1 Introduction

1.1 Overview

The vector autoregressive (VAR) model is a widely used model for multivariate time series analysis. It consists of a system of regression equations. VAR models are estimated by regressing each model variable on lags of its own as well as lags of the other model variables up to some prespecified maximum lag order, p. A VAR model with p autoregressive lags is referred to as a VAR(p) model. VAR models are based on the notion that every model variable depends on its own lags as well as the lags of every other model variable, rendering exclusion restrictions on the interaction of lagged model variables not credible.

VAR models are typically based on monthly or quarterly data. For example, let y_t denote a K-dimensional vector of time series data, consisting of U.S. real GNP growth (Δgnp_t), the U.S. rate of inflation (π_t), and the U.S. short-term nominal interest rate (i_t), for $t = 1, \ldots, T$. Then, suppressing the intercept, a quarterly VAR(2) model for these three variables may be written as a system of three equations

$$\begin{aligned}
\Delta gnp_t &= a_{11,1}\Delta gnp_{t-1} + a_{12,1}\pi_{t-1} + a_{13,1}i_{t-1} \\
&\quad + a_{11,2}\Delta gnp_{t-2} + a_{12,2}\pi_{t-2} + a_{13,2}i_{t-2} + u_{1t} \\
\pi_t &= a_{21,1}\Delta gnp_{t-1} + a_{22,1}\pi_{t-1} + a_{23,1}i_{t-1} \\
&\quad + a_{21,2}\Delta gnp_{t-2} + a_{22,2}\pi_{t-2} + a_{23,2}i_{t-2} + u_{2t} \\
i_t &= a_{31,1}\Delta gnp_{t-1} + a_{32,1}\pi_{t-1} + a_{33,1}i_{t-1} \\
&\quad + a_{31,2}\Delta gnp_{t-2} + a_{32,2}\pi_{t-2} + a_{33,2}i_{t-2} + u_{3t},
\end{aligned} \tag{1.1.1}$$

where $K = 3$ and the zero mean innovations u_{it}, $i = 1, 2, 3$, are serially uncorrelated if the maximum lag order has been chosen appropriately. The model allows these innovations to be mutually correlated with covariance matrix Σ_u. More compactly, we can express this VAR(2) model as

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + u_t, \tag{1.1.2}$$

where

$$y_t = \begin{pmatrix} \Delta gnp_t \\ \pi_t \\ i_t \end{pmatrix}, \quad A_i = \begin{bmatrix} a_{11,i} & a_{12,i} & a_{13,i} \\ a_{21,i} & a_{22,i} & a_{23,i} \\ a_{31,i} & a_{32,i} & a_{33,i} \end{bmatrix}, \; i = 1, 2,$$

$$u_t = \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix}.$$

The innovation vector u_t is the linearly unpredictable component of y_t, given an information set consisting of the lagged values of all three model variables.

In the language of the literature on dynamic simultaneous equations models such a model is known as a reduced form, defined as a model that expresses the current values of the data as a linear function only of its own lagged values and lagged values of the other model variables. The reduced-form VAR model may be viewed as a finite-order approximation to a general linear process. This model has proved useful for summarizing the properties of the data, for forecasting, for testing for the existence of equilibrium relationships tying together two or more economic variables, and for quantifying the speed with which the model variables revert back to the equilibrium following a disturbance.

This book instead focuses on the use of VAR models for structural modeling. The premise is that we can think of the reduced-form VAR(p) model as representing data generated from the structural VAR(p) model

$$B_0 y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + w_t, \tag{1.1.3}$$

where y_t is the $K \times 1$ vector of observed time series data, $t = 1, \ldots, T$, and the deterministic terms have been suppressed for convenience. Furthermore, B_i, $i = 1, \ldots, p$, is a $K \times K$ matrix of autoregressive slope coefficients, the $K \times K$ matrix B_0 reflects the instantaneous relations among the model variables, and the $K \times 1$ vector of mean zero structural shocks w_t is serially uncorrelated with a diagonal covariance matrix Σ_w of full rank such that the number of shocks coincides with the number of variables. The $K \times K$ matrix B_0^{-1} captures the impact effects of each of the structural shocks on each of the model variables. Model (1.1.3) is structural in that the shocks are postulated to be mutually uncorrelated with each element of w_t having a distinct economic interpretation. This fact allows one to interpret movements in the data caused by any one element of w_t as being caused by that shock. The structural shocks in general are not directly observable, but under suitable conditions they may be recovered from the reduced-form representation of model (1.1.3).

The reduced-form representation of model (1.1.3) can be obtained by premultiplying both sides of (1.1.3) by B_0^{-1}, resulting in the model

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{1.1.4}$$

where $A_i = B_0^{-1}B_i$ and $u_t = B_0^{-1}w_t$, and may be estimated by unrestricted least-squares (LS) or maximum likelihood (ML) estimation methods. Model (1.1.1) is easily recognized as an example of such a model. It is readily apparent that, given an estimate of this reduced form, all that is required for recovering the structural model (1.1.3) is knowledge of the structural impact multiplier matrix B_0^{-1} (or, equivalently, of its inverse B_0).

Estimation of the matrix B_0 requires additional restrictions on the data generating process (DGP). If the matrix B_0 can be solved for, given these restrictions and the data, we say that the structural VAR model parameters, $(B_0, B_1, \ldots, B_p, \Sigma_w)$, are identified or, equivalently, that the structural shocks $w_t = B_0 u_t$ are identified. The problem of finding suitable economically credible restrictions on B_0^{-1} or B_0 is known as the identification problem in structural VAR analysis. Much of this book is concerned with alternative strategies for achieving identification. As with any structural econometric model, the validity of structural VAR analysis rests on the credibility of these identifying restrictions. Finding a credible set of restrictions can be challenging. Depending on the identifying assumptions, the structural VAR model may or may not be unique. In the latter case, there is a range of structural models that are observationally equivalent in that they have the same reduced-form representation.

The existence of a structural VAR model allows us to think of the variation in the data as being driven by the cumulative effects of economically interpretable structural shocks. Current observations of the data may be viewed as a weighted average of current and past structural shocks. This insight is important because it helps researchers quantify causal relationships in the data that are obscured in reduced-form VAR analysis. For expository purposes and without loss of generality, suppose that the structural shock of interest involves changes in monetary policy not in response to macroeconomic conditions. After expressing the estimate of the VAR model in a suitable form, one may answer a range of questions about the causal effects of this shock.

- One may ask by how much an unexpected monetary policy tightening in the current quarter will reduce output growth over the next two years, when that policy change occurs all else equal and is not followed by any further monetary policy shocks after the current quarter. The response of output growth to this shock over time can be quantified in the form of an impulse response function.
- One may ask how much of the variability of output growth on average is accounted for by shocks to monetary policy as opposed to other structural shocks. This question can be answered by a forecast error variance decomposition.
- One may ask how much of the recession of 1982, for example, is explained by the cumulative effects of earlier monetary policy

shocks. This question can be answered by constructing a historical decomposition.

- One may also ask by how much the recession of 1982 would have deepened had monetary policymakers not responded to output growth at all. This question, under suitable conditions, may be answered by a policy counterfactual.

This set of examples illustrates why the structural VAR framework is often useful for economic analysis. The chief advantage of the structural VAR model compared with alternative econometric approaches is that it tends to fit the data well and only involves minimal identifying restrictions. In particular, it does not impose cross-equation restrictions or exclusion restrictions on the reduced form that tend not to be robust across alternative specifications of the underlying economic model. The remainder of this book provides a more detailed discussion of the historical evolution of the structural VAR approach, of its implementation, of its properties, and of its pros and cons compared with alternative structural econometric models.

1.2 Outline of the Book

Chapter 2

Chapter 2 of the book deals with the reduced-form specification of VAR models. It introduces stochastic and deterministic trends and shows that, under weak assumptions, the purely stochastic component of a multivariate time series process may be modeled as a finite-order VAR model with white noise errors. This VAR model may be viewed either as the DGP or, more plausibly, as an approximation to a more general linear DGP. If the VAR model is stable, it may equivalently be represented as a moving average (MA) process that expresses the model variables as a weighted average of past regression errors. This MA representation is a good point of departure for studying the implications of temporal or cross-sectional aggregation of the VAR model variables, which can be shown to have important effects on the MA structure and hence on the implied reduced-form VAR representation.

In practice, the parameters of a given VAR model of finite lag order are not known and must be estimated from the data. We review not only the conventional LS and conditional ML estimators of unrestricted stationary reduced-form VAR models but also the bias-corrected LS estimator of unrestricted stationary VAR models and the generalized least-squares (GLS) estimator of restricted stationary VAR models. We then discuss how the presence of integrated variables may affect the convergence rate and asymptotic distribution of the LS and ML estimator, when estimating the model in levels, and what assumptions are required for these estimators to remain asymptotically valid

when the finite-order VAR model is only an approximation to a more general linear process such as an invertible vector autoregressive moving average (VARMA) process.

Next, we explain how to form predictions from the estimated VAR process, and derive the one-step ahead mean-squared prediction error (MSPE) matrix, which allows a formal discussion of the notion of Granger causality (or linear predictability) in the VAR model. The one-step ahead MSPE matrix also plays an important role in the design of commonly used data-based lag-order selection criteria for VAR models. Among the latter methods, we discuss top-down and bottom-up sequential testing procedures, information criteria, and recursive MSPE rankings. The latter approach is conceptually closely related to the use of information criteria. We note that data-based lag-order selection methods not only tend to select lag orders that are too low in small samples, but that they undermine the asymptotic validity of inference about the parameters of the implied VAR model. We make the case that the use of a fixed conservative lag order may circumvent these small-sample and asymptotic problems. Chapter 2 also briefly reviews standard diagnostic tests for non-normality, for serial correlation and for conditional heteroskedasticity in the regression errors, and for the time invariance of VAR model parameters.

Although unrestricted reduced-form VAR models are most common in empirical work, there have been proposals for restricting the lag structure of reduced-form VAR models on economic grounds or on statistical grounds. We conclude with a brief discussion of three classes of restricted reduced-form VAR models, including (1) subset VAR models, which allow for the maximum lag order to differ across model equations; (2) asymmetric VAR (AVAR) models, in which the maximum lag order is the same for each variable in all equations but differs across model variables; and (3) VARX models, where the X refers to the inclusion of one or more exogenous variables. Loosely speaking, a variable is exogenous if it is determined outside of the system of equations under consideration. In other words, an exogenous variable is not subject to current or lagged feedback from other VAR model variables, but depends only on its own lags (or possibly lags of other exogenous variables). For example, in a small open economy, the world interest rate may be considered exogenous with respect to the domestic economic variables.

Chapter 3

Many economic variables can be thought of as exhibiting stochastic trends. Such variables are integrated of order 1, which means that only their first difference is stationary. If two or more integrated variables share a common stochastic trend, they are referred to as cointegrated. More generally, cointegration arises when linear combinations of integrated variables are stationary. Although integrated and cointegrated models may be estimated in levels, as

discussed in Chapter 2, imposing integration and cointegration restrictions on the reduced-form VAR model improves the finite-sample accuracy of the VAR model estimates, if these restrictions are correct. Such restricted VAR parameterizations are known as vector error correction models (VECMs). If all VAR model variables are integrated, but none are cointegrated, the VECM reduces to a VAR model in first differences.

VECMs are discussed in Chapter 3. They are of special interest for structural VAR analysis because they allow economists to identify structural economic shocks by imposing restrictions on the long-run behavior of selected variables, as discussed in Chapter 10, which would not be possible if the level representation of the VAR model were stationary. Moreover, common trends implied by cointegration restrictions may have natural economic interpretations as equilibrium relationships.

We first define cointegration and then show how a reduced-form VAR model in levels may be reparameterized as a VECM with special attention to the role of deterministic terms in cointegrated processes. We discuss how VECMs may be estimated by ML methods or by feasible GLS methods, possibly subject to restrictions on the VECM parameters. Because VECMs are reparameterizations of VAR models in levels, they may alternatively be viewed as the true model or as approximations to a more general linear DGP. In the latter case, additional assumptions are needed to ensure the asymptotic validity of the VECM estimator.

Chapter 3 reexamines the question of how to choose the lag order and how to conduct diagnostic tests in the context of the VECM, stressing that many of the results in Chapter 2 are robust to the presence of integrated and cointegrated variables in the model. We also address the important question of how to specify the cointegrating rank when the nature of the cointegration relationships is not already pinned down by economic theory.

Finally, we examine the costs of estimating VECMs in levels without imposing integration and cointegration restrictions, and discuss the question of how to choose between these two specifications. Given the uncertainty about the validity of integration and cointegration restrictions in practice, and our inability to discriminate between alternative VAR and VEC model specifications reliably in small samples, a number of alternative asymptotic thought experiments have been proposed including local-to-unity asymptotics and asymptotics for fractionally integrated processes. In Chapter 3, we briefly introduce these ideas, which will come up again in Chapters 11 and 12.

Chapter 4

The structural VAR representation expresses the reduced-form VAR errors as a linear combination of structural shocks with economic interpretation. If

we know the structural impact multiplier matrix, which describes the weights attached to each structural shock contributing to the reduced-form error, we can always recover the structural VAR representation from the reduced-form VAR representation, as discussed in Chapter 4. Knowledge of the structural representation of the VAR model (or of the VECM) allows users to construct the responses of each model variable to each structural shock (known as structural impulse responses), to assess the extent to which each structural shock contributes to the variability in the model variables (known as a forecast error variance decomposition), and to assess how the data would have evolved in the absence of one or more of the structural shocks (known as a historical decomposition). The latter decomposition also allows users to simulate counterfactual outcomes, it can be used for the construction of policy counterfactuals which examine how hypothetical changes in policy rules affect economic outcomes, and it facilitates the construction of forecast scenarios which measure the extent to which a baseline forecast would change in response to certain hypothetical future events, expressed as sequences of future structural shocks. Chapter 4 reviews and illustrates each of these tools, highlighting alternative representations used in the literature. The question of how to obtain the structural impact multiplier matrix is deferred to Chapters 8, 9, 10, 11, 13, 14, 15, and 17.

Chapter 5

A substantial part of the VAR literature relies on Bayesian methods. Indeed, many of the leading contributors to the VAR literature have been Bayesian econometricians. For example, the method of using sign restrictions for identification discussed in Chapter 13 was originally developed within a Bayesian framework. It therefore is important for users of structural VAR models to understand these methods, if only to be able to interpret Bayesian estimation results reported in the literature.

Chapter 5 contrasts the central premises of the Bayesian estimation approach with those of the frequentist estimation approach discussed in Chapters 2 and 3. We review the role of the prior density of the model parameters, the likelihood of the model, and the posterior density of the model parameters in Bayesian analysis. We discuss how a Bayesian would construct a point estimate from the posterior distribution, how Bayesian credible sets differ from frequentist confidence intervals, how Bayesian model comparisons differ from classical hypothesis testing, and how model averaging may be used as an alternative to model selection.

Because of the central role played by the posterior distribution of the VAR model parameters in Bayesian estimation and inference, we provide a brief overview of methods that may be used to sample from this distribution

including direct sampling from a known posterior distribution, acceptance sampling, importance sampling, Markov Chain Monte Carlo methods, the Metropolis-Hastings algorithm, and the Gibbs sampler.

We then review the leading methods of specifying priors for the reduced-form VAR parameters in empirical work that are all based on the premise of a Gaussian VAR process. These methods include (1) a Gaussian prior for the slope parameters for a given estimated error covariance matrix, (2) the natural conjugate Gaussian-inverse Wishart prior, and (3) the independent Gaussian-inverse Wishart prior. Making these methods operational is not straightforward without further assumptions. A popular approach known as the Minnesota or Litterman prior reduces the problem of specifying a high-dimensional Gaussian prior distribution for all VAR slope parameters to one of specifying a much smaller set of hyperparameters at the cost of imposing additional parametric structure. Alternatively, priors may be imposed on the VEC representation of a VAR model, allowing for cointegration or for near unit roots. The latter approach is also known as the sum-of-coefficients prior.

Chapter 6

Chapter 6 puts the development of VAR models in historical context and clarifies their relationship with other modeling frameworks used in macroeconometrics. We discuss what structural VAR models have in common with dynamic simultaneous equations models (DSEMs) and how they differ from traditional DSEMs of the type widely used in empirical macroeconomics until the 1970s. We also examine the conditions under which DSGE models, which have been popular since the 1980s, have a reduced-form VAR representation, highlighting the fact that the VAR representation of DSGE model variables, if it exists, typically will not be of finite order. Even more stringent conditions are required for DSGE models to have a structural VAR representation.

DSGE models today are the leading alternative to structural VAR models in macroeconometrics. We briefly review alternative approaches of evaluating DSGE models by calibration, by frequentist, and by Bayesian estimation, stressing the commonalities and differences between calibration methods and GMM estimation, on the one hand, and calibration and Bayesian estimation, on the other. We then contrast the pros and cons of DSGE models compared with structural VAR models. As part of this discussion, we also address common misperceptions about structural VAR models not being "structural" and about DSGE models not requiring auxiliary assumptions about data transformations and lag orders. The implications of the Lucas Critique for policy analysis in DSGE models and in structural VAR models are also discussed. We conclude that DSGE models and structural VAR models are complementary with each approach having its own strengths and weaknesses. The chapter ends with a brief overview of efforts to combine elements of structural VAR models

with traditional DSEMs on the one hand and with DSGE models on the other.

Chapter 7

The central objective in structural VAR analysis is to quantify causal relationships in the data. Chapter 7 studies the precursors of structural VAR models, some of which continue to be used in empirical work to this day. Using the debate over money-income causality as our motivating example, we trace the evolution of the profession's thinking about causality from the narrative approach of Friedman and Schwartz (1963) to Granger causality tests in the 1970s with special attention to the concepts of strict exogeneity and predeterminedness. We explain why the profession lost interest in questions of Granger causality in the 1980s and began to focus on unanticipated changes in economic variables instead. The chapter focuses on the development of direct measures of exogenous monetary policy shocks, of fiscal policy shocks, of OPEC oil supply shocks, of news shocks based on macroeconomic announcements, and of shocks to financial market expectations, for example. We trace the further evolution of this literature from distributed-lag models of the impact of directly observable exogenous shocks to VAR models driven by unobserved exogenous shocks that can only be recovered with the help of additional identifying assumptions.

Chapter 8

Although there have been important advances in how one determines the specification of structural VAR models, in how structural VAR estimates are presented, and in how the estimation uncertainty is captured, the question of the identification of structural economic shocks has always been central in this literature, and this question appropriately receives the most weight in our book, starting with Chapter 8.

The chapter starts by contrasting the structural representation of VAR models with the reduced-form specification discussed in Chapters 2, 3, and 5. We discuss the nature of the identification problem in structural VAR modeling and illustrate how alternative normalizing assumptions affect the type of additional restrictions required for the order and rank conditions for exact identification to hold. Identification of unique structural shocks in a VAR model may be achieved by imposing restrictions on the structural impact multiplier matrix of the model (or alternatively on its inverse). These matrices govern the contemporaneous interaction of the model variables and/or of the structural shocks, conditional on the lagged model variables. Hence, restrictions on these matrices are commonly referred to as short-run identifying restrictions. There are several ways of reducing the number of free parameters in the structural VAR

model to be estimated, but the most common approach is to impose exclusion restrictions that limit the contemporaneous feedback between some of the model variables.

We emphasize that imposing a recursive ordering on the model variables in the impact period, as is common in applied work, amounts to imposing a particular causal chain that results in economically meaningless measures of structural shocks, unless this ordering can be economically motivated. Having reviewed common sources of economically meaningful identifying restrictions, we consider several examples of recursively identified structural VAR models with careful attention to the economic content of their identifying assumptions. We stress that credibly identifying all structural shocks by recursive orderings is feasible only in rare cases, but note that sometimes recursive orderings may be used to identify one of the structural shocks with the other structural shocks remaining unidentified from an economic point of view. We also discuss examples of nonrecursively identified structural VAR models. The chapter concludes with a brief discussion of the graph-theoretic approach to identification. We point out that such data-based approaches to identification are not designed to uncover economically meaningful structures and hence are no substitute for economic reasoning in the construction of structural VAR models.

Chapter 9

Having decided on the identifying restrictions, the question arises of how to estimate the structural VAR model. There are three common approaches. We can estimate the structural VAR model (1) by the method of moments or by instrumental-variable (IV) methods, (2) by full information maximum likelihood (FIML) methods, or (3) by Bayesian methods.

Perhaps the most common approach in applied work is the method of moments. For exactly identified models, one proceeds in two steps. One first estimates the reduced-form VAR model as described in Chapters 2 and 3. One then solves for the structural impact multiplier matrix. If the structural model is recursive, this may be accomplished by simply applying a Cholesky decomposition to the covariance matrix of the reduced-form residuals. More generally, we may use a nonlinear equation solver to solve the system of equations linking the unique elements of the covariance matrix of the reduced-form residuals to the unknown elements of the structural impact multiplier matrix. A third option that is computationally less demanding than a nonlinear equation solver (and hence particularly appealing when working with nonrecursive models) is to use the algorithm proposed by Rubio-Ramírez, Waggoner, and Zha (2010). If there are more restrictions than required for exact identification, rendering the model overidentified, we may solve the model in one step by numerically minimizing the GMM objective function. In some cases, the

method-of-moments estimator may also be constructed using traditional IV regression techniques.

Another common approach is FIML estimation, which also accommodates overidentified models, but in the latter case (like the GMM estimator) requires the use of numerical methods. Finally, Bayesian estimation is common in applied work. There are two alternative approaches. For exactly identified models it is standard to rely on conventional reduced-form priors, as discussed in Chapter 5, to generate draws from the posterior of the VAR model, from each of which an estimate of the structural impact multiplier matrix may be obtained by applying the second step of the method of moments. An alternative to this widely used approach is to specify a prior directly on the structural VAR representation, which also accommodates the overidentified case, as discussed at the end of Chapter 9.

Chapter 10

Finding enough short-run restrictions for identifying the structural shocks of interest can be a challenge in practice. One alternative idea in the literature has been to impose restrictions on the long-run response of the model variables to selected shocks. In the presence of unit roots in some variables, but not in others, this approach may allow us to identify at least some structural shocks. The use of long-run restrictions has been appealing because many economists find it easier to agree on long-run restrictions than on the short-run behavior of the economy. It is not without important drawbacks, however.

Chapter 10 introduces a general framework for imposing both short-run and long-run restrictions. We show how alternative specifications of the same model affect how long-run restrictions are imposed. We discuss a range of empirical examples of the use of long-run restrictions both in isolation and in conjunction with short-run restrictions. We draw attention to the fact that special care is needed in specifying such models to avoid some structural shocks in the model having unintended permanent effects.

The chapter concludes with an overview of the limitations of models based on long-run identifying restrictions including the fact that they require exact unit roots in some model variables, their sensitivity to omitted variables, their lack of robustness at lower data frequencies, their sensitivity to data transformations, and the fact that they yield nonunique solutions without additional normalizations.

Chapter 11

As in the case of models identified by short-run restrictions, we may estimate models identified by long-run restrictions (or by a combination of long-run and short-run restrictions) by the method of moments, by IV methods, or by FIML

methods, depending on the nature of the restrictions. Chapter 11 highlights differences in the estimation of models subject to long-run restrictions, depending on whether the model is expressed as a stationary VAR model or in VECM representation. We also review a number of practical problems with the estimation of models identified by long-run restrictions, including that the estimator of the long-run multiplier matrix may be unreliable, the near-observational equivalence of shocks with permanent effects and shocks with persistent effects, and weak instrument problems in implementing the estimator.

The chapter concludes with a review of an ongoing debate among macroeconomists over the ability of structural VAR models to recover the structural impulse responses implied by DSGE models based on synthetic data generated by these DSGE models. This debate was triggered by Galí (1999) who suggested that evidence based on structural VAR models identified by long-run exclusion restrictions had important implications for DSGE modeling. We reexamine this controversy, drawing on the insights provided in Chapters 10 and 11.

Chapter 12

Estimates of structural VAR models are subject to uncertainty. In practice, users of VAR models are typically not interested in the estimation uncertainty about the model parameters themselves, but in the uncertainty about the implied estimates of the structural impulse responses and forecast error variance decompositions discussed in Chapter 4. For expository purposes, Chapter 12 focuses on inference about structural impulse response estimates. It is understood that the discussion (with some exceptions noted in the chapter) also generalizes to related statistics such as forecast error variance decompositions. The chapter starts with a review of impulse response confidence intervals based on the delta method in stationary VAR models, followed by a detailed review of how to generate bootstrap approximations to the sampling distribution of VAR estimators and of how to use that information for the construction of bootstrap confidence intervals for structural impulse responses.

Next we discuss potential limitations of both delta method and bootstrap approaches in the VAR model and in the VECM context. The chapter highlights the special problems that arise in the possible presence of unit roots and cointegration. We stress that pretesting for unit roots and cointegration undermines the validity of inference in structural VAR and VEC models. We first discuss how standard methods of inference designed for stationary models can be made robust to the possible presence of unit roots and cointegration, at least asymptotically. We then review the use of local-to-unity asymptotics as an alternative asymptotic approximation for the VAR model in levels when the data are persistent. We also discuss nonstandard bootstrap methods designed for local-to-unity processes and other methods designed to improve inference

on structural impulse responses in that framework. Finally, we explain why Bayesian methods of inference are not immune to the possible presence of stochastic trends.

Standard methods of inference for structural impulse responses and related statistics are pointwise. From an economic point of view we are often interested not so much in the value of a given response function at a given horizon as in the overall shape and pattern of sets of structural impulse response functions. Assessing these features requires the user to do joint inference. Chapter 12 reviews recently proposed frequentist and Bayesian methods of conducting joint inference about structural impulse responses and illustrates the importance of these methods by example.

We also briefly explain how the implementation of the bootstrap approach depends on whether we are interested in constructing bootstrap confidence intervals, in conducting predictive inference, or in constructing bootstrap critical values for a test statistic. The chapter concludes with a range of empirical examples that illustrate the implementation of commonly used methods of constructing confidence intervals for structural impulse responses.

Chapter 13

Although there are situations in which exclusion restrictions may be motivated by economic theory or institutional features, it is rare for economists to be able to make a strong case for such restrictions. In many situations, economic theory only speaks to the sign of structural impulse responses on impact. For example, in a typical bivariate model of demand and supply, we expect a negative supply shock to lower quantity and to raise the price, whereas a positive demand shock would raise both quantity and the price. In other words, economic theory has implications for the sign of the impact responses. Only if the short-run supply curve were vertical would imposing a zero restriction on the contemporaneous effect of a demand shock on the quantity make economic sense.

This observation has motivated the idea of identifying structural VAR shocks based on sign restrictions. Because sign restrictions are inequality restrictions, the resulting structural models are no longer exactly identified, but only set identified. Put differently, even with an infinite amount of data, we can only narrow down each structural parameter estimate to a range of values rather than to a point in the parameter space. This property also is shared by the implied structural impulse responses and related statistics. As a result, the methods of inference discussed in Chapter 12 cannot be used for sign-identified models.

Chapter 13 examines in detail methods for approximating the set of identified structural impulse responses. Such approximations are typically constructed using Bayesian methods. We show how sign restrictions on impact responses ("static sign restrictions") may be complemented by sign restrictions

on responses at longer horizons ("dynamic sign restrictions") and by inequality restrictions on linear and nonlinear transformations of structural impulse responses. The chapter reviews several methods designed to evaluate the posterior distribution of set-identified impulse responses with special attention to the role of priors. We highlight the challenges in summarizing the posterior information from sign-identified models, discuss the difficulties in interpreting so-called median response functions, and demonstrate how this problem may be overcome in practice.

A central concern with sign-identified VAR models is that the prior remains asymptotically informative about the structural impulse responses. We discuss alternative frequentist and Bayesian methods recently proposed in the literature designed to make the role of the priors more explicit, to attenuate the role of priors, or to inject more economic information into the priors. The chapter concludes with a range of empirical examples and additional discussion on how to combine sign restrictions and more conventional short-run and long-run restrictions within the same structural VAR model.

Chapter 14

As we stressed in earlier chapters, the identification of structural shocks typically relies on economically motivated restrictions that are imposed on the data. An alternative strand of the literature exploits properties of the data for identification. In particular, changes in the conditional or unconditional volatility of the VAR errors may be used for identification. Chapter 14 discusses these approaches in some detail. We stress that this data-based approach does not provide any guidance as to the economic interpretation of the resulting "structural" shocks, however. It only achieves statistical identification in that it produces a unique set of mutually uncorrelated shocks. Thus, it is not a genuine alternative to the approaches discussed in earlier chapters. It is nevertheless attractive because it allows us to formally test traditional economically motivated exclusion restrictions by means of a test of overidentifying restrictions. Chapter 14 examines this approach in a variety of contexts including models with extraneous volatility changes, with Markov switching in the variances, with smooth transitions in the variances, and with generalized autoregressive conditionally heteroskedastic (GARCH) errors. We note that in some cases, this approach supports conventional identifying restrictions, whereas in others these restrictions can be rejected.

Chapter 14 also discusses an alternative data-based approach to identification that exploits the non-Gaussianity of the VAR errors in many applications. In this case, statistical identification may be achieved by insisting that the structural shocks be stochastically independent rather than just uncorrelated. In the Gaussian model, in contrast, uncorrelatedness implies independence. As in the heteroskedastic model, this approach provides a means for testing

more conventional identifying restrictions, if we are willing to postulate independence.

Chapter 15

Yet another option for identifying structural VAR shocks is to rely on additional extraneous data not already included among the VAR variables. Chapter 15 discusses two such approaches. The first approach relies on high-frequency interest rate futures prices. The change in these prices around the time of a monetary policy shift, suitably scaled to monthly frequency, is interpreted as the policy shock. Responses of the variable of interest to these extraneous shocks may be estimated outside of the VAR model and later imposed when estimating the structural VAR model.

The second approach uses directly observed measures of exogenous shocks as instruments in identifying exogenous variation in the VAR variable of interest. This approach allows us to use the exogenous shock measures discussed in Chapter 7 in the VAR context and is not limited to modeling monetary policy. For example, direct measures of exogenous fiscal policy shocks may be used as instruments to isolate the exogenous variation in fiscal variables.

Chapter 16

Typical VAR models include only a comparatively small number of variables. In recent years, public access to time series data has improved to the point that large panels of time series data are now available, often including dozens or hundreds of variables. A reasonable presumption is that all of these variables potentially include information relevant to forming expectations and to identifying structural shocks. Clearly, such large-dimensional models cannot be estimated without imposing additional structure in estimation, however.

Two popular approaches to this problem have been structural factor-augmented VAR (FAVAR) models and large-scale Bayesian structural VAR models. FAVAR models reduce the dimensionality of the estimation problem by imposing an approximate factor structure on the data, allowing us to approximate any variable as a linear combination of the most important factors contained in the panel of data. The advantage of this approach is that the set of factors is of much lower dimension than the original set of model variables. Its disadvantage is that it is not clear that the economically relevant structural shocks can be identified based on factors or linear combinations of factors. Large Bayesian structural VAR models, in contrast, solve the problem of dimensionality by imposing priors on the estimation problem. An obvious concern with the latter approach is that it is difficult to know how influential that prior is, given that we are not able to estimate the model without the prior.

One reason why such models are used in applied work is that they allow us to trace the responses of a larger set of variables to a given structural shock than would be possible using a conventional VAR model, providing an alternative to traditional DSEMs which had the same ability but at the cost of imposing stronger dynamic restrictions. Another reason is that such models greatly increase the information set used in measuring structural shocks, helping us address concerns about the informational structure of many small-scale structural VAR models. Chapter 16 reviews the derivation, specification, identification, and estimation of these models, highlighting potential problems with each approach. We also briefly discuss related approaches such as panel VAR models and global VAR (GVAR) models.

Chapter 17

Structural VAR analysis is based on the premise that the structural VAR shocks can be recovered from the reduced-form prediction errors. The corresponding reduced-form MA representation is called a fundamental representation. If the shocks in the reduced-form MA representation of the DGP are not the VAR prediction errors, in contrast, the reduced-form MA representation of the VAR model is nonfundamental, and it will in general not be possible to recover the true structural shocks even asymptotically. Such a situation arises when the econometrician's model does not have all the information that economic agents in the real world use in forming expectations. In other words, the reduced-form VAR model is informationally deficient. An important special case of this situation would be a model in which agents have forward-looking expectations that cannot be captured based on the information set of the VAR model.

Most commonly, nonfundamental representations are associated with an omitted-variable problem. The obvious response is to recognize that the root of this problem is the omission of relevant variables and to extend the information set, if possible. Note that the use of large-scale VAR models as discussed in Chapter 16 does not necessarily solve this problem because some of the variables relevant to agents' expectations may simply not be contained in any database. For example, oil market participants may anticipate rising oil prices because of fears about ethnic unrest in the Middle East, yet no database includes time series capturing the determinants of these fears. There are creative solutions to this type of problem in specific cases, however, some of which are discussed in Chapter 17.

Chapter 18

The standard VAR model is linear. In some cases, we may wish to allow the model variables to depend nonlinearly on past observations of the model variables rather than just linearly. Such models are collectively referred to as

nonlinear VAR models. Examples of nonlinear dynamics include models with smoothly evolving time-varying coefficients and models with coefficients that change with the state of the economy. Nonlinear VAR models allow economists to model target zones, stochastically switching regimes in the economy, gradual transitions to new economic regimes, thresholds induced by transaction costs, asymmetries in the responses of model variables to positive and negative shocks, and many other economically relevant phenomena.

An important difference compared with linear VAR models is that nonlinear structural impulse responses depend on the history of the data prior to the time period in which a structural shock occurs as well as on the magnitude and sign of this structural shock. This means that the structural impulse responses must be evaluated by numerical methods. Although the effect of alternative histories may be integrated out to arrive at an unconditional impulse response function, the nature of the structural shock remains important even in that case. Put differently, there is no unique set of structural impulse responses in nonlinear models, but rather a family of responses indexed by the magnitude and sign of the structural shock.

A common simplifying assumption in the literature has been that the nonlinear model is linear conditional on past values of the model variables, allowing the use of standard short-run exclusion restrictions. The use of sign restrictions is feasible as well, although it is not clear how to summarize the posterior distribution of the set of structural impulse responses in that case. In contrast, the use of long-run restrictions is not straightforward. The reason is that the closed-form solutions for the construction of structural impulse responses in linear models are not valid for nonlinear models. Rather, structural impulse responses must be constructed by Monte Carlo integration. It is not clear how to impose long-run restrictions on these numerical estimates. Neither do standard methods of constructing forecast error variance decompositions or historical decompositions apply in nonlinear VAR models.

Chapter 18 discusses the specification of nonlinear VAR models, their estimation, the identification of structural shocks, and inference about statistics such as structural impulse responses. We illustrate how existing methods for VAR models must be adapted in the nonlinear context. We also discuss alternative proposals for constructing impulse responses in nonlinear models such as the generalized impulse response function (GIRF). In addition, the chapter reviews related nonlinear models in the literature, such as nonparametric VAR models that allow for more flexible functional forms and noncausal VAR models that have been used to model nonfundamental representations of the type discussed in Chapter 17.

Finally, we examine models that are linear in the parameters yet imply nonlinear impulse response functions. A case in point is recently proposed structural models that allow for asymmetries in the response to positive and negative shocks, even in the impact period. For example, it is widely thought

that unexpected declines in the price of oil cause the gasoline price to fall less quickly than an unexpected increase in the price of oil of the same magnitude would cause the price of gasoline to increase. Using a common analogy, in the former case, gasoline prices fall slowly like a feather; in the latter, they shoot up like a rocket. We stress that the precise specification of the model matters when modeling asymmetries. For example, widely used censored oil price VAR models are invalid.

Chapter 19

The last chapter discusses practical issues related to trends, seasonality, and structural change. We review alternative more flexible trend models such as the HP filter and band-pass filters. We discuss how to combine variables with different trend specifications within the same model. We summarize in some detail the options for modeling seasonality in VAR models, and we discuss the implications of structural breaks for the specification of VAR models.

2 Vector Autoregressive Models

Structural VAR analysis is based on the premise that the DGP is well approximated by a reduced-form VAR model. In applied work, it is therefore important to choose a suitable VAR specification, taking account of the properties of the data. This chapter is devoted to the question of how to specify and estimate reduced-form VAR models. In Section 2.1 stochastic and deterministic trends in the data are discussed. Section 2.2 outlines the basic linear VAR model and its properties. Section 2.3 examines the estimation of reduced-form VAR models. Section 2.4 discusses how to generate predictions from VAR models, and Section 2.5 introduces the concept of Granger causality. Lag-order selection and model diagnostics are discussed in Sections 2.6 and 2.7. Section 2.8 briefly reviews three classes of restricted reduced-form VAR models.

Given that the linear VAR model is one of the standard tools for empirical research in macroeconomics and finance, there are many previous good expositions of the topics covered in this chapter. Our discussion draws heavily on the material in Lütkepohl (2005, 2006, 2009, 2013)

2.1 Stationary and Trending Processes

We call a stochastic process covariance stationary or simply stationary if it has time invariant first and second moments. Similarly, an economic variable is referred to as covariance stationary if the underlying DGP is covariance stationary. More formally, the scalar process y_t, $t \in \mathbb{N}$ or $t \in \mathbb{Z}$, is covariance stationary if

$$\mathbb{E}(y_t) = \mu \quad \text{and} \quad \text{Cov}(y_t, y_{t+h}) = \gamma_h, \quad \forall t, h.$$

Note that μ and γ_h are constants that do not depend on t. This property is also known as second-order stationarity. If the joint distribution of $y_t, \ldots, y_{t+h}$ is time invariant, the process y_t is strictly stationary.

In practice, an economic variable being stationary is the exception rather than the rule. For example, often the raw data have to be transformed prior

to the analysis by taking natural logs to stabilize the variance of the variable. In addition, there are many variables that have trends that have to be removed or modeled explicitly to ensure stationarity. A trend in a time series variable is thought of as a systematic upward or downward movement over time. For example, a variable y_t may vary about a linear trend line of the form $y_t = \mu_0 + \mu_1 t + x_t$, where x_t is a zero mean stationary stochastic process. The straight line, $\mu_0 + \mu_1 t$, represents a simple deterministic trend function that captures the systematic upward or downward movement of many economic variables reasonably well.

Alternatively, a variable may be viewed as being driven by a stochastic trend. A simple example of a process with a stochastic trend is the univariate AR(1) process

$$y_t = a y_{t-1} + u_t$$

with coefficient $a = 1$ such that

$$y_t = y_{t-1} + u_t.$$

This process is called a random walk. Its AR polynomial has a unit root, i.e.,

$$1 - az = 0 \quad \text{for } z = 1.$$

Its stochastic error u_t (also known as the innovation) is assumed to be a white noise process with mean 0 and variance σ_u^2. In other words, u_t and u_s are uncorrelated for $s \neq t$, $\mathbb{E}(u_t) = 0$, and $\mathbb{E}(u_t^2) = \sigma_u^2$. Given that $y_t - y_{t-1} = u_t$, it is easily seen that the effect of a random change in u_t on future values of y_t is not reversed in expectation. Thus, the effect of u_t on future values of y_t is permanent.

Successive substitution for lagged y_t variables in the defining equation of the random walk, $y_t = y_{t-1} + u_t$, yields

$$y_t = y_0 + \sum_{i=1}^{t} u_i. \tag{2.1.1}$$

Hence, assuming that the process is defined for $t \in \mathbb{N}$, we have

$$\mathbb{E}(y_t) = \mathbb{E}(y_0) \quad \text{and} \quad \text{Var}(y_t) = t\sigma_u^2 + \text{Var}(y_0).$$

In other words, even though $\text{Var}(y_0)$ is finite, the variance of a random walk tends to infinity. Moreover, the correlation

$$\text{Corr}(y_t, y_{t+h}) = \frac{\mathbb{E}\left[\left(\sum_{i=1}^{t} u_i\right)\left(\sum_{i=1}^{t+h} u_i\right)\right]}{[t\sigma_u^2(t+h)\sigma_u^2]^{1/2}} = \frac{t}{(t^2+th)^{1/2}} \xrightarrow[t\to\infty]{} 1 \tag{2.1.2}$$

for any given integer h. Due to this property, even random variables y_t and y_s of the process far apart in time (such that s is much greater than t) are strongly correlated. This property indicates a strong persistence in the time series process. In fact, it turns out that the expected time between two crossings of zero is infinite. Such behaviour is associated with a trend in the data. Clearly, since u_t is stochastic, so is the trend.

A univariate AR(1) process with unit coefficient and a constant term,

$$y_t = \nu + y_{t-1} + u_t,$$

is called a random walk with drift. Successive substitution of lags of y_t shows that in this case

$$y_t = y_0 + t\nu + \sum_{i=1}^{t} u_i$$

and, hence, the process has a linear trend in the mean:

$$\mathbb{E}(y_t) = \mathbb{E}(y_0) + t\nu.$$

Higher-order AR processes such as

$$y_t = \nu + a_1 y_{t-1} + \cdots + a_p y_{t-p} + u_t,$$

where u_t is white noise as before, have stochastic trending properties similar to random walks if the AR polynomial $1 - a_1 z - \cdots - a_p z^p$ has a root for $z = 1$. The AR polynomial can be decomposed as

$$1 - a_1 z - \cdots - a_p z^p = (1 - \lambda_1 z) \times \cdots \times (1 - \lambda_p z), \tag{2.1.3}$$

where $\lambda_1, \ldots, \lambda_p$ are the reciprocals of the roots of the polynomial. If the process has only one unit root or, equivalently, only one of the λ_i roots is 1 and all the others are smaller than 1, the process behaves similarly to a random walk in that it follows a stochastic trend. More precisely, y_t can be decomposed into a random walk and a stationary component such that y_t varies about a stochastic trend generated by its random walk component.

The representation of the AR polynomial shows that the unit root can be removed by taking first differences of the process. Let $\Delta y_t \equiv (1 - L)y_t \equiv y_t - y_{t-1}$, where L is the lag operator such that $Ly_t \equiv y_{t-1}$, and Δ is the difference operator such that $\Delta \equiv 1 - L$ and hence $\Delta y_t = y_t - y_{t-1}$.

An AR(p) process with AR polynomial satisfying the condition

$$1 - a_1 z - \cdots - a_p z^p \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1, \tag{2.1.4}$$

is called stable. Here $|z|$ denotes the modulus of the complex number z. Put differently, $|z|$ is the distance from the origin of the complex plane. If, in addition, the mean of the AR process does not change over time deterministically, as would be the case in the presence of a deterministic time trend, if the error

term u_t has time-invariant variance σ_u^2, and if its first and second moments are bounded, then the AR process is stationary. Sometimes in the literature, condition (2.1.4) is rather imprecisely viewed as a condition ensuring stationarity. Of course, interpreting (2.1.4) as a stationarity condition implicitly assumes that there are no other deviations from stationarity such as a linear deterministic trend in the mean or an innovation variance changing over time.

To ensure finite moments, AR processes with unit roots are assumed to start at some fixed time period, say t_0, if not explicitly stated otherwise. For example, in the foregoing discussion we have assumed that $t_0 = 0$. In contrast, stable AR processes without unit roots are typically assumed to have started in the infinite past to ensure stationarity. Without that assumption they may only be asymptotically stationary in that the moments are not time-invariant, but converge to their limit values only for $t \to \infty$.

If the AR polynomial has $d \in \mathbb{N}$ unit roots and, hence, d of the λ_i roots in (2.1.3) are equal to 1, the process in called integrated of order d ($I(d)$). In that case, the process can be made stable by differencing it d times. For example, if $d = 1$, $\Delta y_t = y_t - y_{t-1}$ is stable. If $d = 2$, $\Delta^2 y_t = (1 - L)^2 y_t = y_t - 2y_{t-1} + y_{t-2}$ is stable, and so forth. If $d = 2$, the original y_t must be differenced twice. For example, if the log price level p_t is $I(2)$, then the inflation rate $\pi_t = \Delta p_t = p_t - p_{t-1}$ is $I(1)$, and the change in the inflation rate $\Delta\pi_t = \pi_t - \pi_{t-1} = p_t - p_{t-1} - (p_{t-1} - p_{t-2}) = p_t - 2p_{t-1} + p_{t-2}$ is $I(0)$. As before, initial values can be chosen such that $\Delta^d y_t = (1 - L)^d y_t$ is stationary, provided the conditions for the mean and for the innovation variance required for stationarity are satisfied.

Stable, stationary processes are referred to as $I(0)$ processes. Generally, for $d \in \mathbb{N}$, a stochastic process y_t is called $I(d)$, if $\Delta^d y_t \equiv z_t$ is a stationary process with infinite-order moving average (MA) representation, $z_t = \sum_{j=0}^{\infty} \theta_j u_{t-j} = \theta(L)u_t$, where the MA coefficients satisfy the condition $\sum_{j=0}^{\infty} j|\theta_j| < \infty$, $\theta(1) = \sum_{j=0}^{\infty} \theta_j \neq 0$, and $u_t \sim (0, \sigma_u^2)$ is white noise. For example, in the case of an $I(1)$ process, this condition implies that $y_t = y_{t-1} + z_t$ has the representation

$$\begin{aligned} y_t &= y_0 + z_1 + \cdots + z_t \\ &= y_0 + \theta(1)(u_1 + \cdots + u_t) + \sum_{j=0}^{\infty} \theta_j^* u_{t-j} - z_0^*, \end{aligned} \tag{2.1.5}$$

where $\theta_j^* = -\sum_{i=j+1}^{\infty} \theta_i$, $j = 0, 1, \ldots$, and $z_0^* = \sum_{j=0}^{\infty} \theta_j^* u_{-j}$ contains initial values. The variable y_t is decomposed into the sum of a random walk, $\theta(1)(u_1 + \cdots + u_t)$, a stationary process, $\sum_{j=0}^{\infty} \theta_j^* u_{t-j}$, and initial values, $y_0 - z_0^*$. The decomposition (2.1.5) is known as the Beveridge-Nelson decomposition (see Beveridge and Nelson 1981).

Of course, our primary interest is in systems of variables. Hence, it is useful to extend the $I(d)$ terminology to that setting as well. Accordingly, we call a vector process $y_t = (y_{1t}, \ldots, y_{Kt})'$ $I(d)$ if stochastic trends can be removed by

differencing y_t d times and if differencing $d-1$ times is not enough for trend removal. It is important to note, however, that in systems of variables even if only one of the variables is $I(d)$ individually, the whole system is viewed as $I(d)$. Moreover, it is possible that a single stochastic trend drives several of the variables jointly. This is the important case of cointegrated variables to be discussed in Chapter 3.

The $I(d)$ terminology has also been extended to non-integer, real numbers d. For general $d \in \mathbb{R}$ the so-called fractional differencing operator Δ^d is defined as a binomial expansion,

$$\begin{aligned}\Delta^d = (1-L)^d &= 1 - dL - \frac{d(1-d)}{2}L^2 - \frac{d(1-d)(2-d)}{6}L^3 - \cdots \\ &= \sum_{i=0}^{\infty}(-1)^i\binom{d}{i}L^i \\ &= \sum_{i=0}^{\infty}(-1)^i\frac{d(d-1)\times\cdots\times(d-i+1)}{1\times 2\times\cdots\times i}L^i.\end{aligned}$$

The infinite sum reduces to a finite sum for $d \in \mathbb{N}$. The process y_t is called fractional or fractionally integrated of order d if $\Delta^d y_t = z_t$ is $I(0)$ with MA representation $z_t = \theta(L)u_t$, $\theta(1) \neq 0$ (see, e.g., Johansen and Nielsen 2012). Such processes were introduced to the time series econometrics literature by Granger and Joyeux (1980) and Hosking (1981b). Fractionally integrated processes are often referred to as long-memory processes because for $d > 0$ they are more persistent and their autocorrelations taper off to zero more slowly than for $I(0)$ processes. Although fractionally integrated processes are not $I(0)$, they may be stationary. Stationarity of a fractionally integrated process requires $|d| < 0.5$.

Integer-valued differences often have a natural interpretation. For example, first differences of the logs of a variable represent growth rates. Such an easy interpretation is lost for fractionally differenced variables. Thus, it is perhaps not surprising that the concept of fractional integration to date has not been used much in structural VAR analysis. More importantly, reliable estimation of fractionally integrated processes requires larger samples than typically available in macroeconomics. Fractional processes therefore do not play an important role in this volume. In the remainder of this book, when we refer to $I(d)$ variables, we always mean non-negative integers d unless explicitly stated otherwise.

2.2 Linear VAR Processes

2.2.1 The Basic Model

Suppose that the relationship between a set of K time series variables, $y_t = (y_{1t}, \ldots, y_{Kt})'$, is of interest and that the DGP can be represented as the sum of

a deterministic part μ_t and a purely stochastic part x_t with mean zero such that

$$y_t = \mu_t + x_t. \tag{2.2.1}$$

In other words, the expected value of y_t is $\mathbb{E}(y_t) = \mu_t$. The deterministic term may contain a constant, polynomial trend terms, deterministic seasonal terms, and other dummy variables. For simplicity, μ_t is usually assumed to contain only a constant such that $\mu_t = \mu_0$. Occasionally a linear trend of the form $\mu_t = \mu_0 + \mu_1 t$ is considered. Generally the additive setup (2.2.1) makes it necessary to think about the deterministic terms at the beginning of the analysis and to allow for the appropriate polynomial order. In some applications trend adjustments are performed prior to a VAR analysis. This approach must be taken, for example, when the detrending procedure cannot be incorporated into the VAR specification. An example is the use of HP-filtered data. Further discussion of these alternative detrending methods can be found in Chapter 19. In that case there may be no deterministic term in the VAR representation in levels, i.e., $\mu_t = 0$ in expression (2.2.1) and $y_t = x_t$.

The purely stochastic part, x_t, of the DGP is assumed to follow a linear VAR process of order p (referred to as a VAR(p) model) of the form

$$x_t = A_1 x_{t-1} + \cdots + A_p x_{t-p} + u_t, \tag{2.2.2}$$

where the A_i, $i = 1, \ldots, p$, are $K \times K$ parameter matrices and the error process $u_t = (u_{1t}, \ldots, u_{Kt})'$ is a K-dimensional zero mean white noise process with covariance matrix $\mathbb{E}(u_t u_t') = \Sigma_u$ such that $u_t \sim (0, \Sigma_u)$. The white noise assumption rules out serial correlation in the errors but allows for conditional variance dynamics such as generalized autoregressive conditionally heteroskedastic (GARCH) errors (see e.g. Chapter 14). Sometimes it is useful to strengthen this assumption, for example, by postulating independent and identically distributed (iid) errors or by postulating that u_t is a martingale difference sequence.[1]

Expression (2.2.2) defines a system of equations. Each model variable in y_t is regressed on its own lags as well as lags of the other model variables up to a lag order p (see Chapter 1). To economize on notation, it is convenient to define the matrix polynomial in the lag operator $A(L) = I_K - A_1 L - \cdots - A_p L^p$ and write the process (2.2.2) as

$$A(L)x_t = u_t. \tag{2.2.3}$$

The observed variables y_t inherit the VAR structure of x_t. This can be seen easily by pre-multiplying (2.2.1) by $A(L)$ and considering $A(L)y_t = A(L)\mu_t +$

[1] The stochastic process v_t is called a martingale sequence if $\mathbb{E}(v_t | v_{t-1}, v_{t-2}, \ldots) = v_{t-1}$ $\forall t$. Then $u_t \equiv \Delta v_t$ is called a martingale difference if it has expectation $\mathbb{E}(u_t | v_{t-1}, v_{t-2}, \ldots) = 0$ $\forall t$. Unlike an iid white noise process, a white noise process that is a martingale difference sequence allows for conditional heteroskedasticity.

u_t. For instance, if the deterministic term is just a constant, i.e., $\mu_t = \mu_0$, then

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{2.2.4}$$

where $\nu = A(L)\mu_0 = A(1)\mu_0 = (I_K - \sum_{j=1}^{p} A_j)\mu_0$. In the terminology of the literature on simultaneous equations models, model (2.2.4) is a reduced form because all right-hand side variables are lagged and hence predetermined.

The VAR process x_t and, hence, y_t is stable if all roots of the determinantal polynomial of the VAR operator are outside the complex unit circle, i.e.,

$$\det(A(z)) = \det(I_K - A_1 z - \cdots - A_p z^p) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1, \tag{2.2.5}$$

where $\mathbb{C}$ denotes the set of complex numbers. Under common assumptions such as a constant mean and white noise innovations with time-invariant covariance matrix, a stable VAR process has time-invariant means, variances, and covariance structure and hence is stationary, as will be seen in the next subsection. Thus, condition (2.2.5) generalizes the stability condition (2.1.4) to the multivariate case.

For later reference we note that the K-dimensional VAR(p) process (2.2.4) can be written as a pK-dimensional VAR(1) process by stacking p consecutive y_t variables in a pK-dimensional vector, $Y_t = (y_t', \ldots, y_{t-p+1}')'$, and noting that

$$Y_t = \boldsymbol{\nu} + \mathbf{A} Y_{t-1} + U_t, \tag{2.2.6}$$

where

$$\boldsymbol{\nu} \equiv \underset{Kp\times 1}{\begin{bmatrix} \nu \\ 0 \\ \vdots \\ 0 \end{bmatrix}}, \quad \mathbf{A} \equiv \underset{Kp\times Kp}{\begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_K & 0 & \cdots & 0 & 0 \\ 0 & I_K & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_K & 0 \end{bmatrix}}, \quad \text{and} \quad U_t \equiv \underset{Kp\times 1}{\begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}}.$$

The matrix $\mathbf{A}$ is referred to as the companion matrix of the VAR(p) process. Using the stability condition (2.2.5), Y_t is stable if

$$\det(I_{Kp} - \mathbf{A}z) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1, \tag{2.2.7}$$

which, of course, is equivalent to condition (2.2.5). It is easy to see that this condition is equivalent to all eigenvalues of $\mathbf{A}$ having modulus less than 1, which provides a convenient tool for assessing the stability of a VAR model and for computing the autoregressive roots. By construction, the eigenvalues of $\mathbf{A}$ are the reciprocals of the roots of the VAR lag polynomial (2.2.5).

2.2.2 The Moving Average Representation

A stable VAR(p) process y_t can be represented as the weighted sum of past and present innovations. This is easily seen for a VAR(1) process,

$$y_t = \nu + A_1 y_{t-1} + u_t.$$

Successive substitution implies

$$y_t = \sum_{i=0}^{\infty} A_1^i \nu + \sum_{i=0}^{\infty} A_1^i u_{t-i} = (I_K - A_1)^{-1}\nu + \sum_{i=0}^{\infty} A_1^i u_{t-i}.$$

The sum on the right-hand side of this infinite-order representation exists if the eigenvalues of A_1 are all less than 1 in modulus. Similarly, a representation in terms of past and present innovations of a VAR(p) model can be obtained via the corresponding VAR(1) representation, resulting in

$$\begin{aligned} y_t &= A(L)^{-1}\nu + A(L)^{-1}u_t \\ &= A(1)^{-1}\nu + \sum_{i=0}^{\infty} J\mathbf{A}^i J' J U_{t-i} \\ &= \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \end{aligned} \tag{2.2.8}$$

where $J \equiv [I_K, 0_{K\times K(p-1)}]$ is a $K \times Kp$ matrix, $\mu = A(1)^{-1}\nu$ and the $K \times K$ coefficient matrices of the inverse VAR operator $A(L)^{-1} = \sum_{i=0}^{\infty} \Phi_i L^i$ are equal to $\Phi_i = J\mathbf{A}^i J'$, $i = 0, 1, \ldots$. These matrices can also be obtained recursively as

$$\Phi_0 = I_K, \quad \text{and} \quad \Phi_i = \sum_{j=1}^{i} \Phi_{i-j} A_j, \qquad i = 1, 2, \ldots,$$

with $A_j = 0$ for $j > p$ (see Lütkepohl 2005, chapter 2).

The existence of the inverse VAR operator is ensured by the stability of the process. The representation (2.2.8) is known as the moving average (MA) representation or more precisely the Wold MA representation or the prediction error MA representation. This qualifier is important because there are infinitely many MA representations of y_t. In fact, any nonsingular linear transformation of the white noise process u_t, say $v_t = Qu_t$, gives rise to a white noise process and can be used for an MA representation of y_t,

$$y_t = \mu + \sum_{i=0}^{\infty} \Theta_i v_{t-i}, \tag{2.2.9}$$

with $\Theta_i = \Phi_i Q^{-1}$, $i = 0, 1, \ldots$. A distinguishing feature of the Wold MA representation is that the weighting matrix Φ_0 of the unlagged error term

is the identity matrix, while Θ_0 is not an identity matrix for nontrivial transformations.

It follows immediately from the Wold MA representation that

$$\mathbb{E}(y_t) = \mu$$

and that

$$\Gamma_y(h) \equiv \text{Cov}(y_t, y_{t-h}) = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)'] = \sum_{i=0}^{\infty} \Phi_{h+i} \Sigma_u \Phi_i'. \tag{2.2.10}$$

Hence, the first and second moments of this VAR process are time invariant and the process is stationary (see Lütkepohl 2005, chapter 2).

2.2.3 VAR Models as an Approximation to VARMA Processes

An important result in this context is due to Wold (1938) who showed that every K-dimensional nondeterministic zero mean stationary process y_t has an MA representation

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \tag{2.2.11}$$

where $\Phi_0 = I_K$. This result follows from the Wold Decomposition Theorem and motivates the terminology used for the MA representation (2.2.8). This result is important because it illustrates the generality of the VAR model. Suppose the Φ_i are absolutely summable and that there exists an operator $A(L)$ with absolutely summable coefficient matrices satisfying $A(L)\Phi(L) = I_K$. Then $\Phi(L)$ is invertible [$A(L) = \Phi(L)^{-1}$] and y_t has a VAR representation of possibly infinite order that can be approximated arbitrarily well by a finite-order VAR(p) if p is sufficiently large.

In particular, under suitable conditions, a VAR(p) process may be used to approximate time series generated from vector autoregressive moving average (VARMA) models of the form

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_{p_0} y_{t-p_0} + u_t + M_1 u_{t-1} + \cdots + M_{q_0} u_{t-q_0},$$

where p_0 and q_0 denote the true autoregressive and moving average lag orders, provided the VAR lag order p is sufficiently large. If the VAR operator $A(z) = I_K - A_1 z - \cdots - A_{p_0} z^{p_0}$ of the VARMA process satisfies the stability condition (2.2.5) and, thus, the VAR operator has no roots in or on the complex unit circle, the VARMA process has a possibly infinite-order MA representation (2.2.11).

Moreover, if the determinant of the MA operator of the VARMA process has all its roots outside the unit circle, i.e.,

$$\det(M(z)) = \det(I_K + M_1 z + \cdots + M_{q_0} z^{q_0}) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1,$$

the process also has an equivalent pure VAR representation of possibly infinite order.[2] Unlike in the univariate case, the inverse of a finite-order operator may also be a finite-order operator in the multivariate case. In other words, $M(z)^{-1}$ may be a finite order operator if $M(z)$ has finite order. Hence, it is possible in the multivariate case that a finite-order MA process has an equivalent finite-order VAR representation and vice versa.

A detailed introductory exposition of VARMA processes is provided by Lütkepohl (2005), and a more advanced treatment can be found in Hannan and Deistler (1988). Since VARMA processes are much more difficult to deal with in practice, we focus on VAR models in the remainder of this book.

If the VAR process of interest has a unit root and, hence, the stability condition is not satisfied, the infinite-order MA representation (2.2.8) does not exist. However, we can still think of the process as starting from $Y_0 = (y_0', \ldots, y_{-p+1}')'$ and obtain a representation

$$y_t = \mu_t + \sum_{i=0}^{t-1} \Phi_i u_{t-i} + J\mathbf{A}^i Y_0$$

by successive substitution. For some purposes this representation is useful, but not for all. In particular, it obscures the long-run properties of the process. These are more easily understood using the so-called Granger representation discussed in Chapter 3.

2.2.4 Marginal Processes, Measurement Errors, Aggregation, Variable Transformations

The reduced-form MA representation is also a good point of departure for studying the implications of dropping variables from a VAR process. Consider a bivariate stationary process for two variables,

$$y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{bmatrix} \phi_{11}(L) & \phi_{12}(L) \\ \phi_{21}(L) & \phi_{22}(L) \end{bmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}. \tag{2.2.12}$$

Thus, the first variable has the representation

$$y_{1t} = \mu_1 + \phi_{11}(L)u_{1t} + \phi_{12}(L)u_{2t}$$

[2] An MA representation with MA operator satisfying this invertibility condition is sometimes called a fundamental MA representation. In Chapter 17 we discuss nonfundamental MA representations that have roots inside the complex unit circle and, hence, do not satisfy the invertibility condition for the MA operator.

in terms of both innovation series. According to Wold's decomposition theorem, it also has an MA representation in terms of a scalar white noise process, v_t:

$$y_{1t} = \mu_1 + \sum_{i=0}^{\infty} \psi_i v_{t-i}.$$

This MA represents the marginal process of y_{1t} that is obtained by integrating out the second variable. If $\phi_{12}(L) \neq 0$, then $v_t \neq u_{1t}$ and $\psi_i \neq \phi_{11,i}$ in general. These facts are important to keep in mind for the analysis of impulse responses in Chapter 4. The point to remember is that, in general, dropping some variables from a multivariate time series process results in a lower-dimensional process with possibly quite different MA coefficients than the process for the original set of variables.

More generally, any transformation of the variables implies changes in the MA coefficients. Consider, for example, a nonsingular transformation matrix F and a transformed process

$$z_t = Fy_t = F\mu + \sum_{i=0}^{\infty} F\Phi_i F^{-1} F u_{t-i} = \mu_z + \sum_{i=0}^{\infty} \Psi_i v_{t-i}, \qquad (2.2.13)$$

where $\Psi_i = F\Phi_i F^{-1}$ and $v_t = Fu_t$. Obviously, such transformations change the MA coefficient matrices and white noise error term, and therefore also affect the lag order of the approximating autoregressive process. The same result also holds when F is not a square matrix. Suppose F is an $M \times K$ matrix of rank M. Then one may add $K - M$ rows to the matrix such that it becomes nonsingular, consider the resulting nonsingular transformation, and finally omit the last $K - M$ components of the transformed vector.

In short, linear transformations of a VAR process have MA representations quite different from that of the original process, but both representations are equally valid. For example, a researcher may be working with a VAR for the interest rate, inflation rate, and real GDP growth. These variables could alternatively be represented as autoregressive-moving average (ARMA) processes for each series separately, which in turn can be approximated by finite-order AR models. Linear transformations are also quite common when aggregating data across households and industries to form macroeconomic aggregates. Likewise, problems of temporal aggregation fall within this framework (see Lütkepohl 2005, chapter 11). For example, it is common to aggregate monthly inflation to quarterly inflation data, which involves taking a linear combination of monthly inflation rates.

In this context it is important to stress that different types of variables require different temporal aggregation methods. For flow variables such as GDP or industrial production, temporal aggregation to a lower frequency

involves accumulating the high-frequency observations over time. For example, quarterly industrial production is obtained by summing the monthly industrial production that has taken place within each quarter. In contrast, stock variables such as the number of unemployed workers or the population of a region are aggregated from monthly data to quarterly frequency by using, for example, the last monthly value of each quarter as the quarterly value and dropping the other monthly observations. In other words, temporal aggregation is performed by what is known as skip-sampling or systematic sampling. Alternatively, one may use the average of the monthly values as a quarterly value, depending on the economic context. The important point to note here is that different temporal aggregation schemes imply different changes in the DGP and hence in the MA representation of the variables. There is an extensive literature discussing these issues, in particular in the forecasting context. Early contributions include Tiao (1972), Amemiya and Wu (1972), Brewer (1973), Abraham (1982), and Wei (1981). A more recent systematic account of this literature is provided in Lütkepohl (1987). An alternative approach has been to combine time series observed at different frequencies within the same econometric model (see Foroni, Ghysels, and Marcellino 2013). Related issues in the context of structural modeling are taken up in Chapter 15.

Another example of a linear aggregation problem is additive measurement error in the data. Suppose that the variable of interest, say y_t^*, is measured with error and denote the measurement error by m_t such that the observed process is $y_t = y_t^* + m_t$. In other words, y_t is a linear transformation of the joint process

$$\begin{pmatrix} y_t^* \\ m_t \end{pmatrix}.$$

Then, assuming that the joint process is stationary, the previous discussion implies that the MA representation of y_t differs from that of y_t^*.

These considerations demonstrate that even linear transformations can have a substantial impact on the MA representation of a stationary process. Although these issues do not invalidate the reduced-form representation of VAR models, they may affect the structural interpretation and identification of VAR models, as discussed in later chapters. Our discussion in this section has been based on the MA representation and, hence, applies to stationary processes more generally. We now return to finite-order VAR processes and discuss parameter estimation within that model class.

2.3 Estimation of VAR Models

VAR models can be estimated by standard methods. Unrestricted least-squares (LS), generalized least-squares (GLS), bias-corrected least-squares, and maximum likelihood (ML) methods are discussed in Sections 2.3.1–2.3.4. Our main focus in this chapter is on stationary VAR processes. The properties of LS and