. Then a graph is defined as the ordered pair $G = (V, E)$.

Given a graph G, the set of vertices of that graph are given by $V(G)$ and the set of edges by $E(G)$. Sometimes, graphs are alternatively defined as triples with a set of vertices, a set of edges disjoint from the set of vertices, and a mapping between these two sets. Either approach will suffice for our purposes. For simplicity, in our case the dependence of E on V and m is not further indicated. It does depend, however, on these two sets and we can place restrictions on the implicit mapping from E to V and m. In this way we can further define certain classes of graphs either through these restrictions or by directly setting up E and m in a suitable way.

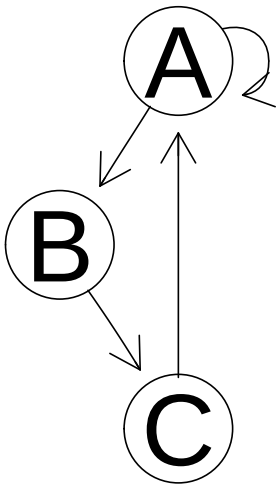

Figure 2.9: Example of a directed graph.

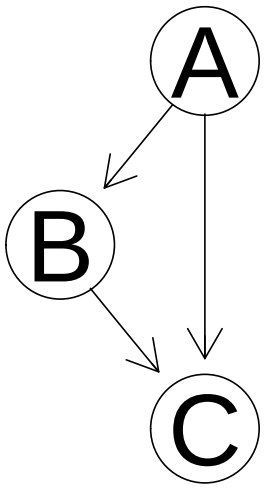

Figure 2.10: Example of a DAG.

Example (Directed Graph). Let us assign $V = \{A, B, C\}$ and $m = \{<, >, \varnothing\}$. Then the graph shown in Figure 2.9 is a member of the class of directed graphs in which the set of marks of each edge is restricted to be either $m_l = (<, \varnothing)$ or $m_l = (\varnothing, >)$. The symbols constituting m can be conveniently interpreted in this case as arrowheads pointing left ($<$), right ($>$) and as an empty head ($\varnothing$).

Directed graphs reflect a hierarchical structure between vertices. Usually, this also implies an asymmetry in that some vertices are being pointed at while others are the source of pointers. It is customary to characterise these relations in terms akin to family relations. From now on we will also shorten the notation and indicate the presence of an edge $e_l = ((v_{l_1}, v_{l_2}), (\varnothing, >))$ in G as $v_{l_1} \to v_{l_2}$.

Definition 2.A.4 (Kinship)**.** For a generic directed edge $v_{l_1} \to v_{l_2}$, v_{l_1} is the parent vertex and v_{l_2} is the child vertex. Recursing through the graph, any parent, parent's parent, and so forth, of a particular vertex v_k is called an ancestor of that vertex. Likewise, any child, children's child, and so forth, of v_k is called a descendant of v_k. For a vertex v_k, the sets of parents, ancestors, children, and descendants are denoted as $\text{PA}(v_k)$, $\text{ANC}(v_k)$, $\text{CH}(v_k)$, $\text{DES}(v_k)$.[7]

A graph can be further characterised as complete or connected. A complete graph is one where there is an edge between any two vertices. A connected graph is one where there is a path between any two vertices that one may traverse along. Figure 2.8 is a

[7] In some parts of the literature, v_k itself is a member of $\text{DES}(v_k)$ and $\text{ANC}(v_k)$ by default for definitional or practical reasons. That usage is not followed here.

connected graph, for example, but it is not complete since the edge U—W is missing. If the edge $U \to V$ were absent as well, Figure 2.8 would not be connected either.

Definition 2.A.5 (Path)**.** A path in graph G refers to a sequence of edges $e_{p_1}, \ldots, e_{p_J}$ such that every element $e_{p_i} \in E(G)$ is unique and for any two consecutive edges e_{p_i} and e_{p_j} we have that $v_{i_2} = v_{j_1}$, with $j = i + 1$. If there are no two edges, then a path is simply an edge. A directed path is a path where in addition all edges are either $\to$ or $\leftarrow$.

In a directed graph, a single vertex may be its own parent and child, as for example vertex A in Figure 2.9 is. Two different vertices may be parent and child to each other at the same time. Or, indeed, a single vertex may be its own ancestor and descendant if one can cycle through vertices, as is for example the case for every vertex in Figure 2.9. We will use another class of graphs to prevent such cycles.

Example (DAG). Let $V = \{A, B, C\}$ and $m = \{<, >, \varnothing\}$ again. Then the graph in Figure 2.10 is a member of the class of directed acyclical graphs (DAGs) in which the mapping from E to V and m is restricted in such a way that no vertex may be its own descendant.

DAGs are especially convenient for causal analysis because every DAG with one or more edges has at least one vertex that is only a receiver of pointers and at least one vertex that is only a source of pointers. Moreover, there exists an ordering of vertices $v_{s_1}, \ldots, v_{s_K}$, not necessarily unique, such that if v_{s_l} is a descendant of v_{s_k}, then $l > k$, and if v_{s_l} is an ancestor of v_{s_k}, then $l < k$. This property adequately mirrors the approach taken in many econometric analyses, where identification of causal effects relies on the assumption that for each target of interest there is some form of exogenous variation available that remains unaffected by alterations further down the line. In addition, in economic theory there is usually an understanding of the specific directionality from right-hand-side to left-hand-side variables within each of the isolated equations that give rise to a full model. Either of these considerations could be summarised in a directed graph by equating variables with vertices and the structural relations between them as directed edges, though whether they comprise a DAG depends on the application at hand.

Nonetheless, if researchers are willing to assume that the system under scrutiny complies with a DAG, then a number of useful results from probabilistic graph theory can be applied. For these results to hold, a few other key assumptions are necessary. First, for every variable v_k, we will assume that we can find a set of variables that shield v_k from the influence of all others, except from those on which v_k has a direct or indirect influence itself.

Definition 2.A.6 (Markov Condition)**.** Let $\boldsymbol{y} = (y_1, \ldots, y_K)'$ be a vector of random variables with joint probability distribution $f_{\boldsymbol{y}}$. The distribution $f_{\boldsymbol{y}}$ and a DAG $G_{\boldsymbol{y}}$ with vertex set $V = \{y_1, \ldots, y_K\}$ and edge set E are said to fulfil the *Markov condition* if

$$y_k \perp\!\!\!\perp \Big(V \setminus \mathrm{DES}(y_k) \cup \mathrm{PA}(y_k) \cup y_k\Big) \;\Big|\; \mathrm{PA}(y_k) \tag{2.15}$$

holds for $k = 1, \ldots, K$.

The condition is sometimes also called the local or parental Markov condition. The condition is the first to establish a relation between the properties of a set of random variables and the elements of a graph. The symbol y_k now represents both a random variable and a vertex. The usefulness of the Markov condition stems from the fact that it offers a straightforward decomposition of the joint probability distribution

$$f_{\boldsymbol{y}}(\boldsymbol{\omega}) = \prod_{k=1}^{K} f_{y_k}\left(\omega_k \middle| \boldsymbol{\omega}_{\mathrm{PA}(y_k)}\right), \tag{2.16}$$

where ω_k represents a realisation of y_k and $\boldsymbol{\omega}_{\mathrm{PA}(y_k)}$ a realisation of the vector of random variables corresponding to $\mathrm{PA}(y_k)$. This decomposition has the potential to reduce the complexity of the process at hand since the number of interactions between variables and thus also the number of parameters governing the process may be lowered. However, its usefulness in encapsulating structural information is limited by the fact that $G_{\boldsymbol{y}}$ need not be unique. In fact, for a given probability distribution, a trivial graph fulfilling the Markov condition is any complete DAG of appropriate order. In a complete DAG, every vertex is either parent or child to every other vertex and there are therefore no independence relations, either conditional or unconditional, implied by the graph under the Markov condition.

To make the use of graphs more fruitful, we will require that existing edges actually reflect a direct interaction between variables.

Definition 2.A.7 (Minimality Condition)**.** For a probability distribution f and a graph G fulfilling the Markov condition, the graph G is said to fulfil the minimality condition if the Markov condition is violated after removing any edge from G.

The minimality condition places more stringent requirements on the graph representing our joint probability distribution. Nonetheless, there may still be a multiplicity of graphs fulfilling the condition. But with both conditions in place, we will give such combinations of distribution and graph a name to refer to.

Definition 2.A.8 (Bayesian Network)**.** Let $f_{\boldsymbol{y}}$ and $G_{\boldsymbol{y}}$ be a distribution and a graph fulfilling the Markov and minimality condition. Then the pair $B_{\boldsymbol{y}} = (G_{\boldsymbol{y}}, f_{\boldsymbol{y}})$ is said to be a Bayesian network over $\boldsymbol{y}$.

The term alludes to Bayes because of the subjectivity that is often entailed in setting up the pair of distribution and graph. The other reason is that the graph may also serve as a guide, in accordance with Bayes Law, for updating beliefs about states of variables once information about their ancestors or descendants become available.

Equipped with these concepts, we are ready to infer, by simple inspection of a graph, that certain variables are not related to each other. If there is no path between two variables y_i and y_j, for example, we know that they must be independent of each other. Less clear are relations between variables that do have a path connecting them. They may or may not be independent depending on the nature of the path(s) between them. A further concern is whether we can make y_i and y_j become independent by conditioning on a third set of variables Z. It will turn out that such conditioning is possible whenever the two variables are *directionally separated* or d-separated by Z in a graph $G_{\boldsymbol{y}}$ associated with a Bayesian network over $\boldsymbol{y}$.

D-separation applies to graphs, but its purpose is to facilitate reasoning about interactions between random variables. Let us leave the precise nature of this interaction open for the moment and instead adopt a more intuitive language. Suppose there is some kind of transmission occurring between adjacent vertices in a directed graph G^d. The transmission may refer to messages or signals, or some other form of asymmetric interaction that travels along consecutive edges. The asymmetry is again reflected in the directionality of the graph. A directed edge $v_k \to v_l$ will only transmit from v_k to v_l but not vice versa. A path $v_k \to v_l - v_m$ transmits from v_k to v_l and perhaps from v_l to v_m or v_m to v_l or both, but not from v_m to v_k. Because of this asymmetry, the occurrence of so-called colliders or v-structures is central to analysing the directional separation of vertices.

Definition 2.A.9 (Collider)**.** Let $B = (G, f)$ be a Bayesian network and let $P = e_{p_1}, \ldots, e_{p_J}$ be a path in G and $y_1, \ldots, y_I$ the sequence of vertices on path P. A vertex y_i on P is called a collider if two arrows collide $y_{i-1} \to y_i \leftarrow y_{i+1}$ on P.

This notion is central to separation because a collider does not transmit any changes occurring in y_{i-1} to y_{i+1} along path P. More generally, we can inspect every path between two vertices y_k and y_l and ask ourselves for each path whether it would transmit any changes occurring in one variable to the other if information can flow freely from vertex to vertex. We can also ask ourselves if this transmission is broken once a certain set of vertices is interfered with. Let us call a path that transmits *unblocked* and a non-transmitting path *blocked*. Perhaps it is blocked by controlling an intermediate vertex v_l, or perhaps it is naturally blocked by a collider. The insight of d-separation is that two vertices are separated whenever every path between them is blocked.

Definition 2.A.10 (Blocked Path)**.** A path P between two vertices v_k and v_l in directed graph G^d is blocked by a set of vertices Z if Z is such that

i) P contains at least one collider that is not included in Z and none of its descendants are included in Z or in $\{v_k, v_l\}$, or

ii) at least one other vertex on P that is not a collider is included in Z.

If neither of these conditions apply, then P is unblocked.

In the graph $U \to V \to W$, the path from U to W would be blocked by $Z = \{V\}$. The path would be unblocked by $Z = \varnothing$. In Figure 2.8, on the other hand, we have $U \to V \leftrightarrow W$. Here, the path from U to W cannot be blocked by either $Z = \{V\}$ or $Z = \varnothing$ since V is a collider and its child is W.

We can now define separation not just between two nodes, but between sets of vertices as well.

Definition 2.A.11 (D-Separation)**.** Let X, Y and Z be distinct sets of vertices in directed graph G^d. The sets X and Y are d-separated by Z, denoted d-sep$(X, Y|Z)$, if for any two vertices $v_k \in X$ and $v_l \in Y$ every path between v_k and v_l is blocked by elements in Z.

D-separation considers all paths running between X and Y and determines for each path whether it is *blocked* or *unblocked* by Z. If all paths are blocked, then X and Y are d-separated. The great insight of d-separation is that colliders block paths by their very nature, whereas paths without colliders are unblocked by default and must be blocked by a shielding vertex. The value of d-separation lies in the fact that, when applied to a Bayesian network, it describes every independence relation that is present in the network over and beyond those which are given by the Markov condition. We will denote the set of independence relations implied by d-separation as $I(G) = \{X \perp\!\!\!\perp Y|Z : \text{d-sep}(X, Y|Z)\}$. Similarly, for a probability distribution f, denote the set of independence relations that hold for f as $I(f)$.

Theorem 2.A.1 (Meek (1995))**.** *Let $B_{\boldsymbol{y}} = (G_{\boldsymbol{y}}, f_{\boldsymbol{y}})$ be a Bayesian network. Then for almost all discrete or normal probability distributions $f_{\boldsymbol{y}}$, we have that $I(f_{\boldsymbol{y}}) = I(G_{\boldsymbol{y}})$.*

The noteworthy feature about d-separation is that it extends the set of local Markov independence relations that we started with to a global set that holds in the underlying distribution and that we can find by scrutinising a graph. The proof of Theorem 7 and 8 in Meek (1995) apply for discrete and normal distributions, though one may conjecture that it holds for other type of distributions, too. The result holds for every probability distribution except for a set of distributions with Lebesgue measure zero, as the following example illustrates.

Example. Let $B_{\boldsymbol{y}}$ be a Bayesian network over $\boldsymbol{y} = (y_1, y_2, y_3)'$ with a graph equivalent to Figure 2.10. Now suppose its data generating process is described by the following three equations

$$y_1 = u_1, \tag{2.17}$$

$$y_2 = ay_1 + u_2, \tag{2.18}$$

$$y_3 = by_2 + cy_1 + u_3, \tag{2.19}$$

where u_1, u_2, u_3 are exogenous random variables following a normal distribution and are excluded from the graph. If $b = 1/a$ and $c = -1$, then $y_1 \perp\!\!\!\perp y_3$ even though $I(G_{\boldsymbol{y}})$ is empty.

In this example, there is an independence in $f_{\boldsymbol{y}}$ which is *not* found by the d-separation criterion and thus $I(f_{\boldsymbol{y}}) \supset I(G_{\boldsymbol{y}})$. The independence arises due to a rather pathological parameter constellation where two effects cancel each other out; a constellation that occurs with probability zero if the parameters a, b, c were drawn at random from a continuous distribution. Nevertheless, it is a caveat that should be kept in mind when analysing the properties of probabilistic graphs.

Sometimes a slightly stronger assumption is made that precludes concerns about inconvenient parameter constellations. This is especially helpful when testing statistical properties implied by the graph or when learning structure from data.

Definition 2.A.12 (Faithfulness)**.** A distribution f is *faithful* to graph G if $I(f) \subset I(G)$.

The faithfulness assumption and the Markov condition together imply that $I(f) = I(G)$. In such a case, graph G is also called a *perfect map* of f. A perfect map may not always be available for every application. Apart from parameter regularities, there are other settings which are inappropriately modelled by Bayesian networks. For example, when independence relations shift in response to which *values* certain variables take, then the faithfulness assumption may not be warranted (see Koller and Friedman, 2009, p. 81 for details). If it does exist, a perfect map need not be unique, either. Think of a complete DAG again, where the set $I(G)$ is empty. A complete DAG is isomorphic to any other complete DAG. Therefore, a distribution represented by a complete DAG would have many valid perfect maps. For this reason, the set of graphs $G_{\boldsymbol{y}}$ that may serve as components of a Bayesian network $B_{\boldsymbol{y}}$ is usually described by its equivalence class.

Definition 2.A.13 (Equivalence Class)**.** Two graphs G_i and G_j belong to the same equivalence class $\mathcal{G}$ whenever $I(G_i) = I(G_j)$.

Colliders are important for characterising an equivalence class, because they are the only pattern associated with a distinct independence relation. A path v_k — v_l — v_m is unconditionally blocked if and only if the middle vertex is a collider. However, they can only be unconditionally d-separated, and thus unconditionally independent, if there is no edge between v_k and v_m. Since lack of such an edge is a necessary (but not sufficient) condition for d-separation, this structure is given a special name. Whenever the vertex v_l is a collider and there is no edge between v_k and v_m, then v_l is called an unshielded collider. A second important structure for analysing equivalence classes is a skeleton. For a directed graph G, the skeleton of G is an undirected graph H such that every edge in G is replaced by an undirected edge in H. We can now characterise the equivalence class associated with a distribution.

Theorem 2.A.2 (Koller and Friedman (2009, Theorem 3.8))**.** *Let f be a probability distribution with at least one perfect map G_i. A graph G_j is in the same equivalence class as G_i, and thus also represents f, if and only if G_j has the same skeleton and the same unshielded colliders as G_i.*

CHAPTER 3

Sign Restrictions and Causal Learning in Structural VARs: A First Case Study Using Oil Market Data

3.1 Introduction

By now, a whole array of identification routines for SVAR models exist, some relying on subject matter considerations, some on features of the data, and others on a combination of both. This article seeks to expand this arsenal further by combining two hitherto unrelated identification routines: sign restrictions on structural parameters on the one hand and graphical modelling of features of the data on the other hand. This combination is applied to the global crude oil market, where it successfully replicates previous findings.

Combining these routines may be advantageous as it addresses drawbacks that either routine, on its own, suffers from. Sign restrictions usually impose mild conditions that are acceptable to a wide audience, at the cost of lower identification and estimation precision. The resulting set of identified structural models may be so large that only few economic conclusions can be reached safely. On the upside, identification via sign restrictions is a widely accepted method, is likely to pin down economic concepts coherently and is under constant development by an active research community.

Graphical modelling, on the other hand, is a rarely used tool among economists. It represents one particular type of reasoning in the literature on causality that seeks to represent causal relations in the data through graphs comprised of nodes and edges. The logic underlying this reasoning is not bound to graphs, but it offers a convenient framework.

This kind of causal modelling stems from considerations mainly entertained by computer scientists to inductively infer structure among a set of variables without the necessity of prior causal knowledge. As such, it is not guaranteed to recover anything that lends itself to intuitive economic interpretation. Rather, graphical algorithms

formalise a search for independence patterns among variables. One common implementation of this is to conduct a sequence of hypothesis tests with little regard for the overall sampling uncertainty. Accepting the outcome of this search as given might place subsequent conclusions on thin ice. In addition, assumptions such as recursiveness are frequently invoked to ease the search task.

For these reasons, graphical modelling has remained on the fringes of structural time series analysis. Yet, ignoring new methodological insights from the causal search literature altogether may prove a missed opportunity. I therefore propose to combine the best of both worlds: the economic intuition provided by sign restrictions, while limiting, in a conservative enough fashion such as not to jeopardise any final conclusions, the scope for admissible models via features of the data. More specifically, I propose to run a 'graphical' pre-analysis, the results of which are incorporated in the subsequent sign-identified VAR analysis. In the spirit of a recent push to break away from separating test results into 'significant' and 'non-significant' outcomes (Wasserstein, Schirm, and Lazar, 2019), I suggest that researchers carefully review the properties highlighted by graphical procedures and decide which aspects are safe to build upon. In this paper, the pre-analysis consists of simple partial correlation patterns that will indicate, under suitable statistical assumptions, whether any two variables are not directly causing each other. In principle, the pre-analysis may also include other recent causal methods that relax some of the invoked assumptions.

The advantages of not just relying on the graphical approach for identification of the VAR are threefold. First, depending on the application at hand, and possibly on the degree to which the data are 'tortured', the method will either produce an over-identified or unidentified system. In the unidentified case the system may still be partially identified, such that a subset of structural shocks can be recovered. This variability of success is irrelevant in the present context, because identification is always achieved via sign restrictions. Second, the unknown statistical uncertainty surrounding graphical modelling is reduced by focusing on those patterns that seem most prevalent. Third, the graphical procedure most often used for structural VARs relies in part on sample characteristics and in part on logical conclusions that arise from the assumption of recursiveness. That assumption can be weakened by leaving those logical conclusions aside.[1]

One challenge in incorporating the results from this prior graphical screening is that most algorithms implementing sign restrictions sample the impact matrix of structural

[1]The logical conclusions arise because the graph associated with a recursive VAR is a directed acyclical graph (DAG). In a DAG some edges may be orientated simply because any other orientation would violate the assumption of acyclicity. Their orientation therefore does not immediately rely on sample information.

innovations, whereas graphical tools focus on the contemporaneous effects between endogenous variables. In this paper, the sign restriction algorithm of Ouliaris and Pagan (2016) is used which allows sampling the contemporaneous effects of endogenous variables and thus to easily incorporate the graphical insights. An alternative remedy would be the Bayesian framework of Baumeister and Hamilton (2015) or to map the prior restrictions from contemporaneous effects to the structural impact matrix.

I apply the method in a case study using the oil market model of Kilian and Murphy (2012). The original study points out the lack of meaningful identification when relying exclusively on impact sign restrictions. The authors therefore also rely on extraneous information to shrink the set of admissible models. Ignoring that additional information, the method I apply successfully replicates the findings of the original study by relying on features of the data instead. This suggests that the proposed identification strategy is a viable option whenever such extraneous information is unavailable or disputed.

The remainder of this paper is organised as follows. Section 3.2 reviews the literature on sign-identified VAR models as well as causal graphs. Section 3.3 discusses the details of these two methodologies. The practical use of combining them is illustrated in Section 3.4, with an example taken from the literature on oil market VARs. Finally, Section 3.5 concludes.

3.2 Literature

Sign restrictions in structural VAR models have been pioneered in particular by Uhlig (2005), Faust (1998), and Canova and De Nicoló (2002). Since then the approach has been widely adopted. It has also seen further methodological improvements, which are discussed in greater detail by Kilian and Lütkepohl (2017, chap. 13) and Uhlig (2017).

There are several challenges for sign-identified VAR models. One challenge is to effectively sample the space of admissible models. Since the restrictions usually do not pin down structural parameters uniquely, it becomes necessary to explore the properties of all models that fulfil the sign restrictions. Of particular interest is the type of economic conclusions that are supported by the set of valid parametrisations. To explore the model space, a popular choice is to orthogonalise the reduced-form errors and then to sample certain linear combinations of these errors such that they fulfil the restrictions (Rubio-Ramírez, Waggoner, and Zha, 2010). Ouliaris and Pagan (2016) have proposed another algorithm that I will adopt in this article. They suggest drawing a specific number of structural parameters at random, estimating the structural form conditional on these draws via instrumental variables and retaining those parametrisa-

tions which meet the sign restrictions. A more detailed discussion is deferred to Section 3.3.

Another challenge is inference. Most studies adopt a Bayesian framework for inference in which set identification can be seamlessly incorporated. But even in a Bayesian setting, one issue of contention is how to formulate prior knowledge in the most convincing way (Baumeister and Hamilton, 2015; Bruns and Piffer, 2019). In frequentist settings, different avenues are beginning to bear fruit (Granziera, Moon, and Schorfheide, 2018; Gafarov, Meier, and Montiel Olea, 2018), but are not yet always applicable. Another debate revolves around how best to summarise inference across a set of sign-identified models (Fry and Pagan, 2011; Inoue and Kilian, 2013) or across multiple forecast horizons or response variables (Inoue and Kilian, 2016, 2020). As Inoue and Kilian (2016) and Uhlig (2017) note, the extent to which this kind of joint inference is important very much depends on the question at hand. But what certainly remains a concern for any kind of inquiry is general estimation precision. The latter can be low when only set identifying assumptions such as sign restrictions are imposed.

Due to this lack of precision in sign-identified VARs, many authors complement sign restrictions with additional identifying assumptions. Baumeister and Hamilton (2020) provide an overview of studies with sign-identified VARs in this regard. They stress the need for including additional information since otherwise the sign-identified analysis is often either meaningless or misleading. A selection of additional means of identification beyond impact sign restrictions include restricting the sign of cross-correlations between impulse response sequences at different horizons (e.g. Canova and De Nicoló, 2002), restricting the sign of impulse response functions over multiple horizons (e.g. Uhlig, 2005), penalising impulse response functions (again Uhlig, 2005), bounds on the magnitude of certain transformations of structural parameters (e.g. Kilian and Murphy, 2012), short-run exclusion restrictions (e.g. Arias, Rubio-Ramírez, and Waggoner, 2018), and using instrumental variables in addition to sign restrictions (e.g. Nguyen, 2019). Many of these additional restrictions are motivated by subject matter considerations.

There are, however, also studies that achieve identification through statistical means, independent of sign restrictions. Most widespread among them are methods that exploit changes in volatility (see Lütkepohl and Netšunajev, 2017) or non-Gaussianity of the error terms (Lanne, Meitz, and Saikkonen, 2017). These methods usually achieve full identification and therefore render any other economic restrictions overidentifying.

Graphical modelling of contemporaneous dependencies, while also a data-driven approach, will either overidentify a VAR or fail to achieve full identification, depending on the data. In addition, overidentification is usually only possible with an unknown,

and therefore potentially large, margin of error. Nonetheless, this approach has been used by a number of researchers in macroeconomics and outside. Graphical path representations of statistical dependence have been used informally as early as Wright (1921) and Wold (1954). Researchers such as Pearl (1988, 2009) and Spirtes, Glymour, and Scheines (2000) have subsequently formalised earlier notions by building a theory that relates graphs to probabilistic concepts. Sometimes, these probabilistic concepts are infused with causal ideas, though they need not be. Such causal graph analyses have gained some critical attention in parts of the microeconometric literature (Imbens, 2019). In macroeconometric VAR analysis, algorithms and ideas born out of causal graph theory have been applied by Swanson and Granger (1997); Demiralp and Hoover (2003); Demiralp, Hoover, and Perez (2008); Hoover, Demiralp, and Perez (2009); Demiralp, Hoover, and Perez (2014); Moneta (2008); Moneta et al. (2011); Fragetta and Melina (2011, 2013), amongst others.

These studies all identify VARs using insights from graph theory that equate certain forms of conditional statistical independence among the set of endogenous variables with causal relations. However, the studies fall short of supplying insights for policy as structural shocks are never identified. This is predominantly so because a meaningful economic interpretation of those structural shocks is often difficult without some economic bearing in the first place.

In practice, independence needs to be tested for. The studies cited above employ a sequence of classical hypothesis tests to detect correlation patterns from which they infer dependence. There is also a vibrant Bayesian literature that has developed certain ways to sample graphs. Ahelegbey, Billio, and Casarin (2016), for instance, implement a Bayesian estimation technique that samples DAGs in the context of structural VAR models. By imposing an acyclicity constraint on both the lag structure and the contemporaneous relations between variables, they achieve two goals: identification of the VAR model and a more parsimonious autoregressive part. The Bayesian graphical VAR (BGVAR) compares favourably to Bayesian VARs and stochastic search variable selection in simulation and forecasting exercises. However, the authors again cannot demonstrate what kind of results the method would provide in structural impulse response analysis and in how far they would differ from other approaches to identification. The reason is again a lack of economic identification.

Augmenting these graphical approaches with sign restrictions will help to overcome this lack of policy-relevant insight. The two approaches are therefore complementary. Graphical identification often lacks structural economic insight, which can more easily be accomplished by relying on sign restrictions. Conversely, sign restrictions neatly pin down economic concepts, but lack precise identification. Improving that precision

may be achieved by adding exclusion restrictions gleaned from a statistical analysis of dependence patterns among the data. By relying on statistically motivated restrictions instead of purely conceptual ones, researchers may reduce the risk of imposing the answer that is sought in the first place. After all, as Uhlig (2005, p. 384) notes, "the answer to the key question [...] is often already substantially narrowed down by a priori theorizing."

3.3 Methods

In this section, I will briefly outline the two approaches used for identification. The first, sign restrictions, is well established in the literature. While the approach has been frequently applied, it is still under active discussion and development. The second part of this section will review the causal learning approach that will later be used for refining identification. It has seen few adoptions in the structural VAR literature and may therefore be less widely known. I will shortly discuss its foundations and build intuition.

Throughout, the focus lies on a standard structural VAR model

$$\boldsymbol{A}_0 \boldsymbol{y}_t = \sum_{p=1}^{P} \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t \tag{3.1}$$

which in reduced-form becomes

$$\boldsymbol{y}_t = \sum_{p=1}^{P} \boldsymbol{A}_p^* \boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \tag{3.2}$$

where $\boldsymbol{y}_t$ is a vector of K endogenous variables and the matrix $\boldsymbol{A}_0$ carries the contemporaneous effects between those variables and is normalised to have a unit diagonal. The ijth element of $\boldsymbol{A}_0$ is denoted as $-a_{ij}$ for $i \neq j$. The vector $\boldsymbol{\varepsilon}_t$ contains K independent structural errors with diagonal covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and the reduced-form residuals are given by $\boldsymbol{u}_t = \boldsymbol{A}_0^{-1}\boldsymbol{\varepsilon}_t$ such that its covariance matrix is $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}_0^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{A}_0^{-1'}$.

3.3.1 Sign Restrictions

Most of the literature by now follows Rubio-Ramírez, Waggoner, and Zha (2010) in drawing and imposing sign restrictions in structural VAR models. They focus on creating draws of the impact matrix $\boldsymbol{A}_0^{-1}$ via repeated rotations of the Cholesky decomposition of the reduced-form residual covariance matrix. This approach may be

implemented in a frequentist or Bayesian setting, but most authors opt for the latter, not least as inference in a Bayesian framework is easier to conduct.

In this study, the focus lies on imposing additional restrictions on elements of the contemporaneous effects matrix $\boldsymbol{A}_0$ as that is the object causal learning is informative about. The restrictions can be implemented in at least three ways. First, one may use the Bayesian methods of Baumeister and Hamilton (2015) to impose prior knowledge on $\boldsymbol{A}_0$ directly. Second, one may translate between the two specifications by working out any implications of restrictions on $\boldsymbol{A}_0$ for $\boldsymbol{A}_0^{-1}$. Third, Ouliaris and Pagan (2016) recently proposed a flexible frequentist framework for imposing sign restrictions on $\boldsymbol{A}_0$, while also allowing for restrictions on any transformations $f(\boldsymbol{A}_0, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_P, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}})$. For simplicity, I adopt the latter framework in this study.

The framework of Ouliaris and Pagan (2016) is inspired by the following observation. Without further restrictions on (3.1), the equation is underdetermined given any sample $\{\boldsymbol{y}_t\}$. The usual order condition states that at least $K(K-1)/2$ restrictions must be imposed for a unique solution. Ouliaris and Pagan (2016) suggest imposing the required number and the required kind of restrictions by drawing sufficiently many elements of $\boldsymbol{A}_0$ at random. Conditional on these draws, the remaining parameters can be estimated via instrumental variables and checked against the imposed sign restrictions on non-linear transformations of $\boldsymbol{A}_0$. Whenever the signs match, the draws are retained and discarded otherwise. The advantage of this approach is its simplicity, both conceptionally and computationally. However, as of now there are also two important downsides to this approach. One is the lack of proper inference, the other is a lack of attention to the implications of the sampling mechanism. Depending on how initial draws for the parameters are created, the final results of impulse response analysis may change. In the extreme case, if initial draws are too concentrated on one particular region, the algorithm may fail to effectively sample any valid parametrisations. In a less extreme case, it might discover admissible models but overweight some of them. This would come to bear on any attempt to summarise the results in a point estimator, since any measure of centrality will be affected by different sampling schemes. These drawbacks may be addressed in principle, but are not the focus of the present study. Nevertheless, the method is useful in the present context for illustrating the benefits of relying on additional identifying restrictions motivated through causal graph theory.

3.3.2 Causal Learning

To build intuition, imagine the following scenario. A large number of geographic land locations from around the globe is sampled at random. Each unit of observation carries

three measurements: latitude[2], altitude, and average temperature. With this setting, a discernible pattern between latitude and altitude is unlikely to emerge. That is because there is no mechanism or theory that would predict that altitude is likely to be very different because of a different latitude, or vice-versa. In this instance, the lack of causation is likely to coincide with a lack of correlation as well. Thus, knowing one measurement carries little information and therefore few implications for the other. But now condition on temperature. Knowing temperature, altitude does suddenly tell a whole lot about which latitude to expect. If it is on average 20° C and we are at 2000 m above sea level, it won't be close to either of the two poles. Why does knowing temperature suddenly shift expectations? In this case we have a clear understanding that both altitude and latitude have a causal impact on temperature, whereas the reverse does not hold: cooling the interior of an aircraft does not lift that aircraft to 30,000 ft above ground; flying to 30,000 ft does cool down the aircraft considerably.

This phenomenon can be paraphrased as *conditioning on a collider connects*. A collider is a node in a directed graph with two directed edges pointing towards that node. The edges essentially *collide* at that node. In the example above, temperature is a collider:

$$\text{latitude} \quad \rightarrow \quad \text{temperature} \quad \leftarrow \text{altitude}.$$

Here, we have encoded our understanding of the causal relations between the three variables with directed edges going from cause to effect. If the two causes—the two variables corresponding to the nodes without incoming edges—are not directly or indirectly linked in any other way, we observe the above pattern. Two independent variables become dependent on each other when conditioning on one of their joint effects. The causal learning technique employed later turns this logic on its head. It infers the direction of causation whenever two variables are uncorrelated, but become correlated once the relation is conditioned on a third variable.

Using graphical terminology is not essential for applying this logic, but it eases keeping track of structural implications and facilitates a non-technical summary. The same logic can also be illustrated in a regression framework, which might be more familiar to economists. Let us abstract from any dynamic effects for the moment and just consider three variables $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, y_{3,t})'$ for $t = 1, \dots, T$, where we do not know the structural relations among them. Then it is unclear whether we could interpret regression coefficients from regressing $y_{1,t}$ on $y_{2,t}$ as the effect an intervention in $y_{2,t}$ would have on $y_{1,t}$ since we do not know if there is any autonomous mechanism in place which would serve as a transfer of such an intervention. But we do know the following. First and foremost, estimating the full system $\boldsymbol{y}_t = (I - \boldsymbol{A}_0)\boldsymbol{y}_t + \boldsymbol{\nu}_t$ will not

[2]suitably scaled to reflect symmetry from the equator

be consistent because of a lack of identification, unless some elements of $\boldsymbol{A}_0$ are zero or otherwise restricted.

But suppose we postulate the DGP as being

$$y_{1,t} = \varepsilon_{1,t} \tag{3.3}$$

$$y_{2,t} = a_{21}y_{1,t} + a_{23}y_{3,t} + \varepsilon_{2,t} \tag{3.4}$$

$$y_{3,t} = \varepsilon_{3,t}, \tag{3.5}$$

where a_{21} and a_{23} are both fixed parameters unequal to zero and $\varepsilon_{i,t}$, $i = 1, 2, 3$, are uncorrelated white noise error terms, each following a normal distribution with variance σ_i^2. Just as above, this process can be summarised graphically as $y_{1,t} \rightarrow y_{2,t} \leftarrow y_{3,t}$. We can now run a sequence of regressions that will help falsify this null model. Let $\hat{\alpha}_{ij}$ be the regression coefficient from regressing $y_{i,t}$ on $y_{j,t}$ and $\hat{\beta}_{ij}$ the coefficient from regressing $y_{i,t}$ on $y_{j,t}$ while conditioning on $y_{k,t}$, $k \neq i, j$. First, if regressing $y_{1,t}$ on $y_{3,t}$ yields a coefficient estimate $\hat{\alpha}_{13}$ significantly different from zero, the model is likely not true. Second, if either $a_{21} = 0$ or $a_{23} = 0$ is found, the system would not appear to be true either. If we are willing to assume that there is no endogeneity bias towards zero that may explain results from the previous two steps, then we can suppose that running regression (3.4) would deliver consistent results.

Can these simple steps be validated any further? Can we, for example, delineate the above case from the case that

$$y_{1,t} = a_{12}y_{2,t} + \varepsilon_{1,t} \tag{3.6}$$

$$y_{2,t} = \varepsilon_{2,t} \tag{3.7}$$

$$y_{3,t} = a_{32}y_{2,t} + \varepsilon_{3,t}, \tag{3.8}$$

where a_{32} and a_{12} are such as to introduce an endogeneity bias towards zero that was previously assumed away in the aforementioned sequence of regressions? It appears there is a way to discriminate the latter case from the former. If regressing $y_{1,t}$ on $y_{3,t}$ while conditioning on $y_{2,t}$,

$$y_{1,t} = \beta_{13}y_{3,t} + \beta_{12}y_{2,t} + u_{1,t} \tag{3.9}$$

does not yield coefficient estimates $\hat{\beta}_{13}$ significantly different from zero, the model (3.3) to (3.5) is unlikely to be true. That is because adding $y_{2,t}$ will introduce an endogeneity bias under the null model since $\text{cov}(y_{2,t}, u_{1,t}) \neq 0$ and $\text{cov}(y_{2,t}, y_{3,t}) \neq 0$ under equations (3.3)–(3.5), but it will not introduce such a bias under (3.6)–(3.8). This latter case is precisely the 'collider connects' case described above.

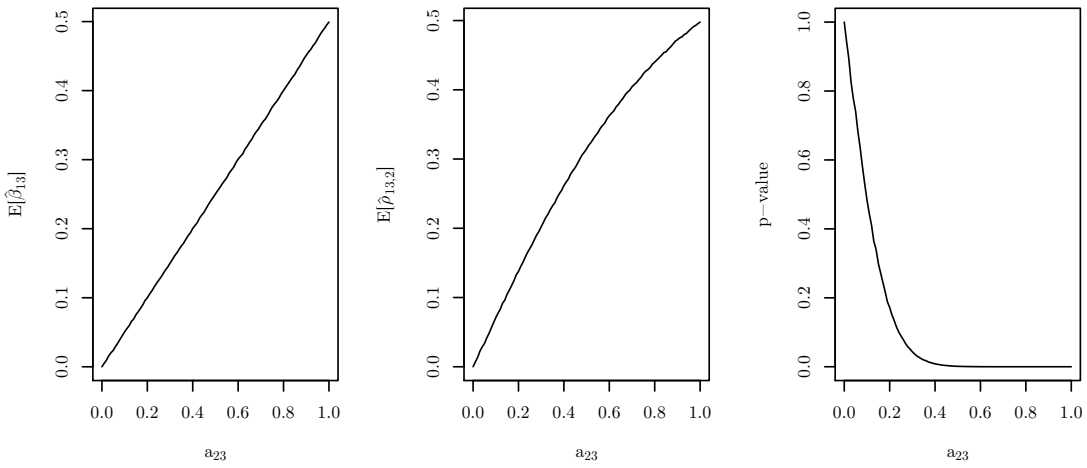

Figure 3.1: Average regression results for (3.9). Even though $y_{1,t}$ and $y_{3,t}$ are not directly related, conditioning on $y_{2,t}$ introduces an endogeneity bias such that $\mathrm{E}[\hat{\beta}_{13}] \neq 0$ whenever $a_{23} \neq 0$ (left panel). For symmetry, the relation between $y_{1,t}$ and $y_{2,t}$ is better judged using the correlation coefficient while conditioning on $y_{2,t}$ (middle panel). The p-value of a Fisher z-test of that partial correlation is to the right.

The phenomenon is illustrated in Figure 3.1. Letting (3.3)–(3.5) hold, setting $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $a_{21} = -1$ and letting a_{23} range from 0 to 1, the effect of different values of a_{23} on $\mathrm{E}[\hat{\beta}_{13}]$ is shown in the first plot to the left. Here, the expectation is $\mathrm{E}[\hat{\beta}_{13}] = a_{23}/2$. In general, regression coefficients are dependent on scale and interchanging regressand and regressor will typically yield different regression coefficients. Since the direction with which to run the regression is not known ex-ante, the partial correlation $\rho_{13,2}$ between $y_{1,t}$ and $y_{3,t}$ conditional on $y_{2,t}$ is a more natural measure in practice. The expected value $\mathrm{E}[\hat{\rho}_{13,2}]$ is plotted in the middle of Figure 3.1. The partial correlation coefficient is just an appropriate rescaling of $\hat{\beta}_{13}$ and in the present context will converge to $1/\sqrt{2}$ as $T \to \infty$ and $a_{23} \to \infty$.

In the following we will denote the unconditional correlation between $y_{i,t}$ and $y_{j,t}$ as ρ_{ij} and conditional on set $\{y_{k,t}\}, k \in \kappa \subseteq K \setminus \{i, j\}$, as $\rho_{ij,\kappa}$. Testing a null hypothesis of zero correlation $\mathrm{H}_0 : \rho = 0$ is conveniently done using Fisher's $z = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$. Under the null, and under suitable conditions on the data, $\sqrt{T}(\hat{z} - z) \xrightarrow{d} N(0, 1)$ as $T \to \infty$ (see Anderson, 2003, ch. 4). The p-value of testing the null hypothesis that $\rho_{13,2} = 0$ using Fisher's z is shown in the right most panel of Figure 3.1. As can be seen, the probability quickly diminishes to zero as a_{23} increases, but for small values of a_{23} of up to 0.1 the data is still in large agreement with $\rho_{13,2} = 0$.

But suppose we are able to reject $\rho_{13,2} = 0$, $\rho_{12} = 0$, and $\rho_{23} = 0$ in favour of $\rho_{13,2} \neq 0$, $\rho_{12} \neq 0$, $\rho_{23} \neq 0$ with sufficiently strong evidence, while willing to accept $\rho_{13} = 0$, then the conclusion that (3.3)–(3.5) hold would seem to be supported by the data. In

contrast, if we found that none of the conditional or unconditional correlations between $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$ are zero, no information about the structural relations between these variables would be gained, except that they all seem directly related to each other. However, if some partial correlations turn out to be zero, say $\rho_{13,2} = 0$ while $\rho_{13} \neq 0$, $\rho_{12} \neq 0$, and $\rho_{23} \neq 0$, then we may conclude that any of these structural forms hold $y_{1,t} \rightarrow y_{2,t} \rightarrow y_{3,t}$ or $y_{1,t} \leftarrow y_{2,t} \leftarrow y_{3,t}$ or $y_{1,t} \leftarrow y_{2,t} \rightarrow y_{3,t}$.

These considerations lie at the heart of the algorithm developed by Spirtes, Glymour, and Scheines (2000) and will be used to shed light on VARs in Section 3.4 by inferring structural relations among a set of endogenous variables. Note that only one of these correlation patterns implies a unique structure: the 'collider' case discussed above. That is the only instance in which a relation between two variables can be orientated in one direction. In other cases a direct connection between two variables may be removed, but no information on edge orientation is obtained.

These test decisions necessitate a set of assumptions. First among these assumptions is the causal Markov condition. This condition is fulfilled if effect z_t does in fact become independent of any 'deeper' causes x_t when conditioned on its direct cause(s) $\mathcal{Y}_t$. In more formal terms, $x_t \perp\!\!\!\perp z_t \mid \mathcal{Y}_t$ whenever for every y_t for which $x_t \rightarrow \cdots \rightarrow y_t \rightarrow z_t$ or $x_t \leftarrow \cdots \leftarrow y_t \rightarrow z_t$, we have $y_t \in \mathcal{Y}_t$, and there are no other (direct) paths from x_t to z_t that do not involve an element of $\mathcal{Y}_t$.[3] Second, it is assumed that the system under scrutiny is causally sufficient. That is, there is no latent variable that affects more than one variable at the same time. Third, the 'stability' or 'faithfulness' condition is fulfilled, which demands that an edge is absent $x_t \neq z_t$ only if $x_t \perp\!\!\!\perp z_t \mid \mathcal{Y}_t$ for some set of nodes $\mathcal{Y}_t$ that are in graph G, possibly an empty set. In particular, this assumption excludes counteracting effects such that $x_t \rightarrow y_t \rightarrow z_t$ and $x_t \rightarrow z_t$ will produce a zero net effect of x_t on z_t. It rules out an endogeneity bias in population precisely towards zero. Fourth, the system is linear such that stochastic dependence may actually be measured by correlation coefficients. Fifth, it is often assumed that the system is recursive. Assuming recursiveness is helpful for orientating edges as some logical conclusions about edge direction can be drawn from certain statistical test decisions. However, for structural VAR analysis that may already be too strong an assumption. The recursiveness assumption can be weakened if the stability condition mentioned above still holds in non-recursive systems, i.e. feedback loops do not cancel each other out completely. Although even that assumption will be contentious in certain cases. Take, for example, quantity and price. It is easily imagined that the correlation between these two variables is close to zero, simply because changes in

[3]Note that $\mathcal{Y}_t$ may contain other variables besides direct causes of z_t for this result to hold. However, specific variables, for example certain colliders or descendants thereof, *must not* be part of $\mathcal{Y}_t$.

quantity and price are driven by a series of demand and supply shifters in such a way that, overall, an evenly scattered cloud of measurement points is produced.

Another contentious issue is the way that causal graph methods rely on null hypothesis significance testing. For causal graph algorithms to work, some null hypotheses need to be *accepted* even though the reason the hypothesis was not rejected may just have been low power or bad luck. Chance is of course always at play in test decisions. But basing structural inferences on a few non-rejections may be too shaky a ground. In addition, most inference algorithms set a fixed threshold for the significance level and automate the search. It is difficult to see why a hypothesis that can be rejected with $p = 0.099$ establishes a direct causal link between two variables, whereas another hypothesis with $p = 0.11$ does not. For that reason this paper will not set a fixed threshold, but instead judge the evidence the data provides in conjunction with domain knowledge about possible structural relations.

3.4 Results From a Small-Scale Crude Oil Market VAR

This section will discuss details on the identification of oil market VARs and showcase the advantage of including graphically inspired sample information in a VAR analysis. Starting with Kilian (2009), several studies have explored the macroeconomic dynamics of the global oil market through vector autoregressions. Kilian (2009) and Kilian and Murphy (2012), for instance, estimate a three-variable oil market VAR covering monthly measures of global real activity, global crude oil production, and the global real price of crude oil. Both papers identify three types of structural shocks: an oil supply shock, an aggregate demand shock, and an oil-specific demand shock. The difference between the two papers is that Kilian (2009) identifies the structural model using short-run exclusion restrictions, whereas Kilian and Murphy (2012) employ sign restrictions. While the latter approach has the advantage of imposing less stringent assumptions, the authors stress that sign restrictions by themselves are insufficient to reach meaningful conclusions about the economic response to structural innovations. Instead, Kilian and Murphy (2012) emphasise that identification can be sharpened if additional identifying information is taken into account. This additional information takes the form of a bound $\xi_s < 0.0258$ on the price elasticity of oil supply. With this addition, the identified set of models imply economic dynamics similar to those identified in Kilian (2009). Kilian and Murphy (2014) extend the approach further by including a fourth variable on oil inventories. This allows them to investigate the effects of another type of oil demand shock born out of forward-looking, speculative motives.

The work of Kilian and Murphy has been fundamental to the study of oil markets and has sparked further debate among economists. It has also served several papers promoting methodological innovations in SVAR analysis as a case study. Among them are Lütkepohl and Netšunajev (2014) and Herwartz and Plödt (2016), who propose two distinct ways of exploiting statistical properties of the data for identification. Lütkepohl and Netšunajev (2014) identify their structural VAR oil market model through changes in the volatility of the reduced-form residuals. Equipped with an identification scheme that reflects properties of the data by construction, they are able to confirm the validity of the exclusion restrictions of Kilian (2009) and to a large extent also the validity of the sign restrictions of Kilian and Murphy (2012). Herwartz and Plödt (2016) take a slightly different approach and focus on the non-Gaussianity of the reduced-form residuals. With the residuals being non-normally distributed, there is, in principle, at most one possible rotation of the error vector that will make its elements stochastically independent. The authors use this fact and search for the linear combination of reduced-form errors for which the null hypothesis of independence is hardest to reject by an independence test. That linear combination is their estimate of the structural errors. Even though the data set is somewhat different, the qualitative results of the original study are largely confirmed also by this approach. Demand shocks, for instance, emerge as the main driver of oil price changes.

More recently, Baumeister and Hamilton (2019) and Zhou (2020) review and refine the global oil market model of Kilian and Murphy (2014), largely corroborating the original results though with important qualifications. Baumeister and Hamilton (2019) impose a Bayesian setting in which 'dogmatic priors' with strict exclusion and inequality restrictions may be relaxed. When the short-run price elasticity of oil supply is allowed to exceed 0.0258, for instance, the authors find that the elasticity is 'considerably larger' with a posterior median of 0.15. With these less restrictive bounds and a few other modifications, Baumeister and Hamilton (2019) find that supply shocks were more important in certain historical episodes, in contrast to findings in Kilian (2009) and Kilian and Murphy (2012), but largely in agreement with Kilian and Murphy (2014). Zhou (2020) studies the robustness of Kilian and Murphy (2014) to another range of modifications. Amongst other things, the author relaxes the elasticity bound to 0.04, which exceeds point estimates from the literature "by about four standard errors," extends the sample length and addresses an earlier error in constructing the index of global real economic activity. Extending the data to 2018 yields a greater role for oil supply shocks in explaining movements of the real price of oil, but conclusions by Kilian and Murphy (2014) about earlier historical episodes remain largely unaffected.

Since this is such a well studied case, I will follow suit by replicating the findings of Kilian and Murphy and by comparing the difference in outcomes when including statistical information prompted by causal graph theory as a means of identification. I use the data set of Kilian and Murphy (2014) for comparability. However, as a first step I focus on the three-variable VAR model of Kilian and Murphy (2012). They collect data on global crude oil production (in log-differences), global real economic activity (de-trended level), and the real price of crude oil (log-level) in the vector $\boldsymbol{y}_t$ and estimate via ordinary least squares (OLS) a VAR(24) model

$$\boldsymbol{y}_t = \boldsymbol{\nu} + \sum_{p=1}^{24} \boldsymbol{A}_p^* \boldsymbol{y}_{t-p} + \boldsymbol{u}_t. \tag{3.10}$$

Furthermore, $\boldsymbol{u}_t = \boldsymbol{A}_0^{-1}\boldsymbol{\varepsilon}_t$ where $\boldsymbol{\varepsilon}_t$ is assumed to be a white-noise error term with diagonal covariance matrix and the signs of the elements of $\boldsymbol{A}_0^{-1}$ are restricted as in Table 3.1. With these impact sign restrictions in place, the elements of $\boldsymbol{\varepsilon}_t$ can be interpreted as a supply shock, an aggregate demand and an oil-specific demand shock. They use the algorithm of Rubio-Ramírez, Waggoner, and Zha (2010) for implementing these restrictions. The structural models produced by the algorithm are ex-ante equally likely, yet still imply a range of different economic dynamics. Therefore, Kilian and Murphy (2012) put forward a number of economic arguments for restricting the model set further. They impose the upper bound of 0.0258 on the short-run price elasticity of oil supply and a lower bound of -1.5 on the immediate reaction of economic activity to an oil-specific demand shock.

In the following I will estimate the same reduced-form VAR model using a lag length of 24 but drawing the structural models with a different algorithm. As a baseline, no additional restrictions are imposed besides the sign restrictions listed in Table 3.1. As such, the model is not uniquely identified. However, following Ouliaris and Pagan (2016), we may identify the model repeatedly as described in Section 3.3 by drawing a sufficient number of parameters at random such that at each iteration the model can be solved conditional on these parameter draws. With sufficiently many draws, this approach will explore the range of parametrisations that satisfy the sign restrictions.

In the baseline case $K(K-1)/2 = 3$ parameters need to be drawn at random. Let $-a_{ij}$ again be the ijth element of $\boldsymbol{A}_0$ in (3.1). Then the values of a_{12}, a_{13} and a_{21} are obtained in the following way. For each parameter, $\theta_{ij} \sim \mathrm{U}(-1, 1)$ is drawn and transformed via $a_{ij} = \theta_{ij}/(1 - |\theta_{ij}|)$. The distribution function of a_{ij} thus obtained has support ranging from $-\infty$ to $+\infty$, is symmetric, has its mode at zero and displays considerably thicker tails than even the Cauchy distribution does. The probability of drawing values below -10 or above $+10$, for instance, is roughly 10 %. Widening

Table 3.1: Sign restrictions of structural impact effects as used by Kilian and Murphy (2012).

	supply	aggregate demand	oil-specific demand
oil production	−	+	+
global activity	−	+	−
oil price	+	+	+

those thresholds by a factor of ten will lower that probability by the same factor. Thus, on average, one a_{ij} draw in a million will be above one million in absolute terms. Whether this distribution is an effective way to explore the parameter space for admissible parametrisations depends very much on the data. Standardising the mean and variance of the data may improve the effectiveness and in some cases the signs of a_{ij} may also be inferred from sign restrictions on the elements of $\boldsymbol{A}_0^{-1}$. Conditional on the draws for a_{12}, a_{13}, and a_{21}, the remaining coefficients are estimated using instrumental variables. For details see Ouliaris and Pagan (2016). Whilst IV estimation does not account for the heteroskedasticity and non-normality of the residuals which has been documented by Lütkepohl and Netšunajev (2014) and Herwartz and Plödt (2016), it can be seen as approximating the DGP well enough, just as least squares commonly does in estimating the reduced-form VAR.

A first glimpse of the results is given in Figure 3.2. The figure shows impulse response functions of 2000 models that fulfil the sign restrictions of Kilian and Murphy (2012) listed in Table 3.1. As the figure vividly illustrates, little can be concluded about the interplay of oil markets and economic activity apart from the fact that every variable responds on impact to each of the shocks in line with the previously imposed sign. The first column displays the response of oil production, activity, and oil price to an oil supply shock. That type of shock is assumed to disrupt production and therefore has by construction a negative on impact effect on production and activity, while increasing prices. An aggregate demand shock (middle column), on the other hand, increases all three variables on impact. Finally, a demand shock that is specific to oil markets (right column) is assumed to drive up production, but to lower economic activity on impact through an adverse oil price increase. For every impulse response function, there seems to be a bound on the range of admissible parametrisations, which is either implied by the data or due to the fact that the algorithm has sampled a limited number of models. Further increasing the number of model draws does not change this conclusion, however. Another visible feature is that some response functions cross the zero line, while others don't. However, whether those functions that are far away from the zero

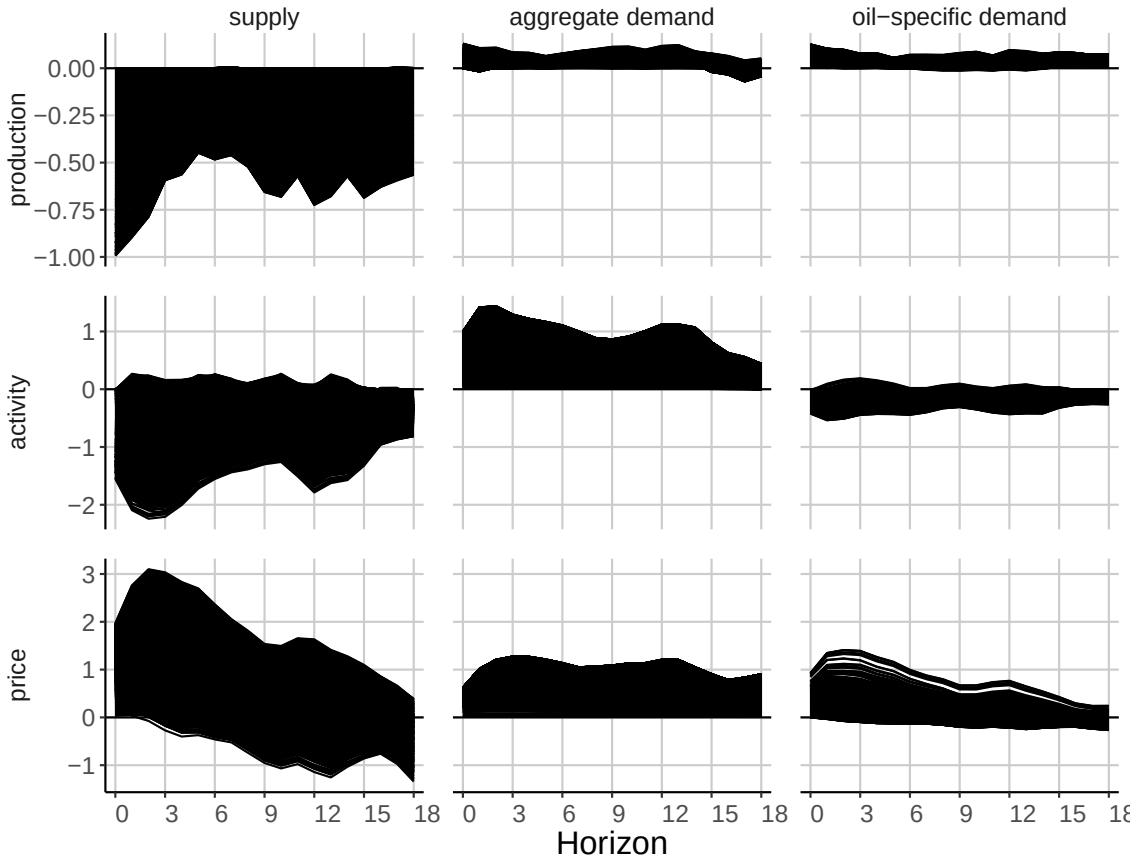

Figure 3.2: Impulse response functions of 2000 models fulfilling the sign restrictions of Kilian and Murphy (2012) and sampled with the algorithm of Ouliaris and Pagan (2016).

line or those which cross zero actually resemble the facts cannot be concluded from this analysis. All model draws are equally admissible.

The results are far more telling if we include information gained during the graphical pre-analysis. This pre-analysis is summarised in Figure 3.3, which indicates that the contemporaneous link between oil production and global economic activity is very weak. Both graphs in Figure 3.3 have three nodes that correspond to the three reduced-form residuals of oil production (prd), oil price (prc), and global activity (act). The edges of the left graph correspond to unconditional correlations, whereas the edges of the right graph correspond to correlation coefficients when conditioning on the remaining variable. The width of the edges have been weighted with the corresponding (partial) correlation coefficient. The higher the correlation in absolute value, the wider is the edge. The colour of each edge reflects the corresponding p-value, which has also been printed as an edge label. Edges with p-values equal to one are white, whereas those with p-values equal to zero are black.

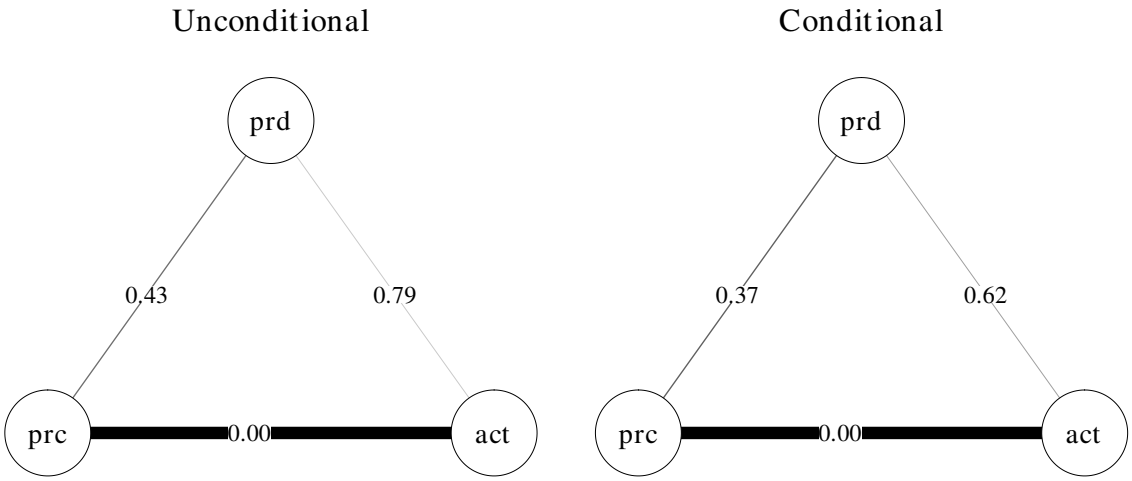

Figure 3.3: Association between reduced-form residuals of a VAR(24) using the Kilian and Murphy (2014) data set. The nodes correspond to oil production (prd), oil prices (prc) and global economic activity (act), which have all been filtered of autocorrelation. The edges represent contemporaneous correlations between nodes. The edge width is proportional to the absolute value of the correlation coefficient. Both the edge colour and the edge label indicate p-values (rounded to two decimal places). Edges in the left graph represent those unconditional correlations. Edges in the right graph represent conditional correlations.

As discussed in Section 3.3, the null hypotheses of zero correlation have been tested using Fisher's z-transformation. To account for the heteroskedasticity of the data, the p-values of those tests have been bootstrapped using the residual-based moving block bootstrap. Amongst other things, Brüggemann, Jentsch, and Trenkler (2016) show that this kind of block bootstrap is capable of estimating the variance of the unconditional residual covariance matrix of a VAR consistently even if the residuals are conditionally heteroskedastic. The bootstrap therefore also allows valid asymptotic inference on elements of the covariance matrix under these conditions. The results extend to continuously differentiable functions of the covariance matrix and therefore also apply to partial correlations and Fisher's z. This technique seems appropriate since models of conditional heteroskedasticity have been shown to fit data on oil markets rather well (see Lütkepohl and Netšunajev, 2014).

Let R be the number of bootstrap samples and let $\hat{z}_r^*$ denote the rth bootstrap replicate of Fisher's z. In the bootstrap world, the null hypothesis of zero correlation is imposed by a level shift $\hat{z}_r^* - \hat{z}$. This is analogous to the construction of Hall's percentile interval[4] that is commonly used in impulse response analysis. A particular condition required for this approach to work is 'translational invariance' of the null

[4]sometimes also know as the 'basic bootstrap confidence interval'

distribution, i.e. the distribution is not affected beyond the first moment by a level shift (see Efron and Tibshirani, 1993, chap. 16). The p-value is then estimated as $p = \frac{1}{R}\sum_{r=1}^{R} I(|\hat{z}| \leq |\hat{z}_r^* - \hat{z}|)$, where $I()$ is the indicator function.

Equipped with these p-values, the graphs in Figure 3.3 succinctly illustrate how the contemporaneous association between oil price and economic activity is fairly strong, while the correlation measures for quantity vs. price and quantity vs. activity are rather low. As previously discussed, a low correlation between quantity and price need not come as a surprise since supply and demand shocks could have negating effects. The low correlation with a p-value of 0.79 between quantity and activity, on the other hand, may indeed reflect the absence of any direct contemporaneous causal effects as both entities are prone to sluggish behaviour. Conditioning on a third variable reduces the p-values slightly, as can be seen on the right of Figure 3.3. However, the reduction is not large enough to justify the orientation of any edges. All in all, the impulse response analysis will be repeated with the additional qualification that the quantity of oil production and global activity cannot influence each other directly within the month. Now, the values of a_{12} and a_{21} are both set to zero and only a_{13} is drawn at random to explore the model space. These two restrictions are the gain from bringing in insights from causal learning.

Figure 3.4 shows the impulse response functions after imposing these zero short-run restrictions motivated by the inspection of (conditional) correlations between reduced-form residuals. Again 2000 models fulfilling the sign restrictions were drawn and their impulse response functions computed. With two zeros imposed on the contemporaneous effects matrix $\boldsymbol{A}_0$, there are now four unknown parameters left. While the structural model is therefore still not uniquely identified, only one parameter, in this case a_{13}, needs to be drawn at random.

With this truncation of the sampling space, the dynamics are now much more clear-cut. The level of oil production shows a pronounced and lasting decline in response to an oil supply disruption, but hardly any response to either of the two demand shocks. The second row of Figure 3.4 indicates a short-lived, negative response of global economic activity to an oil supply disruption after one quarter and a strong and fairly persistent response to an aggregate demand shock. The development of real activity induced by an oil-specific demand shock is unclear, with many response functions straddling the zero line. Finally, the oil price seems to react positively and persistently to all three shocks.

While these figures do not indicate the sampling uncertainty that is due to a finite sample size and therefore do not allow proper inference, the results are remarkably similar to the findings of Kilian (2009) and Kilian and Murphy (2012). This may not be

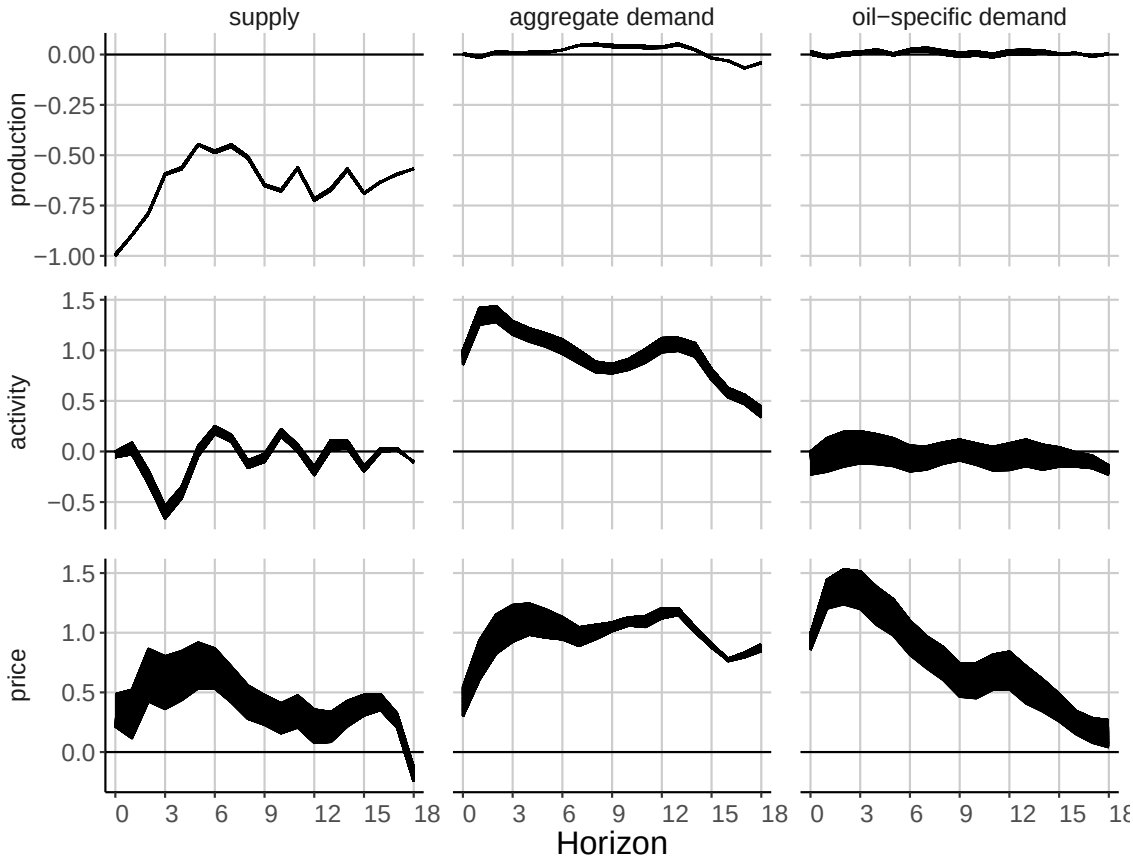

Figure 3.4: Impulse response functions of 2000 models fulfilling the sign restrictions of Kilian and Murphy (2012) and two additional short-run exclusion restrictions between oil production and economic activity.

that surprising, since the imposed restrictions almost combine a strict subset of both studies. Nonetheless, these results are insofar reassuring as they promise an avenue for more precise identification in the absence of widely accepted economic information that may sharpen otherwise loose sign restrictions. In this case, combining some of the implied zeros derived from a graph-theoretical analysis with widely accepted sign restrictions therefore proves to be a viable approach to recover informative and economically meaningful structures.

3.5 Conclusion

This study has explored the potential for structural VAR analysis that combines sign restrictions with a causal graph perspective. The graphical approach rests on insights from the literature on causal learning and infers structural relations among a set of variables inductively, based on a set of assumptions and on properties of the data. In this study, these properties are simple correlation patterns. The data-dependent

approach is of particular interest if other identifying information is either not available or contentious.

Applying this combination to an oil market VAR illustrates how identification can be considerably sharpened. Without the additional short-run exclusion restrictions motivated by correlation patterns, the sign restrictions by themselves are far too loose for any meaningful conclusions. What is more, the identified set of impulse response functions explicitly agrees with VARs identified by other means. The results also enhance usual graphical modelling approaches, for structural shocks would not have been clearly identified in a purely graphical analysis. Moreover, full identification is not guaranteed with the graphical causal learning approach.

These findings encourage more research. In particular, the method may prove especially helpful in larger systems, where meaningful identification through sign restrictions can be even more challenging, even if other prior domain knowledge is available. Secondly, this article has not touched upon inference. Inferential procedures for set-identified VARs estimated with frequentist methods are only beginning to bear fruit. The methods developed by Granziera, Moon, and Schorfheide (2018) are particularly relevant in this regard. Ideally, these inferential methods would also include the uncertainty present in the graphical pre-analysis. Alternatively, one may adopt a Bayesian framework and use the results in Ahelegbey, Billio, and Casarin (2016) in combination with sign restrictions for sharper identification and proper inference. One drawback of that method is, however, that only recursive systems are allowed, whereas this assumption has not been made here. Third, there is a host of other parametric and nonparametric independence tests that relax some of the assumptions required for simple correlation tests. Fourth, advancements in the literature on causal learning include methods that allow for latent variables or feedback loops more explicitly. See Heinze-Deml, Maathuis, and Meinshausen (2018) for a review. These developments may also be worthwhile to incorporate in structural VAR analysis.

Bibliography

Ahelegbey, D.F., Billio, M., and Casarin, R. (2016). Bayesian graphical models for structural vector autoregressive processes. *Journal of Applied Econometrics* **31**(2): 357–386.

Aiolfi, M., Capistrán, C., and Timmermann, A. (2011). Forecast combinations. *In* M.P. Clements and D.F. Hendry (eds.), *The Oxford Handbook of Economic Forecasting*, Oxford University Press, chap. 12, pp. 355–390.

Aka, N. (2014). *Model Selection Sets: Theory, Simulations and an Application.* Master thesis, University of Regensburg.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In* B.N. Petrov and F. Csaki (eds.), *2nd International Symposium of Information Theory*, Akadémiai Kiadó, Budapest, pp. 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716–723.

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Hoboken, NJ, USA.

Angrist, J.D. and Pischke, J.S. (2017). Undergraduate econometrics instruction: Through our classes, darkly. *Journal of Economic Perspectives* **31**(2): 125–44.

Arefiev, N. (2014). A theory of data-oriented identification with a SVAR application. *Working Paper Higher School of Economics Research Paper No. WP BRP 79/EC/2014*, National Research University Higher School of Economics.

Arefiev, N. (2016a). Graphical interpretations of rank conditions for identification of linear Gaussian models. *Working Paper Higher School of Economics Research Paper No. WP BRP 124/EC/2016*, National Research University Higher School of Economics.

Arefiev, N. (2016b). Identification of monetary policy shocks within a SVAR: Using restrictions consistent with a DSGE model. *Working Paper Higher School of Economics Research Paper No. WP BRP 125/EC/2016*, National Research University Higher School of Economics.

Arias, J.E., Rubio-Ramírez, J.F., and Waggoner, D.F. (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. *Econometrica* **86**(2): 685–720.

Athey, S. and Imbens, G.W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics* **11**(1): 685–725.

Basmann, R.L. (1963). The causal interpretation of non-triangular systems of economic relations. *Econometrica* **31**(3): 439–448.

Basmann, R.L. (1965). A note on the statistical testability of 'explicit causal chains' against the class of 'interdependent' models. *Journal of the American Statistical Association* **60**(312): 1080–1093.

Baumeister, C. and Hamilton, J.D. (2015). Sign restrictions, structural vector autoregressions, and useful prior information. *Econometrica* **83**(5): 1963–1999.

Baumeister, C. and Hamilton, J.D. (2019). Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks. *American Economic Review* **109**(5): 1873–1910.

Baumeister, C. and Hamilton, J.D. (2020). Drawing conclusions from structural vector autoregressions identified on the basis of sign restrictions. *Working Paper 26606*, National Bureau of Economic Research.

Beebee, H., Hitchcock, C., and Menzies, P. (eds.) (2009). *The Oxford Handbook of Causation*. Oxford University Press, Oxford, UK.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**(2): 521–547.

Breiman, L. (1996). Stacked regressions. *Machine Learning* **24**(1): 49–64.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*. Springer, New York, USA, 2nd edn.

Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*. Springer, New York, USA, 2nd edn.

Brüggemann, R., Jentsch, C., and Trenkler, C. (2016). Inference in VARs with conditional heteroskedasticity of unknown form. *Journal of Econometrics* **191**(1): 69–85.

Bruns, M. and Piffer, M. (2019). Bayesian structural VAR models: A new approach for prior beliefs on impulse responses. *Discussion Papers 1796*, DIW Berlin, German Institute for Economic Research.

Bryant, H.L., Bessler, D.A., and Haigh, M.S. (2006). Causality in futures markets. *Journal of Futures Markets* **26**(11): 1039–1057.

Bryant, H.L., Bessler, D.A., and Haigh, M.S. (2009). Disproving causal relationships using observational data. *Oxford Bulletin of Economics and Statistics* **71**(3): 357–374.

Canova, F. and De Nicoló, G. (2002). Monetary disturbances matter for business fluctuations in the G-7. *Journal of Monetary Economics* **49**(6): 1131–1159.

Charkhi, A. and Claeskens, G. (2018). Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* **105**(3): 645–664.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**(1): C1–C68.

Chickering, D.M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* **2**: 445–498.

Christiano, L.J., Eichenbaum, M., and Evans, C.L. (1999). Monetary policy shocks: What have we learned and to what end? *In* J.B. Taylor and M. Woodford (eds.), *Handbook of Macroeconomics*, Elsevier, vol. 1 of *Handbook of Macroeconomics*, chap. 2, pp. 65–148.

Christiano, L.J., Eichenbaum, M., and Evans, C.L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* **113**(1): 1–45.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, UK.

Colombo, D. and Maathuis, M.H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* **15**(116): 3921–3962.

Cox, D. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. Chapman & Hall/CRC, Boca Raton, FL, USA.

Demetrescu, M., Hassler, U., and Kuzin, V. (2011). Pitfalls of post-model-selection testing: experimental quantification. *Empirical Economics* **40**(2): 359–372.

Demiralp, S., Hoover, K., and Perez, S. (2014). Still puzzling: evaluating the price puzzle in an empirically identified structural vector autoregression. *Empirical Economics* **46**(2): 701–731.

Demiralp, S. and Hoover, K.D. (2003). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics* **65**: 745–767.

Demiralp, S., Hoover, K.D., and Perez, S.J. (2008). A bootstrap method for identifying and evaluating a structural vector autoregression. *Oxford Bulletin of Economics and Statistics* **70**(4): 509–533.

DiCiccio, C.J. and Romano, J.P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association* **112**(519): 1211–1220.

Diestel, R. (2017). *Graph Theory*. Springer, Berlin, Germany.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY, USA.

Elliot, G. and Timmermann, A. (2016). *Economic Forecasting*. Princeton University Press, Princeton, New Jersey, USA.

Elwert, F. (2013). Graphical causal models. *In* S.L. Morgan (ed.), *Handbook of Causal Analysis for Social Research*, Springer, Dordrecht, Netherlands, chap. 13, pp. 245–273.

Engle, R.F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *In* Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, Elsevier, vol. 2 of *Handbook of Econometrics*, chap. 13, pp. 775 – 826.

Faust, J. (1998). The robustness of identified VAR conclusions about money. *Carnegie-Rochester Conference Series on Public Policy* **49**: 207–244.

Fragetta, M. and Melina, G. (2011). The effects of fiscal policy shocks in SVAR models: A graphical modelling approach. *Scottish Journal of Political Economy* **58**(4): 537–566.

Fragetta, M. and Melina, G. (2013). Identification of monetary policy in SVAR models: a data-oriented perspective. *Empirical Economics* **45**(2): 831–844.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1–22.

Fry, R. and Pagan, A. (2011). Sign restrictions in structural vector autoregressions: a critical review. *Journal of Economic Literature* **49**: 939–960.

Gafarov, B., Meier, M., and Montiel Olea, J.L. (2018). Delta-method inference for a class of set-identified SVARs. *Journal of Econometrics* **203**(2): 316–327.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**(3): 424–438.

Granziera, E., Moon, H.R., and Schorfheide, F. (2018). Inference for VARs identified with sign restrictions. *Quantitative Economics* **9**(3): 1087–1121.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ, USA.

Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**(2): 190–195. 190–195.

Hansen, B. (2016). The risk of James-Stein and lasso shrinkage. *Econometric Reviews* **35**(8–10): 1456–1470.

Hansen, B.E. and Racine, J.S. (2012). Jackknife model averaging. *Journal of Econometrics* **167**: 38–46.

Hansen, P.R., Lunde, A., and Nason, J.M. (2011). The model confidence set. *Econometrica* **79**: 453–497.

Hartsfield, N. and Ringel, G. (2003). *Pearls in Graph Theory: A Comprehensive Introduction*. Dover Publications, Mineola, NY, USA.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, New York, NY, USA.

Hayfield, T. and Racine, J.S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software* **27**(5): 1–32.

Heinze-Deml, C., Maathuis, M.H., and Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application* **5**(1): 371–391.

Hendry, D.F. and Krolzig, H.M. (2004). Resolving three 'intractable' problems using a Gets approach. Manuscript, Oxford University.

Hendry, D.F. and Krolzig, H.M. (2005). The properties of automatic Gets modelling. *The Economic Journal* **115**(502): C32–C61.

Henschen, T. (2018). What is macroeconomic causality? *Journal of Economic Methodology* **25**(1): 1–20.

Hernán, M.A. and Robins, J.M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, USA.

Herwartz, H. and Plödt, M. (2016). The macroeconomic effects of oil price shocks: Evidence from a statistical identification approach. *Journal of International Money and Finance* **61**: 30–44.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**(396): 945–960.

Hoover, K.D., Demiralp, S., and Perez, S.J. (2009). Empirical identification of the vector autoregression: The causes and effects of US M2. *In* J. Castle and N. Shephard (eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, UK, chap. 2, pp. 37–58.

Humphreys, P. and Freedman, D. (1996). The grand leap. *The British Journal for the Philosophy of Science* **47**(1): 113–123.

Imbens, G. (2019). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Working Paper 26104*, National Bureau of Economic Research.

Imbens, G.W. and Rubin, D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge, UK.

Inoue, A. and Kilian, L. (2013). Inference on impulse response functions in structural VAR models. *Journal of Econometrics* **177**(1): 1–13.

Inoue, A. and Kilian, L. (2016). Joint confidence sets for structural impulse responses. *Journal of Econometrics* **192**(2): 421–432.

Inoue, A. and Kilian, L. (2020). Joint Bayesian inference about impulse responses in VAR models. Manuscript, Vanderbilt University.

Jinjarak, Y. and Sheffrin, S.M. (2011). Causality, real estate prices, and the current account. *Journal of Macroeconomics* **33**(2): 233–246.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8**: 613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* **47**(1): 1–26.

Kascha, C. and Trenkler, C. (2015). Forecasting VARs, model selection, and shrinkage. *Working Paper 15-07*, University of Mannheim / Department of Economics.

Keuzenkamp, H.A. (2004). *Probability, Econometrics and Truth: The Methodology of Econometrics*. Cambridge University Press, Cambridge, UK.

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review* **99**(3): 1053–69.

Kilian, L. and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press, Cambridge, UK.

Kilian, L. and Murphy, D.P. (2012). Why agnostic sign restrictions are not enough: Understanding the dynamics of oil market VAR models. *Journal of the European Economic Association* **10**(5): 1166–1188.

Kilian, L. and Murphy, D.P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics* **29**(3): 454–478.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models*. MIT Press, Cambridge, MA, USA.

Koopmans, T.C. (1949). Identification problems in economic model construction. *Econometrica* **17**(2): 125–144.

Kumor, D., Chen, B., and Bareinboim, E. (2019). Efficient identification in linear structural causal models with instrumental cutsets. *In* H. Wallach, H. Larochelle, A. Beygelzimer, F. d,Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., pp. 12477–12486.

Kwon, D.H. and Bessler, D.A. (2011). Graphical methods, inductive causal inference, and econometrics: A literature review. *Computational Economics* **38**(1): 85–106.

Lanne, M., Meitz, M., and Saikkonen, P. (2017). Identification and estimation of non-Gaussian structural vector autoregressions. *Journal of Econometrics* **196**(2): 288–304.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, New York, NY, USA.

Leamer, E.E. (1983). Let's take the con out of econometrics. *The American Economic Review* **73**(1): 31–43.

Leamer, E.E. (1985). Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy* **22**: 255–304.

Leeb, H. and Pötscher, B.M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**: 21–59.

Lucas, R.E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* **1**: 19–46.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Verlag, Berlin, Germany.

Lütkepohl, H. and Netšunajev, A. (2014). Disentangling demand and supply shocks in the crude oil market: How to check sign restrictions in structural VARs. *Journal of Applied Econometrics* **29**(3): 479–496.

Lütkepohl, H. and Netšunajev, A. (2017). Structural vector autoregressions with heteroskedasticity: A review of different volatility models. *Econometrics and Statistics* **1**: 2–18.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**(428): 1535–1546.

Maziarz, M. and Mróz, R. (2019). Response to Henschen: causal pluralism in macroeconomics. *Journal of Economic Methodology* (forthcoming).

Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995)*. Montreal, Quebec, Canada, pp. 411–418.

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., Chlaß, N., Entner, D., and Hoyer, P. (2011). Causal search in structural vector autoregressive models. *In* F. Popescu and I. Guyon (eds.), *Proceedings of the Neural Information Processing Systems Mini-Symposium on Causality in Time Series*. PMLR, Vancouver, Canada, vol. 12, pp. 95–114.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys* **29**(1): 46–75.

Morgan, S.L. (ed.) (2013). *Handbook of Causal Analysis for Social Research*. Springer, Dordrecht, Netherlands.

Nguyen, L. (2019). Bayesian inference in structural vector autoregression with sign restrictions and external instruments. Manuscript, University of California at San Diego.

Ouliaris, S. and Pagan, A. (2016). A method for working with sign restrictions in structural equation modelling. *Oxford Bulletin of Economics and Statistics* **78**(5): 605–622.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, USA.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edn.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference*. MIT Press, Cambridge, MA, USA.

Robins, J.M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika* **90**(3): 491–515.

Robins, J.M. and Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge. *In* C. Glymour and G.F. Cooper (eds.), *Computation, Causation, and Discovery*, MIT Press, Cambridge, MA, USA, pp. 305–321.

Romano, J.P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* **100**(469): 94–108.

Rubio-Ramírez, J.F., Waggoner, D.F., and Zha, T. (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies* **77**(2): 665–696.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**(2): 461–464.

Simon, H.A. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association* **49**(267): 467–479.

Sims, C.A. (1972). Money, income, and causality. *The American Economic Review* **62**(4): 540–552.

Sims, C.A. (1977). Exogeneity and causal ordering in macroeconomic models. *In New Methods in Business Cycle Research: Proceedings from a Conference*. Federal Reserve Bank of Minneapolis, Minneapolis, MN, USA, pp. 23–43.

Sims, C.A. (1980). Macroeconomics and reality. *Econometrica* **48**(1): 1– 48.

Sims, C.A. and Zha, T. (2006). Does monetary policy generate recessions? *Macroeconomic Dynamics* **10**(2): 231–272.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, USA, 2nd edn.

Steiner, P.M., Kim, Y., Hall, C.E., and Su, D. (2017). Graphical models for quasi-experimental designs. *Sociological Methods & Research* **46**(2): 155–188.

Stock, J.H. and Watson, M.W. (2007). *Introduction to Econometrics*. Pearson/Addison Wesley, Boston, MA, USA.

Strotz, R.H. and Wold, H.O.A. (1960). Recursive vs. nonrecursive systems: An attempt at synthesis (part I of a triptych on causal chain systems). *Econometrica* **28**(2): 417–427.

Swanson, N.R. and Granger, C.W.J. (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.

Theil, H. (1957). Specification errors and the estimation of economic relationships. *Review of the International Statistical Institute* **25**(1/3): 41–51.

Timmermann, A. (2006). Forecast combinations. *In* G. Elliot, C.W. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Elsevier, vol. 1, chap. 4, pp. 135–196.

Turlach, B.A., Weingessel, A., and Moler, C. (2019). *quadprog: Functions to Solve Quadratic Programming Problems.* R package version 1.5-8.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics* **41**(2): 436–463.

Uhlig, H. (2005). What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics* **52**(2): 381–419.

Uhlig, H. (2017). Shocks, sign restrictions, and identification. *In* B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson (eds.), *Advances in Economics and Econometrics: Eleventh World Congress*, Cambridge University Press, Cambridge, UK, vol. 2, chap. 4, pp. 95–127.

van der Zander, B., Textor, J., and Liśkiewicz, M. (2015). Efficiently finding conditional instruments for causal inference. *In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015).* AAAI Press, pp. 3243–3249.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. *In Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence.* Cambridge, MA, USA, pp. 220–227.

Wasserstein, R.L., Schirm, A.L., and Lazar, N.A. (2019). Moving to a world beyond "p<0.05". *The American Statistician* **73**(sup1): 1–19.

Weihs, L., Robinson, B., Dufresne, E., Kenkel, J., Kubjas, K., McGee II, R., Nguyen, N., Robeva, E., and Drton, M. (2018). Determinantal generalizations of instrumental variables. *Journal of Causal Inference* **6**(1): 1–21.

West, D.B. (2001). *Introduction to Graph Theory.* Pearson Education, London, UK.

White, H. (2000). A reality check for data snooping. *Econometrica* **68**(5): 1097–1126.

White, H. and Lu, X. (2010). Granger causality and dynamic structural systems. *Journal of Financial Econometrics* **8**(2): 193–243.

White, H. and Pettenuzzo, D. (2014). Granger causality, exogeneity, cointegration, and economic policy analysis. *Journal of Econometrics* **178**: 316–330.

Wold, H. (1954). Causality and econometrics. *Econometrica* **22**(2): 162–177.

Wolpert, D.H. (1992). Stacked generalization. *Neural Networks* **5**(2): 241–259.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**(7): 557–585.

Zellner, A. (1979). Causality and econometrics. *Carnegie-Rochester Conference Series on Public Policy* **10**: 9–54.

Zellner, A. (1988). Causality and causal laws in economics. *Journal of Econometrics* **39**(1): 7–21.

Zhang, J. and Spirtes, P. (2003). Strong faithfulness and uniform consistency in causal inference. *In Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI2003)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 632–639.

Zhang, X., Wan, A.T.K., and Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* **174**(2): 82–94.

Zhou, X. (2020). Refining the workhorse oil market model. *Journal of Applied Econometrics* **35**(1): 130–140.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorgelegte Dissertation auf Grundlage der angegebenen Quellen und Hilfsmittel selbstständig verfasst habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, sind als solche kenntlich gemacht. Die vorgelegte Dissertation hat weder in der gleichen noch einer anderen Fassung bzw. Überarbeitung einer anderen Fakultät, einem Prüfungsausschuss oder einem Fachvertreter an einer anderen Hochschule zum Promotionsverfahren vorgelegen.

Niels Aka
Berlin, den 31. August 2020

Liste verwendeter Hilfsmittel

- RStudio, R, CRAN-Pakete
- GNU/Linux und verwandte freie, quelloffene Programme
- LaTeX
- Verschiedene Intel und AMD Chipsets, Computer, Server, insbesondere die HPC (high-performance computing) Infrastruktur an der FU Berlin und am DIW Berlin.
- Siehe auch Literatur- und Quellenangaben

All den Menschen, die diese oftmals frei verfügbaren Hilfsmittel entwickelt und zur Verfügung gestellt haben, ebenso wie den vielen WissenschaftlerInnen, auf deren Forschung ich aufbaue, gebührt mein Dank.