12 Inference in Models Identified by Short-Run or Long-Run Restrictions

Chapters 9 and 11 reviewed the estimation of the structural impulse responses at horizon i, Θ_i, in structural VAR models subject to short-run and/or long-run restrictions. This chapter focuses on inference in exactly identified structural VAR models. We discuss the leading methods of constructing confidence intervals for structural impulse responses and related statistics such as structural forecast error variance decompositions. Similar methods of inference can be applied to all statistics discussed in Chapter 4 with the exception of historical decompositions. The latter statistics do not permit the construction of confidence intervals because they condition on a particular realization of the data.

Section 12.1 presents the delta method, followed by bootstrap methods in Sections 12.2 and 12.3. Section 12.4 discusses potential pitfalls in conducting inference about structural impulse responses in stationary models. In Section 12.5 we summarize the evidence on the finite-sample properties of alternative confidence intervals for structural impulse responses. Section 12.6 extends the analysis to integrated and cointegrated processes. Section 12.7 discusses local-to-unity asymptotic and bootstrap confidence intervals for structural impulse responses. In Section 12.8, we review an alternative and seemingly simpler approach to inference about impulse responses based on local projections and discuss its practical and conceptual limitations. In Section 12.9, we provide practical recommendations for conducting pointwise inference about structural impulse responses. Section 12.10 contrasts Bayesian and frequentist approaches to pointwise inference. Section 12.11 focuses on joint inference about sets of structural impulse responses. Section 12.12 briefly discusses other bootstrap applications including bootstrap prediction and the construction of bootstrap critical values for test statistics. Section 12.13 provides a number of empirical examples. It also discusses extensions to overidentified structural VAR models. Inference in other classes of structural VAR models including set-identified models is considered in subsequent chapters.

12.1 Delta Method Intervals for Structural Impulse Responses

We begin with a summary of asymptotic results based on the delta method for structural impulse response estimates from exactly identified VAR models, building on the analysis in Lütkepohl (1990). Consider the VAR(p) model

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

where y_t is a K-dimensional vector and u_t is an iid error term with at least four finite moments. Under suitable conditions,

$$\sqrt{T}\begin{pmatrix}\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix}\Sigma_{\widehat{\boldsymbol{\alpha}}} & 0 \\ 0 & \Sigma_{\widehat{\boldsymbol{\sigma}}}\end{bmatrix}\right), \tag{12.1.1}$$

where T is the sample size, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\sigma}}$ denote consistent estimators of $\boldsymbol{\alpha} = \text{vec}([A_1, \ldots, A_p])$ and $\boldsymbol{\sigma} = \text{vech}(\Sigma_u)$, respectively. Block-diagonality of the asymptotic covariance matrix in expression (12.1.1) is assumed here for deriving the asymptotic properties of impulse response estimators because it simplifies the asymptotic distributions. This assumption holds, for example, when the errors are iid and symmetrically distributed with four finite moments. More generally, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\sigma}}$ are not asymptotically independent, and hence, the asymptotic covariance matrix is not block diagonal. Nevertheless, the same techniques may be used to derive the asymptotic distribution of the structural impulse response estimator (see, e.g., Lütkepohl 2005, section 3.7, remark 3, or Brüggemann, Jentsch, and Trenkler 2016, theorem 2.1). The expression for the asymptotic covariance matrix will be different, however.

In the special case of Gaussian iid innovations, the covariance matrix $\Sigma_{\widehat{\boldsymbol{\sigma}}}$ simplifies to $\Sigma_{\widehat{\boldsymbol{\sigma}}} = 2\mathbf{D}_K^+(\Sigma_u \otimes \Sigma_u)\mathbf{D}_K^{+\prime}$, where $\mathbf{D}_K^+ = (\mathbf{D}_K'\mathbf{D}_K)^{-1}\mathbf{D}_K'$ is the Moore-Penrose inverse of $\mathbf{D}_K$, and $\mathbf{D}_K$ is the $K^2 \times \frac{1}{2}K(K+1)$ duplication matrix, as defined in Lütkepohl (2005, appendix A12.2), and $\Sigma_{\widehat{\boldsymbol{\alpha}}} = \Gamma_Y(0)^{-1} \otimes \Sigma_u$, where $\Gamma_Y(0)$ may be estimated as ZZ'/T, given $Z \equiv (Y_0, \ldots, Y_{T-1})$ and $Y_t \equiv (y_t', \ldots, y_{t-p+1}')'$. If the regression includes deterministic regressors, $\Sigma_{\widehat{\boldsymbol{\alpha}}}$ corresponds to the last K^2p rows and columns of the variance-covariance matrix of the VAR regressors.

Note that each structural impulse response coefficient $\theta_{jk,i}$ may be expressed as a component of a continuously differentiable function $g(\boldsymbol{\alpha}, \boldsymbol{\sigma}) = (g_1(\boldsymbol{\alpha}, \boldsymbol{\sigma}), \ldots, g_m(\boldsymbol{\alpha}, \boldsymbol{\sigma}))'$ with values in the m-dimensional Euclidean space, where m is the number of structural impulse response coefficients under consideration. If the partial derivatives $\partial g_i/\partial \boldsymbol{\alpha}'$ and $\partial g_i/\partial \boldsymbol{\sigma}'$ are nonzero at the true parameter values $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ for $i = 1, \ldots, m$, an application of the delta method yields

$$\sqrt{T}\left(g(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\sigma}}) - g(\boldsymbol{\alpha}, \boldsymbol{\sigma})\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g}{\partial \boldsymbol{\alpha}'}\Sigma_{\widehat{\boldsymbol{\alpha}}}\frac{\partial g'}{\partial \boldsymbol{\alpha}} + \frac{\partial g}{\partial \boldsymbol{\sigma}'}\Sigma_{\widehat{\boldsymbol{\sigma}}}\frac{\partial g'}{\partial \boldsymbol{\sigma}}\right). \tag{12.1.2}$$

Standard asymptotic inference requires the limiting variance to be nonzero. That condition may not hold if there are constraints on the model coefficients or if the model variables are integrated and/or cointegrated, in which case $\Sigma_{\widehat{\boldsymbol{\alpha}}}$ is singular (see Chapter 3). Even if $\Sigma_{\widehat{\boldsymbol{\alpha}}}$ itself is nonsingular, however, the partial derivatives $\partial g_i / \partial \boldsymbol{\alpha}'$ may be zero, invalidating inference. This may occur in some points of the parameter space because these derivatives are sums of products of the reduced-form VAR slope parameters. An example would be a finite-order VAR process in which some variable does not respond to a given structural shock such that the impulse response is zero in population (see Benkwitz, Lütkepohl, and Neumann 2000). We will return to this point later in this chapter.

The precise form of the variance-covariance matrices differs, depending on whether we are conditioning on a finite-order VAR DGP or approximating an infinite-order VAR process. The following results draw on Lütkepohl (1990) and Lütkepohl and Poskitt (1991). While we focus on the asymptotic variance-covariance matrix for structural impulse responses, similar results for related statistics such as cumulative impulse responses and structural forecast error variance decompositions can be found in Lütkepohl (2005).

12.1.1 Finite-Order VAR Models

Let y_t be generated by a stationary VAR model of known finite lag order p with iid white noise errors. Extensions of the delta method to VAR models with conditionally heteroskedastic errors are discussed in Brüggemann, Jentsch, and Trenkler (2016). In the latter case, the simple additive structure of the variance in equation (12.1.2) is lost because the asymptotic covariance of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\sigma}}$ in equation (12.1.1) is not zero.

We are interested in conducting inference on the elements of the matrix $\widehat{\Theta}_i$ of estimates of the structural impulse responses at horizon i, where i is fixed with respect to the sample size T. For expository purposes, let B_0^{-1} equal the lower-triangular Cholesky decomposition of Σ_u such that $B_0^{-1} B_0^{-1\prime} = \Sigma_u$. Then

$$\sqrt{T}\,\mathrm{vec}\left(\widehat{\Theta}_i - \Theta_i\right) \xrightarrow{d} \mathcal{N}\left(0, C_i \Sigma_{\widehat{\boldsymbol{\alpha}}} C_i' + \overline{C}_i \Sigma_{\widehat{\boldsymbol{\sigma}}} \overline{C}_i'\right), \quad i = 0, 1, 2, \ldots,$$

where

$$\begin{aligned} C_0 &= 0 \\ C_i &= \left(B_0^{-1\prime} \otimes I_K\right) G_i, \quad i = 1, 2, \ldots \\ \overline{C}_i &= \left(I_K \otimes \Phi_i\right) H, \quad i = 0, 1, 2, \ldots \end{aligned}$$

and

$$G_i = \frac{\partial \text{vec}(\Phi_i)}{\partial \boldsymbol{\alpha}'} = \sum_{m=0}^{i-1} J(\mathbf{A}')^{i-1-m} \otimes \Phi_m, \quad i = 1, 2, \ldots$$

$$H = \frac{\partial \text{vec}(B_0^{-1})}{\partial \boldsymbol{\sigma}'} = \mathbf{L}_K' \left\{ \mathbf{L}_K \left(I_{K^2} + \mathbf{K}_{KK} \right) \left(B_0^{-1} \otimes I_K \right) \mathbf{L}_K' \right\}^{-1},$$

where

$$\Phi_m = J \mathbf{A}^m J',$$

$$\underset{K \times Kp}{J} = \left[I_K, 0_{K \times K(p-1)} \right],$$

$$\underset{Kp \times Kp}{\mathbf{A}} = \begin{bmatrix} A_1 & A_2 & \ldots & A_{p-1} & A_p \\ I_K & 0 & \ldots & 0 & 0 \\ 0 & I_K & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I_K & 0 \end{bmatrix},$$

$\mathbf{L}_K$ is the elimination matrix defined such that for any $m \times m$ matrix F, $\text{vech}(F) = \mathbf{L}_m \text{vec}(F)$, and $\mathbf{K}_{mn}$ is the commutation matrix defined such that for any $m \times n$ matrix G, $\mathbf{K}_{mn}\text{vec}(G) = \text{vec}(G')$ (see Lütkepohl 2005, appendix A12.2). If B_0^{-1} is not recursive, the expression for H must be modified accordingly.

Let diag denote the operator that retains the diagonal elements of a square matrix and stacks them into a column vector. Then the $K^2 \times 1$ vector $\sigma_{\text{vec}(\widehat{\Theta}_i)}$, consisting of the asymptotic standard errors of the elements of $\text{vec}(\widehat{\Theta}_i)$, contains the square roots of the elements of the vector

$$\text{diag}(C_i \Sigma_{\widehat{\boldsymbol{\alpha}}} C_i' + \overline{C}_i \Sigma_{\widehat{\boldsymbol{\sigma}}} \overline{C}_i')/T. \tag{12.1.3}$$

The vector of nominal $(1 - \gamma)$ pointwise asymptotic confidence intervals may be formed as

$$\text{vec}(\widehat{\Theta}_i) \pm z_{\gamma/2} \widehat{\sigma}_{\text{vec}(\widehat{\Theta}_i)}, \quad i = 0, 1, 2, \ldots,$$

where $\widehat{\sigma}_{\text{vec}(\widehat{\Theta}_i)}$ is obtained by replacing the unknowns in $\sigma_{\text{vec}(\widehat{\Theta}_i)}$ by consistent estimates, and $z_{\gamma/2}$ denotes the $\gamma/2$ quantile point of the standard normal distribution. In practice, it is most common to report one-standard error or two-standard error confidence intervals, corresponding, respectively, to (approximate) 68% or 95% two-sided confidence intervals.

The estimates $\widehat{\sigma}_{\text{vec}(\widehat{\Theta}_i)}$ may be constructed by bootstrap methods, as discussed later in this chapter, or using asymptotic closed-form solutions based

on expression (12.1.3). In applied work it is common to rely on closed-form solutions for $\Sigma_{\widehat{\alpha}}$ and $\Sigma_{\widehat{\sigma}}$ that have been derived under the assumption of Gaussian white noise errors. Specifically we replace the unknown parameters in $\Sigma_{\widehat{\alpha}} = \Gamma_Y(0)^{-1} \otimes \Sigma_u$ and $\Sigma_{\widehat{\sigma}} = 2\mathbf{D}_K^+(\Sigma_u \otimes \Sigma_u)\mathbf{D}_K^{+\prime}$ by consistent estimators. The advantage of bootstrap estimates of the standard error is that they allow us to relax the assumption of Gaussian white noise reduced-form errors, provided a suitable bootstrap method is employed.

12.1.2 Infinite-Order VAR Models

Let y_t be generated by a stationary VAR(∞) process with conditionally homoskedastic white noise errors to be approximated by a VAR model of finite lag order p_T, where p_T is assumed to increase with T at an appropriate rate (see Chapter 2). We are interested in conducting inference on the elements of $\widehat{\Theta}_i$ at horizon i, where i is fixed with respect to T. For expository purposes, let B_0^{-1} again equal the lower-triangular Cholesky decomposition of Σ_u such that $B_0^{-1}B_0^{-1\prime} = \Sigma_u$. Then

$$\sqrt{T}\text{vec}\left(\widehat{\Theta}_i(p_T) - \Theta_i\right) \xrightarrow{d} \mathcal{N}(0, \Omega_\theta(i)), \quad i = 0, 1, 2, \ldots,$$

where

$$\Omega_\theta(i) = \left(I_K \otimes \sum_{j=0}^{i-1} \Phi_j \Sigma_u \Phi_j'\right) + \overline{C}_i \Sigma_{\widehat{\sigma}} \overline{C}_i'$$

and $\Sigma_{\widehat{\sigma}}$ is the asymptotic covariance matrix of $\sqrt{T}\text{vech}(\frac{1}{T}\sum_{t=1}^T u_t u_t' - \Sigma_u)$. The $K^2 \times 1$ vector $\sigma_{\text{vec}(\widehat{\Theta}_i)}$, consisting of the asymptotic standard errors of the elements of $\text{vec}(\widehat{\Theta}_i)$, contains the square roots of the elements of the vector

$$\text{diag}(\Omega_\theta(i))/T. \tag{12.1.4}$$

A nominal $(1-\gamma)100\%$ pointwise asymptotic confidence interval may be formed as

$$\text{vec}(\widehat{\Theta}_i) \pm z_{\gamma/2}\widehat{\sigma}_{\text{vec}(\widehat{\Theta}_i)}, \quad i = 0, 1, 2, \ldots,$$

where $\widehat{\sigma}_{\text{vec}(\widehat{\Theta}_i)}$ is obtained by replacing the unknowns in $\sigma_{\text{vec}(\widehat{\Theta}_i)}$ by consistent estimates, and $z_{\gamma/2}$ denotes the $\gamma/2$ quantile of the standard normal distribution. As in the finite-order case, expression (12.1.4) may be estimated by bootstrap methods or using asymptotic closed-form solutions derived under the assumption of Gaussian white noise innovations.

12.1.3 Discussion

Our analysis highlights that different asymptotic thought experiments imply different closed-form solutions for the asymptotic standard errors of the structural impulse responses. In the case of a known and finite lag order, the asymptotic standard errors of the structural impulse responses ultimately decline toward zero as the horizon increases. In sharp contrast, if the VAR model is viewed as an autoregressive sieve approximation, the asymptotic standard errors of the structural impulse responses are a nondecreasing function of the horizon and always nonzero. This fact forces us to choose between one or the other asymptotic thought experiment when constructing delta method confidence intervals. This choice does not arise when using the bootstrap methods discussed later in this chapter. Inoue and Kilian (2002b) and Gonçalves and Kilian (2007) show that confidence intervals based on the standard bootstrap methods remain asymptotically valid under either asymptotic thought experiment. Simulation evidence for structural VAR models based on short-run identifying restrictions of the type discussed in Chapter 8 shows that approximating VAR models can provide reliable estimates of structural impulse responses generated from stationary invertible VARMA models, as long as a suitably large lag order is chosen. For further discussion of the autoregressive sieve bootstrap method, see also Paparoditis (1996) and Kreiss, Paparoditis, and Politis (2011).

12.1.4 Extensions to Other Statistics

Although we focus on inference about structural responses as our leading example, the delta method may be applied to a wide range of statistics of interest in structural VAR analysis, including cumulative structural impulse responses and structural forecast error variance decompositions (see Chapter 4). Closed-form solutions for the asymptotic standard errors of these statistics under the assumption of Gaussian white noise errors are provided in Lütkepohl (2005). Alternatively, these standard errors may be estimated by appropriate bootstrap simulation methods, as discussed in the next section. Of course, confidence intervals based on multiples of bootstrap estimates of the standard error are subject to the same caveats regarding the limiting variance-covariance matrix as the delta method.

Bootstrap methods in practice tend to be the method of choice for two reasons. First, there is no alternative to the use of bootstrap methods when dealing with statistics, for which closed-form solutions for the asymptotic standard error have yet to be derived. Second, bootstrap methods are recommended because, in practice, VAR innovations rarely are Gaussian, calling into question the use of standard closed-form solutions for the asymptotic variance derived under the assumption of Gaussian white noise errors.

12.1.5 On the Choice of the Significance Level

As noted by Leamer (1978) and Sims (1987), there is no universally valid standard for statistical significance. The choice of the significance level in statistical testing and in constructing confidence intervals depends on how tightly parameterized the model in question is and how large the sample size is. The conventional choice of a 5% significance level (corresponding to a 95% confidence interval) is merely a useful convention that facilitates the communication of the reliability of estimates from highly restricted traditional DSEMs. There is no reason to expect this convention to be useful in discussing estimates from largely unrestricted VAR models estimated on comparatively short samples. Many applied users instead favor 68% confidence intervals (corresponding to +/- one standard error bands) over 95% confidence intervals (approximately corresponding to +/- two standard error bands) in reporting VAR estimates of structural impulse responses. Occasionally, researchers report intervals for more than one significance level, but this is not always an option because it tends to make the impulse response plots look too crowded.

When bootstrap methods are used in constructing confidence intervals, as discussed in the next section, an additional consideration is that the number of draws required to accurately estimate the 2.5^{th} and 97.5^{th} percentiles of the sampling distribution used in constructing a 95% confidence interval tends to be much larger than the number of draws required for the 16^{th} and 84^{th} percentiles used in the construction of a 68% confidence interval.

12.2 Bootstrap Intervals for Structural Impulse Responses

The central idea underlying the bootstrap approach is that we approximate the distribution of the statistic of interest based on its sample analogue, allowing us to mimic the underlying sampling experiment (see, e.g., Efron and Tibshirani 1993). Most bootstrap applications rely on simulation methods in practice. One advantage of bootstrap methods in our context is that they allow inference about smooth differentiable functions $g(\boldsymbol{\alpha}, \boldsymbol{\sigma})$ even when closed-form solutions of the variance of the estimator are not available or inconvenient to use.

Another advantage is that the bootstrap approach is more general than standard asymptotic inference. For example, the closed-form solution of the variance of the structural impulse response estimator discussed in Section 12.1 relies on the VAR errors being iid Gaussian, whereas the standard residual-based bootstrap estimator remains asymptotically valid under weaker conditions.

A third advantage is that suitably constructed bootstrap confidence intervals in many applications tend to be more accurate in small samples than asymptotic approximations. In addition, under some conditions, the bootstrap methods may improve the accuracy of the asymptotic approximation.

However, bootstrap methods are no panacea. First, in some cases, they may not be applicable at all. Second, they must be used carefully. Inappropriate bootstrap applications may result in considerably less accurate inference than using first-order asymptotic approximations based on the delta method. Finally, bootstrap methods should only be used when they have asymptotic justification. Their rationale – like that of the delta method – is ultimately asymptotic and relies on the sample being large enough to be representative of the DGP.

In what follows, we again focus on the problem of inference about structural impulse responses. All bootstrap methods discussed can also be used to construct confidence intervals for related statistics such as cumulative structural impulse responses or structural forecast error variance decompositions.

12.2.1 The Standard Residual-Based Recursive-Design Bootstrap

The first study to propose the use of bootstrap methods for structural impulse responses in stationary VAR models was Runkle (1987). In fact, this work pre-dates the development of delta method intervals for structural impulse responses and was borne out of necessity, given the absence of asymptotic closed-form solutions at the time. It was not until Bose (1988) that the first-order asymptotic validity of this residual-based bootstrap method was established for autoregressive slope parameters.

Let u_t be iid white noise with distribution F, or $u_t \overset{iid}{\sim} F$ for short. We approximate the unknown stationary VAR(p) DGP of known order p,

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

where $u_t \overset{iid}{\sim} F$, conditional on the observed data, by the bootstrap DGP

$$y_t^* = \widehat{\nu} + \widehat{A}_1 y_{t-1}^* + \cdots + \widehat{A}_p y_{t-p}^* + u_t^*,$$

with $u_t^* \overset{iid}{\sim} \widehat{F}_T$. Here $[\widehat{\nu}, \widehat{A}_1, \ldots, \widehat{A}_p]$ denotes the LS estimate of the model parameters conditional on the observed sample $\{y_t\}_{t=-p+1}^{T}$ and $\widehat{F}_T$ is the implied estimate of the error distribution F. To distinguish the realizations of the bootstrap DGP from the actual realizations they are marked with a superscript *. The algorithm, as described here, allows for nonzero means in the data, but not for other deterministic regressors such as a linear time trend or seasonal dummies. Such extensions may be handled in the obvious way. Moreover, the least-squares estimator may be replaced by any other consistent estimator without affecting the analysis.

The maintained assumption is that $\mathbb{E}(u_t) = 0$ and that u_t has finite moments of the appropriate order. The estimator $\widehat{F}_T$ may be parametric or nonparametric. If we knew that $u_t \overset{iid}{\sim} \mathcal{N}(0, \Sigma_u)$, for example, then $u_t^* \overset{iid}{\sim} \mathcal{N}(0, \widehat{\Sigma}_u)$, where $\widehat{\Sigma}_u$ is the least-squares estimate of Σ_u. Draws for u_t^* would be generated by drawing from the estimated Gaussian distribution, which represents $\widehat{F}_T$.

In the more likely case that we do not know the parametric family of the distribution of the error term, there is no need to estimate F explicitly. Rather, we draw u_t^* with replacement from the set of regression residuals $\{\widehat{u}_t\}_{t=1}^T$, where $\widehat{u}_t = y_t - \widehat{\nu} - \widehat{A}_1 y_{t-1} - \cdots - \widehat{A}_p y_{t-p}$. This ensures that u_t^* in bootstrap population is iid, conditional on the data, and has the same distribution as the residuals, whatever that distribution might be. If the regression model does not include an intercept, we instead resample the demeaned residuals $\{\widehat{u}_t - \overline{u}\}_{t=1}^T$, where $\overline{u} = \sum_{t=1}^T \widehat{u}_t / T$, to ensure that the bootstrap population mean of u_t^* is zero, consistent with the mean of u_t in the DGP being zero.

Kilian (1998b) compares parametric and nonparametric resampling schemes for u_t^* in vector autoregressions and shows that in practice there is no noticeable efficiency gain from imposing parametric assumptions even when that assumption is true, whereas imposing the wrong parametric structure tends to undermine the accuracy of bootstrap inference. Hence, the nonparametric approach is strictly preferred in practice.

Given the practical importance of drawing residuals at random with replacement, it is useful to briefly review the mechanics of this procedure. For expository purposes, consider a univariate AR(p) process with innovation u_t. The set of residuals $\widehat{u}_t$ represents one sample realization of the iid process for u_t. Our objective is to replicate this sampling process conditional on the observed set of residuals. In practice, we start with the row index of the column vector $\widehat{u}$ of residuals $\widehat{u}_t$. For example, for $T = 5$ we may have

$$index = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \quad \widehat{u} = \begin{pmatrix} 0.10 \\ -0.22 \\ 0.35 \\ 0.01 \\ -0.40 \end{pmatrix}.$$

We specify a uniform distribution over the set of integers $\{1, \ldots, T\}$ in the index, ensuring that each residual has the same probability $1/T$ of being selected. For each bootstrap replication $r = 1, \ldots, R$, of length T, we create a new index, $index^{*r}$, by scaling T draws from a uniform random number generator defined on the interval $[0, 1]$ by T and rounding the results such that the scaled random draws form a column vector of numbers between 1 and T. The elements of u^{*r} are the elements in $\widehat{u}$ with row numbers given by $index^{*r}$.

In our example, we may obtain

$$index^{*r} = \begin{pmatrix} 5 \\ 2 \\ 4 \\ 3 \\ 5 \end{pmatrix} \Rightarrow u^{*r} = \begin{pmatrix} -0.40 \\ -0.22 \\ 0.01 \\ 0.35 \\ -0.40 \end{pmatrix}.$$

Note that this procedure destroys any stochastic dependence among the residuals $\widehat{u}_t$ and ensures that u_t^* is iid, consistent with the premise that u_t is iid. Although the support of the population distribution of u_t is continuous and that of u_t^* is finite, for reasonably large T, the approximation error may be safely ignored.

The ultimate objective of creating draws for u_t^* is to recreate the underlying sampling experiment that generated the observed sample $\{y_t\}_{t=-p+1}^{T}$, with the bootstrap data generating process serving as a proxy for the unknown DGP. This allows us to generate as many sequences of bootstrap realizations $\{y_t^*\}_{t=-p+1}^{T}$ as required to approximate the distribution of the statistic of interest. In practice, we use a finite number R of bootstrap replications of the sampling experiment, where R is so large that approximation error may be ignored. How large R should be depends on the statistic to be bootstrapped. For example, a bootstrap estimate of the bias or the variance of an estimator may require only $R = 200$, but estimating the 95$^{\text{th}}$ percentile of the distribution of an estimator tends to require at least $R = 2{,}000$. The reason is that there are fewer realizations in the tail of the distribution than in the center, necessitating more draws to ensure high enough accuracy.

Given random draws for u_t^{*r}, $t = 1, \ldots, T$, and initial conditions $[y_{-p+1}^{*r}, \ldots, y_0^{*r}]$, we recursively generate for each bootstrap replication $r = 1, \ldots, R$, a sequence of bootstrap realizations $\{y_t^{*r}\}_{t=-p+1}^{T}$ as

$$\begin{aligned} y_1^{*r} &= \widehat{\nu} + \widehat{A}_1 y_0^{*r} + \cdots + \widehat{A}_p y_{-p+1}^{*r} + u_1^{*r}, \\ y_2^{*r} &= \widehat{\nu} + \widehat{A}_1 y_1^{*r} + \cdots + \widehat{A}_p y_{-p+2}^{*r} + u_2^{*r}, \\ &\vdots \\ y_T^{*r} &= \widehat{\nu} + \widehat{A}_1 y_{T-1}^{*r} + \cdots + \widehat{A}_p y_{T-p}^{*r} + u_T^{*r}, \end{aligned}$$

where the additional superscript r reminds us which of the R bootstrap replications this sequence of draws refers to.

The standard approach is to draw the initial conditions $[y_{-p+1}^*, \ldots, y_0^*]$ at random with replacement as a block of p consecutive vector valued observations from the observed data $\{y_t\}_{t=-p+1}^{T}$. For each $r = 1, \ldots, R$, a new draw for $[y_{-p+1}^*, \ldots, y_0^*]$ is selected. This approach ensures that the bootstrap

presample observations are drawn from the same distribution as the remaining observations without making a parametric assumption about that distribution. Alternatively, the initial conditions may be fixed at their sample values, $[y_{-p+1}, \ldots, y_0]$ for all $r = 1, \ldots, R$. This will cause the sampling uncertainty in the statistic of interest to be understated, but in stationary models the approximation error declines with T and is negligible in the limit. Finally, one could also set the bootstrap presample observations equal to an arbitrary value such as the sample average $\bar{y} = \sum_{t=1}^{T} y_t/T$ for all $r = 1, \ldots, R$. Because subsequent realizations of y_t^* depend on the initial values, in the latter case we need to generate $\tau + T$ observations of y_t^*, where τ is chosen to allow the effects of the choice of the initial values on y_t^* to become negligible, and discard the first τ observations. The choice of τ depends both on how far the initial values are from the steady state and how persistent the process is. In practice, τ may be chosen based on visual inspection of the bootstrap data.

In this example we treated the lag order as known. If the lag order p is unknown and has to be estimated from the same data set on which the final model is estimated, we need to reestimate the lag order for each bootstrap replication $r = 1, \ldots, R$ to account for the sampling uncertainty in the lag-order estimate (see Kilian 1998a). It should be noted, however, that this approach does not overcome the problems with post-model selection inference discussed in Leeb and Pötscher (2005) (see Chapter 2).

Given a sequence of data $\{y_t^{*r}\}_{t=-p+1}^{T}$, for $r = 1, \ldots, R$, we proceed exactly as we did when confronted with the observed sample of data $\{y_t\}_{t=-p+1}^{T}$. We fit the VAR(p) model with intercept to obtain the LS estimates $[\widehat{\nu}^{*r}, \widehat{A}_1^{*r}, \ldots, \widehat{A}_p^{*r}]$ and $\widehat{\Sigma}_u^{*r}$. This allows us to construct the implied bootstrap estimates $\widehat{\theta}_{ik,h}^{*r}$, $r = 1, \ldots, R$, of the structural impulse responses, as discussed earlier. From the bootstrap approximation to the unknown empirical distribution of the structural impulse response estimates $\widehat{\theta}_{ik,h}$ we construct confidence intervals. Runkle (1987) in particular reported the percentile interval

$$\left[\widehat{\theta}_{ik,h,\gamma/2}^{*}, \widehat{\theta}_{ik,h,1-\gamma/2}^{*}\right],$$

where $\widehat{\theta}_{ik,h,\gamma/2}^{*}$ and $\widehat{\theta}_{ik,h,1-\gamma/2}^{*}$ are the critical points defined by the $\gamma/2$ and $1 - \gamma/2$ quantiles of the distribution of $\widehat{\theta}_{ik,h}^{*r}$. Note that the validity of this approach hinges on the distribution of $\widehat{\theta}_{ik,h}^{*r}$ becoming close to the asymptotic normal distribution of $\widehat{\theta}_{ik,h}$, as the sample size goes to infinity. This will typically be the case in the stationary VAR(p) model with some exceptions noted below. We will return to the question of how to evaluate the distribution of the bootstrap estimator of interest in more detail later in the chapter. For now our focus is on the question of how to generate bootstrap approximations to the sampling distribution.

12.2.2 The Standard Residual-Based Fixed-Design Bootstrap

An alternative and less common approach is to treat the design matrix as fixed in repeated sampling. Under stationarity, ignoring the randomness of the regressors does not affect the asymptotic validity of the bootstrap. This allows us to approximate the unknown stationary VAR(p) data generating process

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

where $u_t \overset{iid}{\sim} F$, conditional on the observed data, by the bootstrap DGP

$$y_t^* = \widehat{\nu} + \widehat{A}_1 y_{t-1} + \cdots + \widehat{A}_p y_{t-p} + u_t^*$$

with $u_t^* \overset{iid}{\sim} \widehat{F}_T$. Bootstrap samples are created by adding the resampled residuals to the fitted model, holding the regressor matrix fixed in repeated sampling. Given a sequence of data $\{y_t^{*r}\}_{t=-p+1}^{T}$, we regress y_t^{*r} on the original regressors $(1, y_{t-1}, \ldots, y_{t-p})$ to obtain the least-squares estimates $[\widehat{\nu}^{*r}, \widehat{A}_1^{*r}, \ldots, \widehat{A}_p^{*r}]$ and $\widehat{\Sigma}_u^{*r}$. After repeating this procedure for $r = 1, \ldots, R$, we construct the empirical distribution of $\widehat{\theta}_{ik,h}^{*r}$, $r = 1, \ldots, R$, and the implied bootstrap confidence intervals as before. Simulation evidence in Gonçalves and Kilian (2004) suggests that this algorithm is slightly less accurate in finite samples than the recursive-design bootstrap for autoregressive processes.

12.2.3 The Residual-Based Wild Bootstrap

Both bootstrap approaches discussed so far require the regression error to be iid. This assumption is quite strong, especially when working with monthly data that are often characterized by conditional heteroskedasticity of unknown form. An alternative approach is based on the wild bootstrap. The wild bootstrap was originally developed in the context of the classical linear regression model (see Wu 1986). Its main use has been to make inference in cross-sectional regressions robust to unconditional heteroskedasticity in the error term (see Davidson and Flachaire 2008). Its asymptotic validity for the parameters of dynamic regression models with conditional heteroskedasticity was established in Gonçalves and Kilian (2004, 2007).

The wild bootstrap may be used in the context of fixed regressor bootstrap methods as well as recursive bootstrap methods. The only difference compared with the methods discussed earlier is how we create draws for u_t^*. The central idea is to multiply each element of the residual vector by a scalar draw, η_t, from an auxiliary distribution (or external distribution) that has mean zero and variance 1 such as the $\mathcal{N}(0, 1)$ distribution. Thus

$$u_t^{*r} = \left(u_{1t}^{*r}, u_{2t}^{*r}, \ldots, u_{Kt}^{*r}\right)',$$

where $u_t^{*r} \equiv \widehat{u}_t \eta_t$ and $\eta_t \overset{iid}{\sim} (0, 1)$. Other examples of possible η_t distributions include the two-point distribution $\eta_t = -(\sqrt{5} - 1)/2$ with probability $\mathcal{P} = (\sqrt{5} + 1)/(2\sqrt{5})$ and $\eta_t = (\sqrt{5} + 1)/2$ with probability $1 - \mathcal{P}$, as proposed by Mammen (1993), and the two-point distribution $\eta_t = 1$ with probability $\mathcal{P} = 0.5$ and $\eta_t = -1$ with probability $\mathcal{P} = 0.5$, as proposed by Liu (1988). There is no indication that the choice of η_t makes a difference in practice (see Gonçalves and Kilian 2004).

When applying the wild bootstrap, there is no need to demean the residuals even when no intercept is included in the regression model. The vector $\widehat{u}_t$ is treated as a constant in the bootstrap world. Hence, conditional on the data, $\mathbb{E}^*(u_t^*) = 0$ and $V(u_t^*) = \mathbb{E}^*(u_t^{*2}) = \widehat{u}_t^2 \cdot 1 = \widehat{u}_t^2$, where $\mathbb{E}^*$ is the expectation under the bootstrap probability measure. This feature of the wild bootstrap is closely related to the robust variance estimator proposed by Nicholls and Pagan (1983). The resulting bootstrap distributions may be evaluated in the usual manner. Any t-statistics or Wald statistics based on the wild bootstrap, of course, will have to be computed based on heteroskedasticity-robust standard errors, not the usual LS standard errors.

The wild bootstrap approach is preferable to modeling the error term of the VAR model as a parametric multivariate GARCH model or stochastic volatility model. Multivariate GARCH models, for example, typically require strong and ad hoc restrictions on the conditional covariance matrix for estimation to be computationally feasible. Moreover, there are many different GARCH and stochastic volatility specifications that are mutually exclusive. The wild bootstrap is nonparametric, it accounts for shifts in the volatility of the data, it avoids ad hoc modeling choices, and it has high finite-sample accuracy even when the regression error is iid (see Gonçalves and Kilian 2004, 2007).

The residual-based wild bootstrap was originally designed to approximate the distribution of the autoregressive slope parameters. One important drawback of this method is that it will recover the asymptotic distribution of the reduced-form impulse responses in the limit, but not the asymptotic distribution of the error variance-covariance matrix, as shown in Brüggemann, Jentsch, and Trenkler (2016). Intuitively, the problem is that the wild bootstrap does not correctly replicate the relevant fourth-moment structure of the VAR innovations. As a result, the wild bootstrap fails to recover the second component of the limiting variance in

$$\sqrt{T}\left(g(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\sigma}}) - g(\boldsymbol{\alpha}, \boldsymbol{\sigma})\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g}{\partial \boldsymbol{\alpha}'} \Sigma_{\widehat{\boldsymbol{\alpha}}} \frac{\partial g'}{\partial \boldsymbol{\alpha}} + \frac{\partial g}{\partial \boldsymbol{\sigma}'} \Sigma_{\widehat{\boldsymbol{\sigma}}} \frac{\partial g'}{\partial \boldsymbol{\sigma}}\right).$$

Because the wild bootstrap underestimates the asymptotic variance of the error variance, for statistics of the form $g(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\sigma}})$ it produces confidence intervals that are too narrow and have coverage accuracy that is too low asymptotically. It is

worth pointing out that, in practice, the uncertainty about the error covariance matrix quickly dies out as the horizon lengthens, such that the residual-based wild bootstrap adequately captures the uncertainty about structural impulse responses at all but the lowest horizons. Although the residual-based wild bootstrap is a clear improvement on the iid residual-based bootstrap in the presence of conditional heteroskedasticity, it, nevertheless, is unsatisfactory. Brüggemann, Jentsch, and Trenkler (2016) propose an alternative bootstrap method intended to ensure high asymptotic coverage accuracy for structural impulse responses at all horizons that will be discussed in the section on block bootstrap methods.

12.2.4 Bootstrapping Tuples of Regressands and Regressors

An alternative approach that is robust to heteroskedasticity of unknown form involves bootstrapping tuples consisting of the regressand and the regressors. This method is a natural generalization of the pairwise bootstrap for linear regression first suggested by Freedman (1981). Its asymptotic validity for autoregressions is established in Gonçalves and Kilian (2004).

Consider the unknown stationary VAR(p) DGP

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

It is clear that when the data are dependent over time, one cannot resample each date t observation by drawing with replacement without destroying the dependence structure in the data. This problem may be circumvented by forming tuples

$$x_t \equiv \left(y_t', 1, y_{t-1}', \ldots, y_{t-p}'\right),$$

of dimension $1 \times (Kp + K + 1)$ for $t = 1, \ldots, T$ and stacking them into a matrix X. Note that x_t contains a 1 to account for the intercept term in the model. As long as the regression error is not serially correlated, we can draw with replacement from the T rows of the matrix X to form a bootstrap sample matrix X^{*r} with elements

$$x_t^{*r} \equiv \left(y_t^{*r\prime}, 1, y_{t-1}^{*r\prime}, \ldots, y_{t-p}^{*r\prime}\right), \quad t = 1, \ldots, T.$$

The first K columns of the resulting matrix X^{*r} correspond to the $T \times K$ dependent variable $\{y_t^{*r\prime}\}_{t=1}^T$, while the remaining columns form the $T \times (Kp + 1)$ matrix of bootstrap regressors. This is all we need to form the bootstrap LS estimates $[\widehat{\nu}^{*r}, \widehat{A}_1^{*r}, \ldots, \widehat{A}_p^{*r}]$ and $\widehat{\Sigma}_u^{*r}$. After repeating this procedure for $r = 1, \ldots, R$, we construct the empirical distribution of $\widehat{\theta}_{ik,h}^{*r}$, $r = 1, \ldots, R$, and the implied bootstrap confidence intervals as before.

Note that we never specified the error term u_t^* in this algorithm. This means that this algorithm will be valid for regression models with white noise errors,

even if the error term is not iid, but conditionally heteroskedastic. Of course, methods that exploit the iid assumption when it is true in population will tend to be more efficient and hence more accurate in small samples. Like the fixed-design bootstrap, this alternative bootstrap is less parametric and hence slightly less accurate when the autoregressive model is the data generating process, as shown in Gonçalves and Kilian (2004). Like the wild bootstrap, however, it fails to capture the full uncertainty in the residual covariance matrix and hence is not designed for inference on structural impulse responses (see Brüggemann, Jentsch, and Trenkler 2016).

12.2.5 Block Bootstrap Methods

If the innovations of the model are serially correlated, none of the preceding bootstrap methods are applicable. An obvious response would be to increase the VAR lag order appropriately before applying conventional bootstrap methods. An alternative approach, designed to protect against serial correlation in the error term, is the use of block bootstrap methods. Block bootstrap methods require the underlying DGP to be $I(0)$. It may seem that the use of the block bootstrap would always be preferred, given its robustness to departures from the iid error assumption, but this is not the case because this robustness comes at a cost. For example, the use of block bootstrap methods would be inefficient if we were concerned with serial correlation of known form in the error term.

There are different types of block bootstrap methods in the literature, only some of which are relevant for working with VAR models. In explaining the block bootstrap and how it may be applied in the context of VAR models, it is useful to start with a much simpler example.

Naive Block Bootstrap. The most common form of the bootstrap is based on the exchangeability of iid observations. Usually, in the case of dependent data, we parameterize the dependence by imposing an autoregressive structure, so that the problem can be reduced to one of resampling residuals that are iid in population.

An alternative strategy is to re-sample blocks of observations rather than individual observations. This simplest block bootstrap approach is asymptotically valid for statistics that are symmetric functions of time series observations generated from covariance stationary processes. Symmetry here means that the statistic does not depend on the order of the observations. A good example is the sample mean:

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t.$$

A good counterexample is the autocorrelation coefficient. To cement ideas, we begin with a simple example. Consider a sample $y_1, \ldots, y_T$ generated by a covariance stationary scalar process $\{y_t\}_{t=-\infty}^{\infty}$, which may or may not be an autoregressive process. The procedure involves three simple steps:

1. Suppose that we observe the sample

$$(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}, y_{12}, y_{13}, y_{14}, y_{15})'.$$

We begin by defining $b = T - l + 1$ overlapping blocks of length l.[1] Let $l = 5$. Then $b = 15 - 5 + 1 = 11$ and the set of blocks we will draw from is:

$$\begin{aligned}
x_1 &= (y_1, y_2, y_3, y_4, y_5) \\
x_2 &= (y_2, y_3, y_4, y_5, y_6) \\
x_3 &= (y_3, y_4, y_5, y_6, y_7) \\
x_4 &= (y_4, y_5, y_6, y_7, y_8) \\
&\vdots \\
x_9 &= (y_9, y_{10}, y_{11}, y_{12}, y_{13}) \\
x_{10} &= (y_{10}, y_{11}, y_{12}, y_{13}, y_{14}) \\
x_{11} &= (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})
\end{aligned}$$

Bootstrap replicates of the original sample may be generated by sampling with replacement from the set $\{x_t\}_{t=1}^{b}$. We then concatenate the iid draws, denoted by x_t^*, to form a bootstrap replication of the original time series of length $T = sl$ of the form $\{y_t^*\}_{t=1}^{T} = \left(x_1^*, x_2^*, \ldots, x_s^*\right)'$, where $s = T/l$. For example,

$$(y_9, y_{10}, y_{11}, y_{12}, y_{13} | y_2, y_3, y_4, y_5, y_6 | y_5, y_6, y_7, y_8, y_9)'$$

could be the r^{th} bootstrap replication of the original time series. In the example, $s = 15/5 = 3$, so each bootstrap time series of length T consists of three blocks.

2. Given this bootstrap time series, we construct

$$\bar{y}^{*r} = \frac{1}{T} \sum_{t=1}^{T} y_t^*.$$

3. This procedure is repeated for $r = 1, \ldots, R$, allowing us to build up the bootstrap distribution of the sample mean.

[1] Although we focus on the case of overlapping blocks, which is most common in practice, it should be noted that block bootstrap methods can also be applied to nonoverlapping blocks. The latter approach makes less efficient use of the data, however.

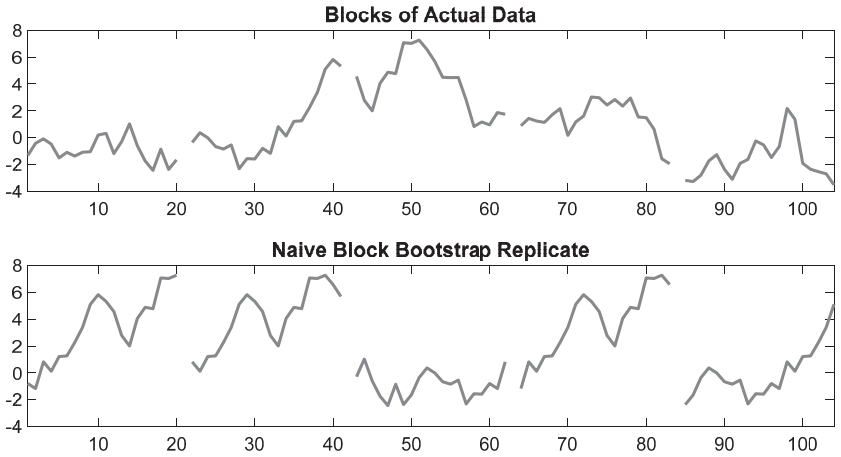

Figure 12.1. Example of how resampling overlapping blocks may destroy the dependence structure in the bootstrap data at the point of transition from one block to the next.

Notes: The data were generated from a stationary AR(1) model. The plot was created by inserting an extra space in between any two blocks to allow visual separation of each block.

This procedure was originally introduced by Künsch (1989). The asymptotic validity of this bootstrap approach for stationary data requires $s \to \infty$ and $l \to \infty$ at a suitable rate, as $T \to \infty$, where s is the number of blocks of length l such that $T = sl$.

Limitations of the Naive Block Bootstrap. It is easy to see why this idea does not generalize easily to statistics such as autoregressive coefficients that are not symmetric functions of the data. For illustrative purposes, consider the time series shown in the upper panel of Figure 12.1, which was generated from an AR(1) process with a parameter of 0.95. The iid innovations of this process were drawn from a $\mathcal{N}(0, 1)$ distribution. The sample size is $T = 100$. For expository purposes the sample has been divided into five blocks of length $l = 20$.

The lower panel of Figure 12.1 displays a representative replicate of bootstrap data obtained by drawing with replacement from the set of overlapping blocks of length $l = 20$ constructed from the original time series. This figure illustrates a major problem with the naive block bootstrap applied to statistics such as autocorrelation coefficients or autoregressive slope coefficients. As we paste one block of data to the next, we destroy the time series dependence. Clearly, at these points there are discontinuous jumps in the time series not found in the original data, so the autocorrelation structure of the bootstrap data differs systematically from that found in the original DGP.

If one lets the block length l increase with the sample size, this problem can be mitigated at least asymptotically. This is of little help, however, when the sample is short and the data are persistent. On the one hand, we need the blocks to be long enough for the original time series dependence to be preserved in the resampled series. On the other hand, in short samples, large l makes for few estimates of the statistic of interest and hence poor approximations to the empirical distribution. Thus the researcher finds himself between a rock and a hard place.

Points of discontinuity are not a problem for statistics such as the sample mean that do not depend on the order of the time series observations, but they call into question applications of the naive block bootstrap to nonsymmetric statistics such as autocorrelations or estimators of autoregressive slope coefficients. There are two modifications of the original block bootstrap idea designed to deal with this problem. One is the blocks-of-blocks bootstrap; the other is the residual based block bootstrap.

Blocks-of-Blocks Bootstrap. Nonsymmetric statistics such as autocorrelations or estimators of autoregressive slope coefficients can be expressed as a symmetric function of m-dimensional blocks of consecutive observations, $m > 1$. This fact suggests that, when dealing with nonsymmetric statistics, we resample blocks of l consecutive m-tuples drawn from the set of all possible m-tuples. This amounts to a blocks-of-blocks bootstrap. The asymptotic validity of this procedure for stationary processes requires letting the block size l increase at a suitable rate, as $T \to \infty$.

Suppose that the sample data $\{y_t\}_{t=1}^{T}$ have already been demeaned. We are interested in the first-order autocorrelation coefficient which depends on y_t and y_{t-1}. This fact tells us that $m = 2$. Further suppose that $T = 16$. Construct the (15×2) matrix of m-tuples by stacking y_t next to y_{t-1}:

$$X = \begin{bmatrix} 0.10 & -0.3 \\ -0.5 & 0.1 \\ 0.4 & -0.5 \\ 0 & 0.4 \\ -0.1 & 0 \\ \vdots & \vdots \\ -0.2 & 0.11 \\ 0.05 & -0.2 \\ 0.1 & 0.05 \end{bmatrix}.$$

Rather than drawing with replacement from the rows of this matrix, we draw with replacement from the set of overlapping blocks of rows of this matrix. For

example, for $l = 3$, we might draw a bootstrap replicate:

$$index^* = \begin{pmatrix} 13 \\ 14 \\ \underline{15} \\ 2 \\ 3 \\ \underline{4} \\ 3 \\ 4 \\ 5 \end{pmatrix} \Rightarrow X^* = \begin{bmatrix} -0.2 & 0.11 \\ 0.05 & -0.2 \\ 0.1 & 0.05 \\ -0.5 & 0.1 \\ 0.4 & -0.5 \\ 0 & 0.4 \\ 0.4 & -0.5 \\ 0 & 0.4 \\ -0.1 & 0 \end{bmatrix}.$$

Having generated X^*, we substitute the first and second column of X^* into the formula for the first-order autocorrelation coefficient to generate the r^{th} replicate of that statistic. Note that when we use the blocks-of-blocks bootstrap, we never actually construct the bootstrap realizations of the data $\{y_t^*\}_{t=1}^T$. Rather, we directly substitute the re-sample into the function that defines the statistic of interest in terms of the m-tuples.

In applications of the blocks-of-blocks bootstrap, the statistic of interest determines the choice of m. In the example of the first-order autocorrelation, $m = 2$. When considering an AR(3) model, on the other hand, we would have $m = 4$. This leaves the task of choosing l. This problem is analogous to the choice of the truncation lag for heteroskedasticity and autocorrelation consistent (HAC) standard errors (see Andrews 1991). There are no good rules of thumb for choosing l in general. In practice, it is advisable to check the sensitivity of the results with respect to l.

The blocks-of-blocks bootstrap idea may be easily generalized to the problem of bootstrapping structural impulse responses in vector autoregressions. In that case

$$X = \begin{bmatrix} y_1' & 1 & y_0' & \cdots & y_{-p+1}' \\ y_2' & 1 & y_1' & \cdots & y_{-p+2}' \\ \vdots & \vdots & \vdots & & \vdots \\ y_T' & 1 & y_{T-1}' & \cdots & y_{T-p}' \end{bmatrix}.$$

It is convenient to include the intercept in X, such that X is of dimension $T \times (Kp + K + 1)$. The choice of l depends on how much serial correlation we expect in the error term. Typically, in VAR(p) models with large p, there is not much serial correlation left in the error term, so l tends to be small. Having determined l, we generate X^* by drawing with replacement from blocks of l adjacent rows of X. Note that this approach may also be viewed as a generalization of the earlier idea of bootstrapping tuples of regressands and

regressors that allows for serial correlation in the error term of an autoregression.

Given X^{*r}, we recover the regressand and regressors, reestimate the VAR(p) model on the bootstrap data, and compute $\widehat{\theta}_{ik,h}^{*r}$ from $[\widehat{\nu}^{*r}, \widehat{A}_1^{*r}, \ldots, \widehat{A}_p^{*r}]$ and $\widehat{\Sigma}_u^{*r}$. After repeating this procedure for $r = 1, \ldots, R$, we construct the empirical distribution of $\widehat{\theta}_{ik,h}^{*r}$, $r = 1, \ldots, R$, and the implied bootstrap confidence intervals as before. Although the implementation of this method is straightforward, we are not aware of applications to vector autoregressions.

The block bootstrap is based on blocks of fixed length l. The stationary bootstrap is a variation of the block bootstrap with the block size l modeled as a random variable with a geometric distribution with mean l. The stationary bootstrap in practice does not improve the accuracy of the block bootstrap approximation for autoregressions and may worsen it in small samples (see Berkowitz, Birgean, and Kilian 2000).

Residual-Based Block Bootstrap. An alternative approach is to apply the original block bootstrap not to the data directly, but rather to the residuals of an autoregression. The idea of the residual-based block bootstrap can be traced to Lahiri (1996). This approach has also been referred to as a pre-whitened bootstrap because fitting the autoregression removes much of the persistence of the data that calls into question the application of the original block bootstrap (see, e.g., Berkowitz, Birgean, and Kilian 2000). Although a number of closely related approaches have been discussed in the literature, there is no reference that explicitly discusses this method and derives its asymptotic properties for the stationary VAR(p) model.

This residual block bootstrap procedure may be summarized as follows. Consider the VAR approximation

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

where u_t is assumed to be a linear stationary process with innovations that are mean zero and satisfy suitable restrictions on their moments. The error u_t is also assumed to satisfy a suitable strong mixing condition (see, e.g., Jentsch, Politis, and Paparoditis 2015).[2] Having fitted the VAR(p) model to the data and having recovered estimates of the model parameters and of the model residuals

$$\widehat{u}_t = y_t - \widehat{\nu} - \widehat{A}_1 y_{t-1} - \cdots - \widehat{A}_p y_{t-p}$$

[2] Strong mixing conditions ensure a sufficiently fast rate of decay in the dependence between u_t and u_s as the distance between t and s grows, such that observations sufficiently far apart in time are nearly independent (see Davidson 1994).

for $t = 1, \ldots, T$, we arrange the residuals in the form of a matrix

$$\begin{bmatrix} \widehat{u}_1 & \widehat{u}_2 & \ldots & \widehat{u}_l \\ \widehat{u}_2 & \widehat{u}_3 & \ldots & \widehat{u}_{l+1} \\ \vdots & \vdots & & \vdots \\ \widehat{u}_{T-l+1} & \widehat{u}_{T-l+2} & \ldots & \widehat{u}_T \end{bmatrix},$$

where each row denotes a block of l consecutive residuals and the blocks are overlapping. The number of blocks is $s = T - l + 1$. The bootstrap innovations are obtained by drawing at random with replacement from the rows of this matrix, laying these blocks of residuals end to end, and retaining the first T bootstrap innovations.

The resulting bootstrap innovation draws, denoted by $[\widetilde{u}_1^*, \ldots, \widetilde{u}_T^*]$, however, must be recentered to ensure that $\mathbb{E}^*(u_t^*) = 0$ for all $t = 1, \ldots, T$, where $\mathbb{E}^*$ denotes the expectation under the bootstrap probability measure. The reason is that the bootstrap innovations must have the same population mean as the original population innovations for the bootstrap analogy to hold. This may be accomplished by constructing the bootstrap innovations, $[u_1^*, \ldots, u_T^*]'$, as

$$u_{jl+i}^* = \widetilde{u}_{jl+i}^* - \mathbb{E}^*(\widetilde{u}_{jl+i}^*) = \widetilde{u}_{jl+i}^* - \frac{1}{T-l+1} \sum_{r=0}^{T-l} \widehat{u}_{i+r}$$

for $i = 1, 2, \ldots, l$ and $j = 0, 1, \ldots, s-1$. The corresponding bootstrap data, $[y_{-p+1}^*, \ldots, y_T^*]$, are then generated recursively from

$$y_t^* = \widehat{\nu} + \widehat{A}_1 y_{t-1}^* + \cdots + \widehat{A}_p y_{t-p}^* + u_t^*.$$

Bootstrap realizations of the estimators of interest are obtained by fitting a VAR(p) model to this sequence of bootstrap data. Finite-sample evidence for this approach can be found in Berkowitz, Birgean, and Kilian (2000).[3]

A generalization of the residual-based block bootstrap idea was proposed by Yeh (1998). The residual-based wild block bootstrap allows for the errors of the regression model to be nonstationary. Nonstationarity here does not refer to allowing for unit roots in the process governing u_t, but to allowing for heterogeneity in the innovations across time. An example would be a serially correlated error process subject to unconditional heteroskedasticity. The theoretical results in Yeh (1998) are limited to a fixed regression design and nonoverlapping blocks of residuals. As before, given model residuals

$$\widehat{u}_t = y_t - \widehat{\nu} - \widehat{A}_1 y_{t-1} - \cdots - \widehat{A}_p y_{t-p}$$

[3] Jentsch, Politis, and Paparoditis (2015) discuss extensions of the block bootstrap approach to multivariate integrated and cointegrated models. There are, at present, no formal results, however, about the validity of conducting inference about structural impulse responses in cointegrated VAR models based on the residual-based block bootstrap.

for $t = 1, \ldots, T$, we arrange the residuals in the form of a matrix

$$\begin{bmatrix} \widehat{u}_1 & \widehat{u}_2 & \ldots & \widehat{u}_l \\ \widehat{u}_{l+1} & \widehat{u}_{l+2} & \ldots & \widehat{u}_{2l} \\ \vdots & \vdots & & \vdots \\ \widehat{u}_{T-l+1} & \widehat{u}_{T-l+2} & \ldots & \widehat{u}_T \end{bmatrix},$$

where each row denotes a block of l consecutive residuals, but now the blocks are nonoverlapping and the number of blocks is $s = [T/l]$, where s denotes the largest integer smaller or equal to T/l.

Bootstrap innovations are created by post-multiplying each row of this matrix by an independent draw from an external random variable with mean zero and variance 1 such as a draw from a $\mathcal{N}(0, 1)$ distribution. Let $\eta_1, \ldots, \eta_s$ denote the sequence of realizations of these $\mathcal{N}(0, 1)$ draws. Then the bootstrap innovations, $[u_1^*, \ldots, u_T^*]$, are obtained by laying the rows of the matrix

$$\begin{bmatrix} \widehat{u}_1\eta_1 & \widehat{u}_2\eta_1 & \ldots & \widehat{u}_l\eta_1 \\ \widehat{u}_{l+1}\eta_2 & \widehat{u}_{l+2}\eta_2 & \ldots & \widehat{u}_{2l}\eta_2 \\ \vdots & \vdots & & \vdots \\ \widehat{u}_{T-l+1}\eta_s & \widehat{u}_{T-l+2}\eta_s & \ldots & \widehat{u}_T\eta_s \end{bmatrix},$$

end to end. No centering is required. The corresponding bootstrap data, $[y_{-p+1}^*, \ldots, y_T^*]$, are generated from

$$y_t^* = \widehat{v} + \widehat{A}_1 y_{t-1} + \cdots + \widehat{A}_p y_{t-p} + u_t^*,$$

treating the regression design as fixed in repeated sampling. Bootstrap realizations of the estimators of interest are obtained by regressing $\{y_t^*\}_{t=1}^T$ on the original regressor matrix. Unlike the residual-based block bootstrap, this procedure is not applicable to structural impulse response analysis. It only captures the uncertainty about the VAR slope parameters (and smooth functions thereof).

The Residual-Based Block Bootstrap in Conditionally Heteroskedastic VAR Models. Brüggemann, Jentsch, and Trenkler (2016) show that the residual-based bootstrap may also be used to achieve accurate inference about structural impulse responses in stationary VAR models with conditionally heteroskedastic errors, but no serial correlation in the errors. A precise statement of these conditions can be found in their paper. Their proposal involves fitting to the data the VAR(p) model

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

where u_t is serially uncorrelated, but conditionally heteroskedastic, and to recover estimates of the model parameters and the model residuals

$$\widehat{u}_t = y_t - \widehat{\nu} - \widehat{A}_1 y_{t-1} - \cdots - \widehat{A}_p y_{t-p}$$

for $t = 1, \ldots, T$. They then choose a block length $l < T$ such that $s = [T/l]$ is the number of blocks where $ls \geq T$, and arrange blocks of length l of the model residuals in the form of the matrix

$$\begin{bmatrix} \widehat{u}_1 & \widehat{u}_2 & \ldots & \widehat{u}_l \\ \widehat{u}_2 & \widehat{u}_3 & \ldots & \widehat{u}_{1+l} \\ \vdots & \vdots & & \vdots \\ \widehat{u}_{T-l+1} & \widehat{u}_{T-l+2} & \ldots & \widehat{u}_T \end{bmatrix}.$$

Bootstrap innovations are generated by drawing s times with replacement from the rows of this matrix. The time series of bootstrap innovations, $[\widetilde{u}_1^*, \ldots, \widetilde{u}_T^*]$, is obtained by laying these randomly chosen blocks end to end and retaining the first T bootstrap innovations. The resulting bootstrap innovation draws again must be recentered to ensure that $\mathbb{E}^*(u_t^*) = 0$ for all $t = 1, \ldots, T$. We construct the bootstrap innovations, $[u_1^*, \ldots, u_T^*]$, from

$$u_{jl+i}^* = \widetilde{u}_{jl+i}^* - \mathbb{E}^*(\widetilde{u}_{jl+i}^*) = \widetilde{u}_{jl+i}^* - \frac{1}{T-l+1} \sum_{r=0}^{T-l} \widehat{u}_{i+r}$$

for $i = 1, 2, \ldots, l$ and $j = 0, 1, \ldots, s-1$. Given an appropriate draw for the bootstrap presample observations, one can generate the bootstrap sample, $[y_{-p+1}^*, \ldots, y_T^*]$, from

$$y_t^* = \widehat{\nu} + \widehat{A}_1 y_{t-1}^* + \cdots + \widehat{A}_p y_{t-p}^* + u_t^*,$$

and compute the bootstrap estimators of interest after fitting a VAR(p) model to this sequence of data. Simulation evidence suggests that the coverage accuracy of structural impulse response confidence intervals based on the residual-based block bootstrap method is consistently high only in fairly large samples.

12.2.6 Alternative Bootstrap Confidence Intervals

So far we have focused mainly on the question of how to generate bootstrap approximations to the distribution of the statistic of interest. An important question is how to construct bootstrap confidence intervals from this distribution. There are a number of options ranging from bootstrap standard error bands to percentile and percentile-t intervals, all of which could be employed in conjunction with any of the resampling methods discussed so far.

For notational simplicity, in what follows, we refer to the scalar structural VAR impulse response coefficient $\theta_{jk,i}$ as θ, suppressing the subscripts.

$\widehat{\theta}$ denotes a consistent and asymptotically normal estimator of this scalar parameter θ, and $\widehat{\theta}^*$ is the bootstrap version of this estimator.

Intervals Based on Bootstrap Standard Errors. A particularly simple approach is to compute the standard deviation of the bootstrap draws of $\widehat{\theta}^*$ numerically. This allows one to form confidence intervals that are centered on the point estimate of the structural impulse response:

$$\widehat{\theta} \pm z_{\gamma/2}\widehat{\sigma}(\widehat{\theta}^*).$$

Effectively, the only difference to the delta method interval is that we have replaced the closed-form estimator of the standard deviation. The bootstrap estimator and the closed-form estimator are asymptotically equivalent under Gaussian white noise errors. One advantage of the bootstrap estimator is that it allows us to relax the assumption of Gaussian iid innovations underlying the conventional delta method interval, for example, by resampling the residuals with replacement or by using the residual-based block bootstrap method. The other advantage is that the bootstrap estimator may be used even in the absence of a closed-form solution for the asymptotic variance.

Efron's Percentile Interval. The most common form of bootstrap confidence intervals is the percentile interval proposed by Efron (1979), which is also the interval employed in Runkle (1987). Because subsequently, alternative "percentile intervals" have been proposed in the literature, we refer to this interval as Efron's percentile interval. Let $\widehat{\theta}^*_{\gamma/2}$ and $\widehat{\theta}^*_{1-\gamma/2}$ be the critical points defined by the $\gamma/2$ and $1-\gamma/2$ quantiles of the distribution of $\widehat{\theta}^*$. Then Efron's percentile interval is

$$CI^{\text{Efron}}_{PER} = \left[\widehat{\theta}^*_{\gamma/2},\ \widehat{\theta}^*_{1-\gamma/2}\right].$$

This interval was designed to be asymptotically valid for inference about consistent estimators with an asymptotic normal approximation of the form

$$\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2(\widehat{\theta})).$$

It is readily apparent that the interval endpoints of the percentile interval will agree with the interval endpoints of the delta method interval asymptotically, if the statistic of interest is asymptotically Gaussian and centered on its population value. As stressed by Efron, however, this percentile interval (as well as its variants discussed later) remains valid even when the distribution of $\widehat{\theta}$ is not normal, as long as normality can be achieved by applying a monotonic transformation to θ and to its estimator $\widehat{\theta}$.[4] The percentile interval automatically recovers the appropriate interval endpoints without requiring the user to

[4] An example would be Fisher's inverse hyperbolic tangent transformation for the correlation coefficient from a bivariate normal sample (see Efron 1982).

know the correct transformation. In this sense, it is more general than conventional asymptotic normal approximations. If there is no suitable transformation that satisfies this condition, in contrast, as would be the case when the statistic of interest is biased, the CI_{PER}^{Efron}-interval will not be valid and may have poor coverage accuracy.

It immediately follows that caution is called for in applying this method when conducting inference about structural impulse responses. The finite-sample distribution of structural impulse response estimators, $\widehat{\theta}_{ik,h}$, may be far from Gaussian (see Kilian 1999b). A particular concern is the small-sample bias of the structrual impulse response estimator. When modeling autoregressions, bias is not a concern asymptotically, because the small-sample bias of the least-squares estimator is of order T (see Section 2.3.3), but it is an obvious concern for the finite-sample accuracy of the CI_{PER}^{Efron}-interval, which is based on the premise of an unbiased estimator.

Efron (1982) proposed a generalization of the original percentile interval intended to allow for bias in the estimator, while retaining the assumption of approximate normality (possibly after a monotonic transformation), referred to as the bias-corrected or *BC* percentile interval. His proposal was to adjust the endpoints of the percentile interval depending on the median bias of the estimator $\widehat{\theta}$. This solution presumes that we are dealing with a pure location shift and that the variance of the estimator is invariant with respect to the value of θ. This assumption is clearly violated when dealing with smooth functions of slope parameters in vector autoregressions, suggesting that such bias corrections will not help in our context. Indeed, Kilian (1998c) showed by simulation that the *BC*-interval is not reliable in small samples when conducting inference about structural impulse responses.

In related work, Efron (1987) therefore proposed a further refinement of the percentile interval, referred to as the accelerated *BC*-interval or BC_α interval (to which an analytic approximation was developed in later years under the name of *ABC* interval). Compared with the *BC*-interval, the BC_α-interval involves an additional acceleration constant, α, that captures changes in the variance of the estimator, as the location of the distribution changes. The BC_α-interval was designed for inference on statistics of the form

$$\frac{\widehat{\theta}-\theta}{\sigma(\widehat{\theta})} \sim \mathcal{N}(-z_0, 1),$$

where $\sigma(\widehat{\theta}) = \sigma(\theta_0)[1+\alpha(\widehat{\theta}-\theta_0)/\sigma(\theta_0)]$ and θ_0 is any convenient reference point on the scale of θ values. Let $\Phi(\cdot)$ denote the standard normal cumulative distribution function and z_γ denote the $100\gamma^{\text{th}}$ percentile point of the standard normal distribution. Then the BC_α-interval is defined as

$$CI_{BC_\alpha} = \left[\widehat{\theta}^*_{\gamma_1},\ \widehat{\theta}^*_{\gamma_2}\right],$$

with

$$\gamma_1 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{\gamma/2}}{1 - \widehat{\alpha}\left(\widehat{z}_0 + z_{\gamma/2}\right)}\right)$$

and

$$\gamma_2 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{1-\gamma/2}}{1 - \widehat{\alpha}\left(\widehat{z}_0 + z_{1-\gamma/2}\right)}\right),$$

where the bias-correction term z_0 may be estimated as

$$\widehat{z}_0 = \Phi^{-1}(f),$$

with f denoting the fraction of bootstrap estimates $\widehat{\theta}^*$ smaller than $\widehat{\theta}$ and $\Phi^{-1}(\cdot)$ denoting the inverse function of $\Phi(\cdot)$ such that, for example, $\Phi^{-1}(0.95) = 1.645$. The acceleration constant α corrects for the skewness of the distribution of the estimator and is typically estimated by the delete-one jackknife method as

$$\widehat{\alpha} = \frac{\sum_i \left(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}\right)^3}{6\left\{\sum_i \left(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}\right)^2\right\}^{3/2}},$$

where $\widehat{\theta}_{(i)}$ denotes the estimator obtained by deleting the i^{th} observation from the original sample and $\widehat{\theta}_{(\cdot)}$ is the average of $\widehat{\theta}_{(i)}$ over all elements i of the sample. If $\alpha = 0$, the BC_α-interval reduces to the BC-interval, and if both $\alpha = 0$ and $z_0 = 0$, the BC_α-interval reduces to the original percentile interval. For further discussion the reader is referred to Efron and Tibshirani (1993).

The delete-one jackknife method utilized by Efron to estimate α is based on the assumption of independent observations. This assumption is violated in autoregressive processes. One alternative may be to estimate $\widehat{\alpha}$ instead based on the block-jackknife method proposed in Künsch (1989) for the analysis of time-dependent stationary observations. Alternatively, the jacknife estimator $\widehat{\theta}_{(i)}$ may be constructed by deleting the i^{th} row of the matrix X obtained by stacking

$$x_t \equiv \left(y_t', 1, y_{t-1}', \ldots, y_{t-p}'\right), \quad t = 1, \ldots, T,$$

under the assumption of an autoregressive DGP with independent errors. None of these alternatives have been explored in the literature to date. As of now, the BC_α-interval does not appear to have been used for inference in autoregressive models.

Applied researchers sometimes utilize the original percentile interval ignoring the bias of the impulse response estimator altogether. This approach cannot be recommended in general. Kilian (1998c) demonstrates that the standard residual-based bootstrap based on the percentile interval, as implemented by

Runkle (1987), tends to suffer from severe coverage deficiencies in practice, when the data are highly persistent. Its coverage accuracy is much lower than that of the delta method interval, in fact. It is not uncommon for a nominal 95% interval to have effective coverage as low as 10%. A tell-tale sign of these bias problems in practice is that the impulse response estimate often lies outside of the 95% bootstrap confidence interval.

The reason is that the central premise of Runkle's bootstrap method, namely that the least-squares estimates of the VAR model parameters are not systematically different from their true values, is violated in typical applications. For expository purposes, consider the AR(1) model. Suppose that

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

where in population $a = 0.98$ and $T = 100$. The bootstrap approximation is based on

$$y_t^* = \widehat{\nu} + \widehat{a} y_{t-1}^* + u_t^*,$$

where $\mathbb{E}(\widehat{a} - a) < 0$ due to the small-sample bias of the LS estimator. In expectation, $a = 0.98$ is associated with an estimate of $\widehat{a} = 0.95$ (see Shaman and Stine 1988). This means that the bootstrap DGP is centered on the wrong location. More importantly, the bootstrap estimates $\widehat{a}^*$ conditional on this biased DGP are in turn downward biased relative to $\widehat{a}$, amplifying the overall bias in the bootstrap estimator. As a result, in expectation, the bootstrap estimate $\widehat{a}^*$ is only about 0.92.[5] Put differently, the bootstrap method, when applied inappropriately, pulls down the distribution of the estimator even further away from a, which causes the coverage rate of these bootstrap confidence intervals to become very low in practice. In contrast, the delta method, while also centered on the initial biased estimate of 0.95, imposes symmetry on the confidence intervals, preventing the second-round bias and resulting in higher, if still inadequate, coverage accuracy.

This insight goes a long way toward explaining the coverage deficiencies of the Runkle (1987) bootstrap intervals. In practice, there are two potential remedies for this problem in the stationary model. One approach is the bias-adjusted bootstrap method of Kilian (1998c) applied in conjunction with the traditional percentile interval. Another approach is to rely on alternative bootstrap confidence intervals that allow for bias and asymmetry in the distribution of the estimator without requiring explicit bias estimates. Both approaches are discussed later in this chapter.

Equal-Tailed Percentile-t Intervals. There are two types of percentile-*t* intervals. One was proposed by Efron (1982). This equal-tailed percentile-*t* interval

[5] Given that we are conditioning on 0.95 rather than 0.98 and given that the bias is increasing in the autoregressive parameter, the additional bias is slightly smaller than the bias in $\widehat{a}$.

is based on the same idea as the traditional asymptotic intervals, except that we create our own "bootstrap t"-table to replace the critical points based on the $\mathcal{N}(0, 1)$ table. We approximate the distribution of the asymptotically pivotal t-statistic

$$\frac{\widehat{\theta} - \theta}{\widehat{\sigma}(\widehat{\theta})}$$

by

$$\frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\sigma}(\widehat{\theta}^*)},$$

where the estimate $\widehat{\theta}$ is treated as a fixed parameter in the bootstrap DGP. Let $\widehat{t}^*_{\gamma/2}$ and $\widehat{t}^*_{1-\gamma/2}$ be the critical points defined by the $\gamma/2$ and $1-\gamma/2$ quantiles of the distribution of $\widehat{t}^* \equiv (\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}(\widehat{\theta}^*)$. Then

$$CI_{PER-t} = \left[\widehat{\theta} - \widehat{t}^*_{1-\gamma/2}\widehat{\sigma}(\widehat{\theta}),\ \widehat{\theta} - \widehat{t}^*_{\gamma/2}\widehat{\sigma}(\widehat{\theta})\right].$$

Superficially, this bootstrap interval looks similar to the usual asymptotic interval, but note that the bootstrap t-values allow for possible asymmetry in the distribution and implicitly correct for bias. This fact is important. It explains why the equal-tailed percentile-t interval tends to be much more accurate in practice without any explicit bias adjustments than the percentile interval without further adjustments.

How reliable this percentile-t interval is in finite samples also depends on the accuracy of $\widehat{\sigma}(\widehat{\theta})$. Here $\widehat{\sigma}(\widehat{\theta})$ is the estimated standard error of $\widehat{\theta}$ and may be computed analytically or by bootstrap simulation. In the latter case, one requires an additional nested bootstrap loop not only for the estimation of $\widehat{\sigma}(\widehat{\theta})$ but also for the estimation of each $\widehat{\sigma}(\widehat{\theta}^*)$. In practice, it is common to rely on closed-form solutions for the asymptotic variance, whenever possible. Even if a few hundred bootstrap iterations are enough for estimating the second moment of the distribution, this increases the computational cost by a factor of a few hundred.

The reliability of this interval also depends on how close to pivotal the finite-sample distribution of the t-statistic is. A statistic is pivotal if its distribution is the same, regardless of the values of the population parameters in the DGP. In our context, we know that the t-statistic is asymptotically $\mathcal{N}(0, 1)$ distributed under stationarity and hence pivotal in the limit, but this approximation need not work well in small samples and, in any case, can be expected to work well only for processes that are not very persistent. Nevertheless, in the AR(1) model the finite-sample approximation provided by the equal-tailed percentile-t interval remains reasonably accurate even fairly close to the unit root. Additional simulation evidence for VAR models shows, however, that the small-sample accuracy of the equal-tailed percentile-t interval can be erratic

in some regions of the parameter space (see, e.g., Kilian 1999b; Benkwitz, Lütkepohl, and Neumann 2000).

Symmetric Percentile-t Intervals. The other type of percentile-t interval is the symmetric percentile-t interval proposed by Hall (1992). Let $\widehat{t}^*_{1-\gamma}$ be the critical point defined by the $1-\gamma$ quantile of the distribution of $\widehat{t}^* \equiv |\widehat{\theta}^* - \widehat{\theta}| \,/\widehat{\sigma}(\widehat{\theta}^*)$. Then

$$CI^{\text{symmetric}}_{PER-t} = \left[\widehat{\theta} - \widehat{t}^*_{1-\gamma}\widehat{\sigma}(\widehat{\theta}),\ \widehat{\theta} + \widehat{t}^*_{1-\gamma}\widehat{\sigma}(\widehat{\theta})\right].$$

Note that this approach simply amounts to bootstrapping the two-sided t-test statistic as opposed to the one-sided t-statistic. The same caveats regarding implementation and computational cost apply that we already discussed for the equal-tailed percentile-t interval. In simulation studies, the symmetric percentile-t interval for autoregressions is often, but not always more accurate than the equal-tailed percentile-t interval (see Kilian 1999b).

Hall's Percentile Interval. Finally, Hall (1992) proposed bootstrapping yet another statistic obtained by replacing the denominator of the t-statistic underlying the equal-tailed bootstrap-t interval,

$$\frac{\widehat{\theta} - \theta}{\widehat{\sigma}(\widehat{\theta})},$$

by 1. Put differently, he proposed bootstrapping the statistic

$$\widehat{\theta} - \theta.$$

Let $\widehat{\theta}^*_{\gamma/2}$ and $\widehat{\theta}^*_{1-\gamma/2}$ be the critical points defined by the $\gamma/2$ and $1-\gamma/2$ quantiles of the distribution of $\widehat{\theta}^*$. Then Hall's percentile interval is

$$\begin{aligned} CI^{\text{Hall}}_{PER} &= \left[\widehat{\theta} - \left(\widehat{\theta}^*_{1-\gamma/2} - \widehat{\theta}\right),\ \widehat{\theta} - \left(\widehat{\theta}^*_{\gamma/2} - \widehat{\theta}\right)\right] \\ &= \left[2\widehat{\theta} - \widehat{\theta}^*_{1-\gamma/2},\ 2\widehat{\theta} - \widehat{\theta}^*_{\gamma/2}\right]. \end{aligned}$$

Hall (1992) somewhat misleadingly refers to the resulting bootstrap interval as the percentile interval, and to the original percentile interval proposed by Efron many years earlier as the "other" percentile interval.

Unlike Efron's percentile interval, Hall's percentile interval implicitly builds in an automatic bias correction, in that it has by construction a long left tail if the underlying distribution has a long right tail. This feature is not enough to ensure accurate inference in autoregressions, however. Although Hall's percentile interval in some cases is quite accurate, simulation evidence suggests that its small-sample coverage accuracy may be quite poor in other cases, especially compared with percentile-t intervals. Nor is it systematically more accurate than Efron's percentile interval in its original form. For further discussion the reader is referred to the comparison in Kilian (1999b).

Another interesting feature of Hall's percentile interval is that it has the correct size asymptotically when the statistic of interest is consistent and asymptotically normal, as assumed here, but that it may also be asymptotically valid in non-Gaussian settings, in which the Efron percentile interval is known to be invalid.

12.3 Bootstrap Intervals Based on Bias-Adjusted Estimators

Kilian (1998a, 1998b, 1998c, 1999b) proposed a bootstrap method based on bias-adjusted slope parameters designed to address the concerns about the lack of small-sample accuracy of the recursive bootstrap method in Runkle (1987). Like Runkle, Kilian relies on the percentile interval.[6]

The proposal is to approximate the unknown stationary VAR(p) data generating process

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

where $u_t \overset{iid}{\sim} F$, conditional on the observed data, not by the bootstrap DGP

$$y_t^* = \widehat{\nu} + \widehat{A}_1 y_{t-1}^* + \cdots + \widehat{A}_p y_{t-p}^* + u_t^*$$

with $u_t^* \overset{iid}{\sim} \widehat{F}_T$, but rather by the bootstrap DGP

$$y_t^* = \widehat{\nu} + \widehat{A}_1^{BC} y_{t-1}^* + \cdots + \widehat{A}_p^{BC} y_{t-p}^* + u_t^*,$$

where $[\widehat{A}_1, \ldots, \widehat{A}_p]$ has been replaced by first-order mean-bias-corrected LS estimates $[\widehat{A}_1^{BC}, \ldots, \widehat{A}_p^{BC}]$ already discussed in Chapter 2 and where $u_t^* \overset{iid}{\sim} \widehat{F}_T$. Let $\boldsymbol{\alpha} = \text{vec}([A_1, \ldots, A_p])$. Then, under some regularity conditions, $\mathbb{E}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = b(\widehat{\boldsymbol{\alpha}})/T + O(T^{-3/2})$, so $\widehat{\boldsymbol{\alpha}}^{BC} = \widehat{\boldsymbol{\alpha}} - b(\widehat{\boldsymbol{\alpha}})/T$, where $b(\widehat{\boldsymbol{\alpha}})/T$ may be estimated based on the closed-form solutions discussed in Chapter 2 or by bootstrap methods. The motivation for this bias correction is that we want the bootstrap DGP to be centered on parameter values that are not systematically different from their population values.

Sometimes $\widehat{\boldsymbol{\alpha}}$ is already close to the unit circle, and adjusting for small-sample bias would push $\widehat{\boldsymbol{\alpha}}^{BC}$ outside the stationary region. Such an adjustment

[6] This method of bias correcting the bootstrap is not related to the *BC* interval of Efron. In the context of bootstrapping structural impulse responses, the *BC* interval would adjust the endpoints of the interval for the conventional bootstrap impulse response estimator proposed by Runkle to account for the median bias in the impulse response estimator. This procedure does not involve bias-adjusting either the estimator of the impulse response or that of any of the autoregressive slope parameters. In contrast, Kilian adjusts for mean bias in the original as well as in the bootstrap slope parameter estimates before computing the bootstrap impulse responses, while evaluating the bootstrap distribution using Efron's original percentile interval.

would be inconsistent with the premise of a stationary DGP. To guard against this possibility, Kilian proposes a stationarity correction of the bias estimate (see Chapter 2). One first computes the modulus of the largest root of the companion matrix $\widehat{\mathbf{A}}$. Denote this quantity by $m(\widehat{\boldsymbol{\alpha}})$. If $m(\widehat{\boldsymbol{\alpha}}) \geq 1$, set $\widehat{\boldsymbol{\alpha}}^{BC} = \widehat{\boldsymbol{\alpha}}$. If $m(\widehat{\boldsymbol{\alpha}}) < 1$, however, compute $\mathbf{b} = b(\widehat{\boldsymbol{\alpha}})/T$ and $\widehat{\boldsymbol{\alpha}}^{BC} = \widehat{\boldsymbol{\alpha}} - \mathbf{b}$. If the implied $m(\widehat{\boldsymbol{\alpha}}^{BC}) \geq 1$, iterate on $\widehat{\boldsymbol{\alpha}}_i^{BC} = \widehat{\boldsymbol{\alpha}}^{BC} - \mathbf{b}_i$, $i = 1, 2, \ldots$, until $m(\widehat{\boldsymbol{\alpha}}_i^{BC}) < 1$, where $\mathbf{b}_i$ is shrunk by 1% for each i. This stationarity correction has no effect asymptotically.

The most convenient way of estimating the first-order mean bias in the VAR model is to rely on the closed-form solutions provided by Pope (1990). An alternative would be to estimate the bias by a nested bootstrap loop as in Kilian (1998c). If the regression model includes additional deterministic regressors such as a linear time trend, existing closed-form solutions for the LS bias in VAR models do not apply and we have to rely on the bootstrap. Especially when using closed-form solutions, it is possible to iterate the first-order bias estimates for improved accuracy. As discussed in Kilian (1998c), the magnitude of these higher-order bias terms is typically too small to matter in practice.

It may seem that we should also bias correct $\widehat{\boldsymbol{\sigma}}$, because the structural impulse response estimator $\widehat{\theta}_{ik,h}$ depends on both $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\sigma}}$. Such a modification would only have second-order effects, however, and would require us to recompute and recenter the residuals as well to preserve the internal consistency of the bootstrap algorithm. Given these complications, the small order of the effects, and the favorable performance of the algorithm even without this correction, it makes sense to retain the original $\widehat{\boldsymbol{\sigma}}$.

A subtle but important point to keep in mind is that bias corrections also affect the mean of the process because $\mathbb{E}(y_t) = (I_K - A_1 - \cdots - A_p)^{-1}\nu$. This problem does not arise when $\nu = 0$. In practice, however, ν may be far from zero, in which case the bootstrap replicates of the data will have a different mean after the bias correction, which affects the stationary distribution and causes transitory dynamics when generating bootstrap data in the conventional manner. These transitory dynamics undermine the accuracy of the bootstrap approximation, unless we discard the transients. The best way to avoid this problem in practice is to demean the data prior to estimating the model.

Having generated a sequence of data $\{y_t^{*r}\}_{t=-p+1}^{T}$, one fits the VAR($p$) model with intercept to obtain the LS estimates $[\widehat{\nu}^{*r}, \widehat{A}_1^{*r}, \ldots, \widehat{A}_p^{*r}]$ and $\widehat{\Sigma}_u^{*r}$, one implements the bias corrections on $\widehat{\boldsymbol{\alpha}}^*$ exactly as discussed earlier, and one constructs the implied bootstrap estimates $\widehat{\theta}_{ik,h}^{*r} = g(\widehat{\boldsymbol{\alpha}}^{BC*r}, \widehat{\boldsymbol{\sigma}}^{*r})$ for $r = 1, \ldots, R$. From the bootstrap approximation to the unknown empirical distribution of the structural impulse response estimates one may construct the $100(1-\gamma)$ percentile confidence interval

$$\left[\widehat{\theta}_{\gamma/2}^{*},\ \widehat{\theta}_{1-\gamma/2}^{*}\right],$$

where $\widehat{\theta}^*_{\gamma/2}$ and $\widehat{\theta}^*_{1-\gamma/2}$ refer to the lower and upper interval endpoints of the distribution of $\widehat{\theta}^{*r}_{ik,h}$.

The asymptotic validity of the bias-adjusted bootstrap follows from standard arguments in the literature by observing that the first-order bias is of order T and does not affect the limiting distribution of $\sqrt{T}\,(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ and $\sqrt{T}\,(\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})$. Hence, none of the proposed adjustments affect the first-order asymptotic validity of the recursive bootstrap method (see Kilian 1998c).

Originally, Kilian (1998c) proposed a computational shortcut to the algorithm described above, which is referred to as the bootstrap-after-bootstrap method. The proposal was to use the first-order bootstrap bias estimate $b(\widehat{\boldsymbol{\alpha}})/T$ as a proxy for the bias of the bootstrap estimates $b(\widehat{\boldsymbol{\alpha}}^{*r})/T$ for $r = 1, \ldots, R$. Kilian (1998c) proves that $b(\widehat{\boldsymbol{\alpha}})/T$ and $b(\widehat{\boldsymbol{\alpha}}^{*r})/T$ differ only by $O_p(T^{-3/2})$. This difference vanishes faster than the bias estimate itself and hence can be ignored in practice. The bootstrap-after-bootstrap speeds up the implementation of the bias-adjusted bootstrap method by a factor roughly corresponding to the number of bootstrap replications used in estimating the bias by bootstrap. This computational shortcut was essential when Kilian (1998c) was written, but had become largely obsolete by the time the article appeared in print, as computers had become much more powerful in the intervening years. Indeed, all subsequent papers on this topic, including Kilian (1998a, 1998b, 1999b), Kilian and Chang (2000), and Inoue and Kilian (2002b), use the unabbreviated algorithm described above.

An important caveat is that the closed-form bias estimates of Pope (1990) do not allow for conditional or unconditional heteroskedasticity. In fact, there are no analytic results about the mean bias in that case, and the order of the bias is unknown. It is possible, of course, to estimate the bias using the wild bootstrap, for example. The theoretical rationale for applying the bias-adjusted bootstrap approach in this case has yet to be established, however.

12.4 Potential Pitfalls in Impulse Response Inference

As already mentioned, one of the crucial assumptions in applying the Gaussian approximation is that the variance of the structural impulse response estimator is not zero (see Benkwitz, Lütkepohl, and Neumann 2000). This caveat applies not only to the delta method but also to all bootstrap methods. The nature of this problem is best understood in the context of a univariate stationary AR(1) process

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

where $|a| < 1$ and u_t is iid white noise. In this model, the response to a unit shock at horizon h is

$$\phi_h = a^h.$$

Under general conditions, the limiting distribution of the least-squares estimator $\widehat{a}$ can be expressed as

$$\sqrt{T}\,(\widehat{a} - a) \xrightarrow{d} \mathcal{N}(0, 1 - a^2),$$

so, by the delta method,

$$\sqrt{T}\left(\widehat{a}^h - a^h\right) \xrightarrow{d} \mathcal{N}\left(0, h^2 a^{2(h-1)}(1 - a^2)\right)$$

provided $a \neq 0$. If $a = 0$, in contrast, such that all impulse responses are zero in population,

$$\sqrt{T}\left(\widehat{a}^h - a^h\right) = \sqrt{T}\widehat{a}^h \xrightarrow{d} 0$$

for $h > 1$. In other words, $(\widehat{a}^h - a^h)$ converges to zero at a rate more rapid than $T^{1/2}$ and the distribution of $\sqrt{T}\left(\widehat{a}^h - a^h\right)$ is degenerate.

Because the estimate of the limiting variance is nonzero almost everywhere in the parameter space, it may be tempting to use the quantity

$$\sqrt{T}\left(\widehat{a}^2 - a^2\right)/2\widehat{a}(1 - \widehat{a}^2)^{1/2}$$

for constructing a confidence interval for ϕ_2, for example. However, for $a = 0$, the t-ratio reduces to

$$\sqrt{T}\widehat{a}/2\sqrt{(1 - \widehat{a}^2)},$$

where plim $\sqrt{(1 - \widehat{a}^2)} = 1$ such that the t-ratio converges to $\mathcal{N}(0, 1/4)$ asymptotically, given that

$$\sqrt{T}\widehat{a} \xrightarrow{d} \mathcal{N}(0, 1)$$

for $a = 0$. Hence, a confidence interval based on the asymptotic normal distribution would lack coverage accuracy.

As a result, standard asymptotic inference which ignores this possible singularity in the asymptotic variance-covariance matrix of the impulse response estimator will be misleading. This problem generalizes to structural impulse response analysis in vector autoregressions. Analogous problems arise for related statistics such as structural forecast error variance decompositions (see Benkwitz, Lütkepohl, and Neumann 2000).

The bootstrap approach does not help solve the problem of degenerate asymptotic distributions.[7] In fact, in some cases bootstrap intervals may be even less accurate than the delta method in the presence of singular covariance structures. For example, the CI_{PER}^{Efron}-interval has zero coverage probability for

[7] A common view is that changes in the rate of convergence at the point of discontinuity necessarily cause the bootstrap to fail. Inoue and Kilian (2003) prove by counterexample that the bootstrap may be valid even in the presence of such discontinuities. Nevertheless, in the current setting, the bootstrap approximation fails.

even h, when $a = 0$, because in that case the bootstrap impulse response estimator $(\widehat{a}^*)^h$ will be strictly positive with probability 1. For odd h, this problem does not arise. To illustrate this point, consider the example of $h = 2$. Given the asymptotic standard normal distribution of $\sqrt{T}\widehat{a}$, we know that

$$T\widehat{a}^2 \xrightarrow{d} \chi^2(1),$$

which has positive support. Consequently, the lower confidence band of the CI_{PER}^{Efron}-interval for $(\widehat{a}^*)^2$ cannot include $a^2 = 0$, resulting in a coverage rate of zero in repeated sampling. This result is unaffected by modifications of the original percentile interval such as small-sample bias adjustments.

Although other bootstrap confidence intervals such as CI_{PER}^{Hall}, CI_{PER-t}, or $CI_{PER-t}^{\text{symmetric}}$ will have higher coverage than the CI_{PER}^{Efron}-interval for even h, they also lack asymptotic justification. Moreover, some of the intervals that perform well for $a = 0$ in finite samples may perform poorly in asymptotically unproblematic parameter regions, making it difficult to offer advice to applied researchers.

Benkwitz, Lütkepohl, and Neumann (2000) explore a number of alternative approaches that remain asymptotically valid in this setting. For example, one possibility is to circumvent the problem by postulating a VAR(∞) process and allowing the lag order of the approximating VAR model to grow with the sample size. Another possibility is the use of subsampling methods. Benkwitz et al. conclude that either the small-sample accuracy of these procedures is not impressive or that it is not obvious how they can be extended to higher-dimensional processes. Yet another possibility would be to eliminate all points of singularities by fitting suitably restricted VAR models. That approach may not be practical, however, unless it is already known which impulse responses are zero. The use of pretests for identifying subset VAR structures also would raise concerns about the validity of asymptotic inference.

How practically important the example in Benkwitz, Lütkepohl, and Neumann (2000) is for structural VAR analysis depends on one's perspective. Benkwitz et al.'s argument is that there are situations where zero impulse responses are of particular interest in applied work. It may seem that zero impulse responses would arise easily in higher-dimensional stationary VAR models, but this is not the case, given that there are no exogenous observables in general equilibrium models of the economy and only exogenous variables would exhibit responses that are literally zero. There are some situations in which economic theory implies a zero response, however, as in Eichenbaum and Evans' (2005) analysis of uncovered interest parity. In the latter situation, Benkwitz et al.'s (2000) point that there may be problems of inference in VAR models due to singularities in the asymptotic distribution clearly applies.

However, there are many situations in which zero impulse responses are implausible. For example, no one would expect the response of real GDP to its

own innovation to be zero at all horizons, given the high persistence of the real GDP data. Thus, the problem identified by Benkwitz et al. does not necessarily invalidate standard methods of inference.

12.5 Finite-Sample Properties of Bootstrap Confidence Intervals

Abstracting from the problem of singular covariance structures, it can be shown that the symmetric percentile-t interval has theoretical properties similar to the infeasible BC_α and ABC intervals. Hall (1992) proves, albeit in a much simpler context, that among all two-sided confidence intervals apart from the BC_α and ABC intervals only the symmetric percentile-t interval achieves improved coverage accuracy asymptotically. All other two-sided confidence intervals are no more accurate asymptotically than the delta method interval. In particular, bootstrapping a studentized estimator alone is not sufficient for improved asymptotic coverage accuracy.

Hall's theoretical results, however, assume that an Edgeworth expansion of the distribution function of the estimator of interest exists, which has yet to be established in many applications. They also are based on the premise that the sample is large enough to invoke this asymptotic approximation. A simulation study in Kilian (1999b) suggests that even in a stylized bivariate VAR(1) model it may take sample sizes as large as $T = 2{,}000$ for the first-order asymptotics for $\widehat{\theta}_{ik,h}$ to provide a good approximation. In contrast, actual sample sizes in empirical macroeconomics tend to be between 160 and 720, not to mention that actual VAR models tend to have more lags and variables. Hence, it is not clear how relevant Hall's results about asymptotic refinements are for practitioners.

A small-scale simulation study based on AR(1) and VAR(p) models in Kilian (1999b) provides no evidence that the symmetric percentile-t interval is systematically more accurate than the equal-tailed percentile-t interval or than the percentile interval based on the bias-adjusted bootstrap method of Kilian (1998b). In fact, there are cases when it is less accurate than either alternative in finite samples. It is also shown that Hall's percentile interval tends to perform poorly in many cases, as does Efron's percentile interval when applied without explicit bias adjustment, confirming the earlier simulation results in Kilian (1998c). Related results for the AR(1) model can also be found in Benkwitz, Lütkepohl, and Neumann (2000).

In related work, Kilian (1998c, 1999b) demonstrates that, when applying the standard residual-based bootstrap method for VAR models with iid errors, small-sample bias adjustments greatly improve the coverage accuracy of Efron's percentile interval for structural impulse responses. Among the bootstrap intervals considered so far, it tends to be the most robust and most reliable approach, as long as the conditions for applying the delta method are satisfied. In models with only an intercept, the effective coverage tends to be reasonably close to the nominal coverage even for borderline stationary processes with

roots close to unity. In models with an added deterministic linear time trend, the improvement in coverage accuracy compared with Runkle's (1987) approach remains striking, but the actual coverage accuracy may be significantly lower than the nominal coverage accuracy. None of the methods discussed so far can be considered satisfactory in the latter class of models. This is a small-sample problem, not a problem with their asymptotic validity.

Coverage accuracy is the most important criterion for selecting between confidence intervals, but there are also other differences to keep in mind. For example, percentile-t intervals are not transformation respecting. This means that we cannot infer the interval endpoints for a statistic $g(\widehat{\theta})$, where $g(\cdot)$ denotes a nonlinear transformation, from the endpoints of the interval constructed for $\widehat{\theta}$. Moreover, like the delta method confidence interval, the percentile-t interval is not range-respecting. For example, when estimating a proportion, the interval may include values outside of the interval [0, 1]. On the other hand, Efron's percentile interval, while range-respecting, by construction does not include values of 0 or 1, which are on the boundary of the parameter space of the proportions estimator. They are valid only for inference about interior solutions. This point also applies to the problem of bootstrapping structural forecast error variance decompositions, for example.

12.6 Inference for Integrated and Cointegrated VAR Processes

If the variables in the VAR model are known to be integrated and/or cointegrated, it is straightforward to adapt the methods of inference discussed so far. As before, we assume that the structural model is exactly identified.

12.6.1 VAR Models in Differences

In the simplest example of variables that are integrated of order 1, one simply expresses these variables in first differences. If all variables are in first differences, the lag order of the VAR model to be bootstrapped is reduced by 1, resulting in the VAR($p-1$) model

$$\Delta y_t = \nu + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t.$$

In this case, inference would be based on the cumulative impulse responses of Δy_t, which correspond to the responses of the level of y_t. Lütkepohl (1990) provides closed-form solutions of the asymptotic variance-covariance matrices of the cumulative structural impulse responses that facilitate the application of the delta method.

Bootstrap inference in this model is almost identical to the methods outlined earlier, except that we bootstrap the cumulative sum of the impulse responses

of Δy_t. We approximate the DGP by

$$\Delta y_t^* = \widehat{\nu} + \widehat{\mathbf{\Gamma}}_1 \Delta y_{t-1}^* + \cdots + \widehat{\mathbf{\Gamma}}_{p-1} \Delta y_{t-p+1}^* + u_t^*$$

conditional on the data. Bias adjustments of the type discussed in Kilian (1998c, 1999b) tend to be of lesser importance after imposing the unit roots in estimation.

12.6.2 Vector Error Correction Models

As shown in Chapter 3, there are three different representations of VECMs used in empirical work.

VECMs with Known or Consistently Estimated Cointegrating Vectors. If the cointegrating vector is known, we can rewrite the VAR model as a VECM:

$$\Delta y_t = \alpha z_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t,$$

where $z_{t-1} = \beta' y_{t-1}$, $\mathbf{\Gamma}_j = -\left[A_{j+1} + A_{j+2} + \cdots + A_p\right]$ for $j = 1, 2, \ldots, p-1$, $\alpha\beta' = -\left(I_K - A_1 - A_2 - \cdots - A_p\right)$, and where we have assumed that there are no deterministic terms in the DGP for expository purposes. This VECM can be estimated by multivariate LS using a suitably specified regressor matrix. Because $z_t \sim I(0)$ and $\Delta y_{t-i} \sim I(0)$, $i = 1, \ldots, p-1$, all regressors can be written as stationary regressors and the slope estimator is $\sqrt{T}$-consistent and jointly asymptotically normal.

Alternatively, if the cointegrating vector is unknown, we can estimate the VECM

$$\Delta y_t = \alpha \widehat{z}_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t,$$

where $\widehat{z}_{t-1} = \widehat{\beta}' y_{t-1}$ may be obtained prior to estimating the VECM by one of the estimators described in Chapter 3. Because the estimator $\widehat{\beta}$ is superconsistent in that it converges at rate T rather than $\sqrt{T}$, we can condition on $\widehat{z}_{t-1}$ in estimating the VECM as though it were known. Whether β is known or estimated makes no difference for the asymptotic distribution of the implied structural impulse responses.

For asymptotic inference on the structural impulse responses in the VECM, we first need to express the VECM as a VAR model in levels:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t,$$

where $A_1 = \alpha\beta' + I_K + \mathbf{\Gamma}_1$, $A_i = \mathbf{\Gamma}_i - \mathbf{\Gamma}_{i-1}$, $i = 2, \ldots, p-1$, and $A_p = \mathbf{\Gamma}_{p-1}$. Let

$$A = \left[A_1, \ldots, A_p\right] = [\mathbf{\Pi}, \mathbf{\Gamma}]\, W + J, \tag{12.6.1}$$

where

$$\mathbf{\Pi} = -\left(I_K - A_1 - \cdots - A_p\right),$$

$$\underset{K\times Kp}{J} = \left[I_K,\ 0_{K\times K(p-1)}\right],$$

and

$$\underset{Kp\times Kp}{W} = \begin{bmatrix} I_K & 0 & 0 & \dots & 0 & 0 \\ I_K & -I_K & 0 & \dots & 0 & 0 \\ 0 & I_K & -I_K & & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & I_K & -I_K \end{bmatrix}.$$

Lütkepohl (2005, section 7.2) shows that $\widehat{\mathbf{\Gamma}} = [\widehat{\mathbf{\Gamma}}_1, \dots, \widehat{\mathbf{\Gamma}}_{p-1}]$ and $\widehat{\mathbf{\Pi}}$ in the VECM have a joint asymptotic normal distribution, as has the least-squares estimator $\widehat{A}$ of A:

$$\sqrt{T}\text{vec}(\widehat{A} - A) \xrightarrow{d} \mathcal{N}(0, \Sigma_\alpha^{co}).$$

The variance-covariance matrix Σ_α^{co} may be consistently estimated by

$$\widehat{\Sigma}_\alpha^{co} = (ZZ')^{-1} \otimes \left[\left(Y - \widehat{A}Z\right)\left(Y - \widehat{A}Z\right)'\right],$$

where $Y = [y_1, \dots, y_T]$ and $Z \equiv [Y_0, \dots, Y_{T-1}]$ with $Y_{t-1} \equiv \left[y'_{t-1}, \dots, y'_{t-p}\right]'$. Moreover, in the Gaussian model,

$$\sqrt{T}\text{vech}\left(\widehat{\Sigma}_u - \Sigma_u\right) \to \mathcal{N}\left(0, 2\mathbf{D}_K^+(\Sigma_u \otimes \Sigma_u)\mathbf{D}_K^{+\prime}\right).$$

Asymptotically, $\widehat{\Sigma}_u$ is independent of $\widehat{A}$. Given these estimates, inference about the structural impulse responses based on the delta method can be conducted analogously to the stationary Gaussian VAR model.

Bootstrap inference for known β involves approximating the VECM DGP by

$$\Delta y_t^* = \widehat{\alpha}\beta' y_{t-1}^* + \widehat{\mathbf{\Gamma}}_1 \Delta y_{t-1}^* + \cdots + \widehat{\mathbf{\Gamma}}_{p-1}\Delta y_{t-p+1}^* + u_t^*$$

conditional on the data. For each bootstrap replication of the VECM estimates, the bootstrap estimates $\widehat{\mathbf{\Gamma}}^*$and $\widehat{\mathbf{\Pi}}^*$ are converted to bootstrap estimates $\widehat{A}^*$ using (12.6.1), allowing us to build up the bootstrap approximation to the empirical distribution of the structural impulse responses exactly as in the stationary case. In the absence of an intercept, care must be taken to recenter the residuals.

If the VECM relies on a first-stage estimate of the cointegrating vector β, one option is to first cumulate the Δy_t^* realizations implied by the bootstrap

DGP

$$\Delta y_t^* = \widehat{\alpha}\widehat{\beta}' y_{t-1}^* + \widehat{\mathbf{\Gamma}}_1 \Delta y_{t-1}^* + \cdots + \widehat{\mathbf{\Gamma}}_{p-1} \Delta y_{t-p+1}^* + u_t^*$$

to obtain $\{y_t^*\}_2^T$ and then estimate $\widehat{\beta}^*$ from these bootstrap data before reestimating the VECM conditional on $\widehat{\beta}^*$. The other option is to condition on the initial $\widehat{\beta}$ throughout as though β were known.

VECMs with Known Cointegrating Rank. If we know only the number of cointegrating relationships in the data, the VECM may equivalently be written as

$$\Delta y_t = \mathbf{\Pi} y_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t.$$

As discussed in Chapter 3, this model may be estimated efficiently by Gaussian maximum likelihood estimation if we have a priori information about the number of cointegrating relationships (or equivalently about the rank of $\mathbf{\Pi}$). Johansen (1995) proves that the maximum likelihood estimators $\widetilde{\mathbf{\Gamma}}$, $\widetilde{\mathbf{\Pi}}$, and $\widetilde{\Sigma}_u$ in this VECM have the same joint asymptotic normal distribution as the Gaussian ML estimator in the restricted VECM we discussed earlier.[8]

Bootstrap inference may be based on the DGP,

$$\Delta y_t^* = \widetilde{\mathbf{\Pi}} y_{t-1}^* + \widetilde{\mathbf{\Gamma}}_1 \Delta y_{t-1}^* + \cdots + \widetilde{\mathbf{\Gamma}}_{p-1} \Delta y_{t-p+1}^* + u_t^*,$$

where the known rank of $\widetilde{\mathbf{\Pi}}^*$ is used in each bootstrap replication to construct the estimates $\widetilde{\mathbf{\Gamma}}^*$ and $\widetilde{\mathbf{\Pi}}^*$, which are converted to $\widetilde{A}^*$ in order to construct the bootstrap estimates of the structural impulse responses.

The VECM can also be used if the cointegrating rank is unknown. In that case, statistical tests are used to determine the cointegrating rank first, and the bootstrap is conditioned on the cointegrating rank obtained in this way (see Chapter 3). Unlike in the case of a known cointegrating rank, the asymptotic properties of this procedure are not known.

VECMs with Known Cointegrating Vector in Triangular Representation. For cointegrated models with only two variables and a known cointegrating vector $(1, -1)$, there is yet another approach, called the triangular VAR representation of the VECM. For expository purposes, consider a VECM for spot exchange rates (s_t) and economic fundamentals (f_t). The presumption is that in the long run, s_t and f_t move together such that $s_t - f_t \sim I(0)$. It can be shown that this

[8] Equivalently, one could have written the VECM as $\Delta y_t = \Pi y_{t-p} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + u_t$, where the first term involves y_{t-p} rather than y_{t-1}. This does not affect the asymptotic results except for some slight changes in notation. For details, see Lütkepohl and Reimers (1992a).

VECM is algebraically equivalent to a VAR model of the form:

$$\begin{pmatrix} \Delta s_t \\ s_t - f_t \end{pmatrix} = C_1 \begin{pmatrix} \Delta s_{t-1} \\ s_{t-p} - f_{t-p} \end{pmatrix} + \cdots + C_p \begin{pmatrix} \Delta s_{t-p} \\ s_{t-p} - f_{t-p} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

with some restrictions on the lagged coefficients that in practice tend to be ignored. The latter model directly incorporates the known cointegrating vector (see, e.g., Campbell and Shiller 1987; Kilian 1999a). Estimation is by unrestricted LS. This example illustrates that, in many cases, known cointegrating vectors can be embedded as ratio variables in regular VAR models. The advantage is that standard residual-based bootstrap or asymptotic methods of inference may be applied to this stationary model. The responses of s_t correspond to the cumulative response of Δs_t, and the response of f_t may be inferred from the difference of the responses of s_t and $s_t - f_t$. Thus, asymptotic inference about the response of f_t requires knowledge of the covariance between the responses. In practice, inference is usually based on bootstrap methods.

12.6.3 Integrated and/or Cointegrated VAR Models in Levels

Although imposing unit roots and cointegration restrictions improves the efficiency of impulse response estimators, in many cases consistent estimators with conventional asymptotic distributions may also be constructed from the unrestricted levels specification of the VAR model.

Gaussian Asymptotic Approximations. As already discussed in Chapters 2 and 3, any VAR model may be consistently estimated in levels, regardless of the possible presence of unit roots or cointegration, provided the model includes an intercept and enough lags such that the slope coefficients can be written as coefficients on zero mean stationary regressors (see Sims, Stock, and Watson 1990). There are two situations in which the use of this level specification may be preferred. One situation is that we are not sure about the presence of unit roots in the model variables. In this case, a suitably specified VAR model in levels with intercept encompasses both the integrated and possibly cointegrated VAR model and the stationary VAR model without trend. It does not allow for trending behavior in the stationary variables, however. Such trending behavior would call for the inclusion of an additional deterministic time trend in the regression in the event the model is stationary. The inclusion of a deterministic time trend does not affect the asymptotic results, but it magnifies the small-sample bias in the slope parameters of the VAR model and hence may seriously undermine the accuracy of the asymptotic approximation in small samples. Perhaps for this reason, in applied work, users of VAR models typically exclude the deterministic time trend even for trending variables such as real output. In the frequentist framework, this amounts to treating real output as a near-unit

root (and possibly near-cointegrated) process in which the intercept acts as a near-drift.

Another important situation is a VAR model in which there may or may not be cointegration between the unit root variables. Rather than relying on pretests for the cointegration rank with the pretests outcome being subject to error, or simply assuming the existence of a cointegration relationship, it makes sense in this case to estimate the VAR without these restrictions, even if we are sure that there are unit roots in some of the variables.

As discussed in Chapters 2 and 3, asymptotic inference in level VAR(p) models is complicated by the fact that the presence of integrated variables may render the asymptotic variance of the statistic of interest singular. Although the individual slope parameters retain their marginal Gaussian asymptotic distribution after $\sqrt{T}$-scaling provided $p > 1$, their joint variance-covariance matrix is singular and inference about certain linear combinations of slope parameters is nonstandard (see Chapter 3). This fact also has implications for inference about smooth nonlinear functions of VAR(p) slope parameters, $g(\boldsymbol{\alpha})$. Under certain conditions, it may render the asymptotic covariance matrix of the estimator singular, invalidating standard inference.

This problem is best illustrated in the simpler context of a univariate autoregression. Let the scalar time series y_t be generated by the autoregressive process

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

where $u_t \overset{iid}{\sim} F$, and $\rho \equiv a_1 + \cdots + a_p = 1$. Define $a = \left(a_1, \ldots, a_p\right)'$. Consider the first-order linear approximation to the response of y_{t+h} to a unit shock in u_t at horizon $h \geq p$, given by

$$\phi_h = g(a_1, \ldots, a_p) \approx \frac{\partial g}{\partial a'} \left(\widehat{a} - a\right).$$

Then it is immediately apparent that when this linear approximation is proportionate to the parameter ρ, i.e.,

$$\frac{\partial g}{\partial a_1} a_1 + \cdots + \frac{\partial g}{\partial a_p} a_p \propto a_1 + \cdots + a_p,$$

which happens when all elements of the derivative vector $\partial g / \partial a'$ are identical in population, the distribution of the impulse response estimator $\widehat{\phi}_h$ becomes nonstandard with the estimator converging at rate T (like the estimator of ρ in a Dickey-Fuller regression) rather than $\sqrt{T}$. In this case, the limiting variance of $\sqrt{T}(\widehat{\phi}_h - \phi_h)$ is zero and standard asymptotic and bootstrap inference fails even when $p > 1$. Analogous results can be derived for vector-valued unit root processes with and without cointegration.

How much of a concern this situation is in practice is not clear. One response, therefore, is to simply ignore this possibility, when constructing

pointwise confidence intervals for smooth functions $g(a)$. An alternative response is to adapt the approach of Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996), as discussed in Chapter 3. If y_t consists of $I(0)$ and $I(1)$ variables only, it suffices to add an extra lag to the VAR process fitted to the data to obtain a nonsingular covariance matrix for the parameters associated with the first p lags. In other words, if the DGP is a VAR(p) process and a lag-augmented VAR($p+1$) model,

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

is fitted by LS, then the estimator of $[A_1, \ldots, A_p]$ has a nonsingular variance-covariance matrix. Hence, inference on smooth functions of the slope parameters such as reduced-form impulse responses may be conducted by the delta method based on $[\widehat{A}_1, \ldots, \widehat{A}_p]$. More generally, if $I(d)$ variables are present the singularity problem of the covariance matrix can be resolved by augmenting the VAR model by d extra lags for estimation. This approach not only suffices to rule out singularities in the pointwise asymptotic variance of $g(a)$; it is, in fact, required if we are interested in conducting joint rather than pointwise inference on statistics that are smooth functions of the form $g(a)$.

Of course, most statistics of interest in structural VAR analysis involve smooth functions of the form $g(\boldsymbol{\alpha}, \boldsymbol{\sigma})$ rather than $g(\boldsymbol{\alpha})$. A case in point are structural impulse responses and structural forecast error variance decompositions. In this case, the concern regarding forbidden linear combinations typically does not arise. Recall that, under suitable conditions,

$$\sqrt{T}\left(g(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\sigma}}) - g(\boldsymbol{\alpha}, \boldsymbol{\sigma})\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g}{\partial \boldsymbol{\alpha}'} \Sigma_{\widehat{\boldsymbol{\alpha}}} \frac{\partial g'}{\partial \boldsymbol{\alpha}} + \frac{\partial g}{\partial \boldsymbol{\sigma}'} \Sigma_{\widehat{\boldsymbol{\sigma}}} \frac{\partial g'}{\partial \boldsymbol{\sigma}}\right).$$

Even if the leading term of the limiting variance,

$$\frac{\partial g}{\partial \boldsymbol{\alpha}'} \Sigma_{\widehat{\boldsymbol{\alpha}}} \frac{\partial g'}{\partial \boldsymbol{\alpha}},$$

were zero due to some singularity in $\Sigma_{\widehat{\boldsymbol{\alpha}}}$, the other term,

$$\frac{\partial g}{\partial \boldsymbol{\sigma}'} \Sigma_{\widehat{\boldsymbol{\sigma}}} \frac{\partial g'}{\partial \boldsymbol{\sigma}},$$

will always be nonsingular, ensuring that the limiting variance of the structural impulse response estimator is nonsingular, unless the value of the reduced-form impulse response is zero.

This result provides the basis for conducting asymptotically valid pointwise inference on structural impulse responses in VAR models using the delta method. Provided the structural forecast error variance components are bounded away from the limiting values of 0 or 1, this result also allows asymptotic inference on structural forecast error variance decompositions. Of course,

the limit of this decomposition as $h \to \infty$ is not well defined in the presence of a unit root because the Wald MA representation of y_t does not exist.

Only if we are interested in conducting joint inference for sets of structural impulse responses will the use of the lag augmentation approach be required. It should be noted that the lag augmentation approach involves a loss of efficiency in that we include a redundant lagged regressor in the VAR model. In practice, however, this effect is likely to be negligible, as the width of confidence intervals of structural impulse responses tends to be insensitive to small changes in the lag order, especially if a reasonably large number of lags is used in the first place.

Bootstrap Approximations. Results analogous to those in Sims, Stock, and Watson (1990) also apply to conventional residual-based bootstrap methods for VAR models in levels, as shown in Inoue and Kilian (2002a). For expository purposes, consider first the simpler univariate context. Suppose the DGP is a random walk,

$$y_t = ay_{t-1} + u_t,$$

where $a = 1$, and we do not impose the unit root in estimation. Then the standard bootstrap estimator of a based on the AR(1) model (with or without deterministic regressors) will not recover the correct asymptotic Dickey-Fuller (DF) distribution. Rather, it will converge to a random distribution. In other words, the standard bootstrap based on the unrestricted AR(1) regression model fails (see Basawa, Mallik, Cormick, Reeves, and Taylor 1991; Datta 1996).[9]

By augmenting this AR(1) model with additional lags, however, say by fitting an AR(2) model rather than an AR(1) model, we can overcome the inconsistency of the bootstrap, at least for statistics such as individual slope parameters. It is irrelevant for this argument that the DGP is an AR(1) model. What matters is the number of lags in the fitted model. This approach does not ensure high accuracy of bootstrap confidence intervals in small samples, but it ensures their first-order asymptotic validity.[10]

More generally, the results in Inoue and Kilian (2002a) apply to all higher-order AR(p) models, even when a deterministic time trend is included in the regression in addition to the intercept. The reason is that the bootstrap estimate $\widehat{a}^*$, where a is the sum of the autoregressive coefficients, converges at a rate

[9] Interestingly, the unrestricted bootstrap remains asymptotically valid when we fit an AR(1) with intercept to data from a random walk with drift, although there is a similar discontinuity in the asymptotic distribution at the unit circle (see Inoue and Kilian 2003).

[10] One practical concern is that residual-based bootstrap methods for structural impulse responses under these conditions may have poor coverage accuracy without small-sample bias adjustments of the type discussed in Kilian (1998c, 1999b). Although first-order bootstrap mean-bias adjustments in simulation studies tend to improve the small-sample accuracy of bootstrap intervals in the presence of unit roots, their theoretical validity so far has been established only for stationary VAR models (see Kilian 1999b).

faster than $\sqrt{T}$. The fact that $\widehat{a}^*$ converges to a random distribution is still true even in AR(p) models. This fact does not matter, however, because $\widehat{a}^*$ does not enter the limiting distribution of the level slope parameter estimates after $\sqrt{T}$-scaling. This means that we can conduct valid bootstrap inference on individual slope parameters and linear combinations of individual slope parameters except for combinations that are proportionate to a (see Inoue and Kilian 2002a).

This reasoning also generalizes to vector autoregressive processes of order $p > 1$, whether the data are $I(0)$, $I(1)$, or cointegrated, provided an intercept is included in estimation. More generally, residual-based bootstrap inference remains first-order asymptotically valid for smooth nonlinear functions $g(\boldsymbol{\alpha}, \boldsymbol{\sigma})$, as long as the limiting variance of the impulse response is not degenerate. As in the case of the delta method, singularities in the joint variance-covariance matrix of the slope parameters may be overcome by relying on the lag-augmented VAR model if necessary.

An important assumption underlying the derivation of the asymptotic validity of the bootstrap for structural impulse responses in the presence of unit roots is that the horizon of the impulse response remains fixed as the sample size increases. Without that assumption, which is also used by the delta method intervals of Lütkepohl (1990), bootstrap inference for structural impulse responses based on the VAR model in levels would be invalid (see Phillips 1998). In practice, this assumption is commonly interpreted as requiring the horizon to be small relative to the sample size. In this regard, Kilian and Chang (2000) document by simulation that the coverage accuracy of conventional asymptotic and bootstrap intervals for VAR impulse responses tends to deteriorate at longer horizons when the data are highly persistent. For typical large-dimensional VAR models estimated in levels, the usual pointwise confidence intervals often remain accurate only for a horizon of about one year. This finite-sample result is not surprising, because the data have roots close to unity and these intervals are based on asymptotic approximations derived under the assumption of a fixed horizon.

12.7 Inference in Local-to-Unity VAR Processes

Many macroeconomic time series are highly persistent. An obvious concern for applied work is that we do not know whether these time series are integrated or not. This is not a question that can be resolved empirically. Even the most powerful unit root tests lack the power to help us discriminate between these hypotheses for empirically relevant sample sizes.

So-called confirmatory analysis based on tests of the $I(0)$ null hypothesis does not remedy this problem and has been shown to be misleading (see, e.g., Caner and Kilian 2001; Müller 2005). Confirmatory analysis refers to the

practice of testing both the null hypothesis that a given process is $I(1)$ and that it is $I(0)$, and interpreting a failure to reject the $I(1)$ null and a simultaneous rejection of the $I(0)$ null as confirmation of the $I(1)$ hypothesis. This practice tends to be misleading, because tests of the $I(1)$ hypothesis lack power against persistent alternatives in small samples, typically resulting in a non-rejection, while a spurious rejection of the $I(0)$ hypothesis is made likely by the fact that the standard critical values of tests of the $I(0)$ hypothesis do not allow for persistent processes under the null hypothesis, causing these tests to suffer from severe size distortions in small samples, with rejection rates approaching 70% at the nominal 5% level in some realistic settings. Moreover, even if the data were known to be integrated of order 1, we would not know whether they are cointegrated, and if they were known to be cointegrated, we would need to know at least the cointegrating rank, if not also the cointegration vectors. Statistical tests for cointegration often lack power, and identifying the correct cointegrating vectors in higher-dimensional VAR systems remains a challenge in practice.

It may be tempting simply to impose unit roots and/or cointegration restrictions in estimation, but imposing these restrictions when they do not hold will render the VAR estimates inconsistent and inference invalid. In contrast, the failure to impose unit root and cointegration restrictions when they do hold will render the model estimates less efficient, but the estimates will remain consistent. Moreover, as we have discussed, impulse response inference in suitably specified VAR models estimated in levels remains asymptotically valid in most practically relevant situations even without imposing these restrictions, and lag augmentation may be used to overcome the remaining problems.

These asymptotic arguments suggest that in case of doubt we should err on the side of not imposing unit roots and cointegration on higher-order VAR models. In finite samples, however, that strategy is not without its own drawbacks because the high persistence of the levels data increases the small-sample bias of the VAR slope parameters and may undermine the accuracy of conventional asymptotic and bootstrap approximations even in higher-order VAR models.

12.7.1 Local-to-Unity Asymptotics

The questionable accuracy of conventional delta method and bootstrap intervals near the unit circle has motivated the development of an alternative asymptotic approximation for estimators based on VAR models in levels, referred to as local-to-unity asymptotics (see Chapter 3). This method is designed to be robust to the possible presence of unit roots in VAR(1) models as well as higher-order VAR models. By construction, it is intended for VAR models based on short-run identifying restrictions only, as departures from exact unit roots would immediately invalidate the use of long-run identifying restrictions.

Early applications of this modeling framework to the construction of impulse response confidence intervals for univariate autoregressions include Wright (2000) and Gospodinov (2004). Here we focus on the vector model. The standard procedure of generating confidence intervals for structural VAR impulse responses within the local-to-unity framework was proposed by Pesavento and Rossi (2006). Consider the K-dimensional reduced-form model

$$(I_K - CL)y_t = \Psi(L)u_t, \tag{12.7.1}$$

where $C = I_K + \Lambda/T$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_K)$, and $\Psi(L) = \sum_{i=0}^{\infty} \Psi_i L^i$ is an MA operator. This assumption means that the diagonal elements of C are in the neighborhood of 1, where that neighborhood depends on a constant that shrinks toward zero at rate T. That rate T is chosen to match the rate of convergence of the least-squares estimator of C. If λ_j were 0, y_{jt} would follow an exact unit root process. Allowing $\lambda_j < 0$ means that the root is near unity, but below unity (see Chapter 3).

Our interest is in inference on the structural impulse responses. Unlike in conventional asymptotics, we parameterize the horizon h of the response as an increasing function of the sample size. Specifically, let $h = [\delta T]$ for some fixed $\delta > 0$, where $[\cdot]$ denotes the integer part of the number in brackets. Then, given that

$$\lim_{h\to\infty} \left(1 + \frac{\delta\lambda_k}{h}\right)^h = e^{\delta\lambda_k},$$

we have

$$C^h \to e^{\delta\Lambda}$$

as $h \to \infty$, where $e^{\delta\Lambda}$ is a diagonal matrix with the vector $(e^{\delta\lambda_1}, e^{\delta\lambda_2}, \ldots, e^{\delta\lambda_K})$ on the main diagonal. The reduced-form impulse response at horizon h is:

$$\frac{\partial y_{t+h}}{\partial u_t'} = \Phi_h = C^h \left(I_K + C^{-1}\Psi_1 + C^{-1}\Psi_2 + \cdots\right) + o(1) \approx C^h\Psi(1), \tag{12.7.2}$$

and the corresponding structural response of variable j to shock k is:

$$\frac{\partial y_{j,t+h}}{\partial w_{k,t}} \approx e_j' C^h \Psi(1) B_0^{-1} e_k \to e_j' e^{\delta\Lambda} \Psi(1) B_0^{-1} e_k \tag{12.7.3}$$

as $T \to \infty$, where e_j and e_k are columns j and k of the matrix I_K, and B_0^{-1} is the structural impact multiplier matrix. Note that this limiting response depends on the largest roots in Λ as well as the cumulative short-run dynamics $\Psi(1)$. At long horizons, the uncertainty associated with $\Psi(1)$ and B_0^{-1} will be negligible compared with the uncertainty about Λ, so we can simply replace these terms

by consistent estimates $\widehat{\Psi}(1)$ and $\widehat{B}_0^{-1}$. Although Λ cannot be estimated consistently, the confidence limits for the localizing constants λ_j can be obtained by inverting a unit root test such as the augmented Dickey-Fuller test (see, e.g., Stock 1991).[11] Let $(\lambda_{L,j}, \lambda_{U,j})$ denote the lower and upper bounds of this confidence interval for λ_j, $j = 1, \ldots, K$. Because the components of the limiting response are monotone functions of λ_j, the implied confidence interval for the structural impulse response in (12.7.3) is

$$\left(e^{\delta\lambda_{L,j}} e_j' \widehat{\Psi}(1)\widehat{B}_0^{-1} e_k, \quad e^{\delta\lambda_{U,j}} e_j' \widehat{\Psi}(1)\widehat{B}_0^{-1} e_k\right).$$

The difference from the pointwise impulse response confidence intervals discussed earlier is that coverage is not pointwise, but joint for all long horizons. Simulations suggest that at horizons shorter than about 10% of the sample size, the asymptotic approximation breaks down. One caveat regarding this interval is that the assumption of a diagonal C, while sometimes reasonable in macroeconomic applications, may not hold in all empirical applications. Relaxing this assumption would considerably complicate inference. The other obvious drawback of this procedure is that it provides accurate approximations only for responses at long horizons.

The procedure in Pesavento and Rossi (2006) is a two-step procedure in that we first construct a confidence interval for λ_j and then form the confidence interval for the structural impulse response based on (12.7.3). In related work, Gospodinov (2004) proposes an alternative, more computationally intensive approach that involves only one step. His proposal is to construct the interval based on a one-step inversion of the LR test statistic for the impulse response. Here we consider a generalization of this approach to structural impulse responses discussed in Gospodinov (2003). As in Chapter 3, let

$$y_t = Cy_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t,$$

where $C = I_K + \Lambda/T$, and define $\chi \equiv (\text{vec}(C)', \text{vec}(\mathbf{\Gamma}_1)', \text{vec}(\mathbf{\Gamma}_2)', \ldots, \text{vech}(\Sigma_u)')'$. Consider the null hypothesis $\mathbb{H}_0 : \Theta_h(\chi) = \Theta_{0,h}$ and let

$$LR_T = T \log\left[\det(\widetilde{\Sigma}_u^{-1}) \det(\widetilde{\Sigma}_u^r)\right]$$

denote the likelihood ratio statistic, where $\widetilde{\Sigma}_u^r$ and $\widetilde{\Sigma}_u$ denote the variance-covariance matrices of the restricted and unrestricted residuals, respectively, in

[11] Phillips (2014) shows that the inversion of the usual ADF statistic leads to invalid inference when the DGP is stationary. This observation suggests that caution is called for when constructing the first-stage confidence interval for the local-to-unity parameter. Phillips (2014) discusses possible solutions based on the switching approach of Elliott, Müller, and Watson (2015) and based on the approaches of Hansen (1999) and Mikusheva (2007).

the approximating VAR model. Then, given $h = [\delta T]$, under $\mathbb{H}_0$

$$LR_T \xrightarrow{d} \text{tr}\left[\left(\int_0^1 \mathbf{J}_c(s)d\mathbf{W}(s)'\right)'\left(\int_0^1 \mathbf{J}_c(s)\mathbf{J}_c(s)'ds\right)^{-1}\left(\int_0^1 \mathbf{J}_c(s)d\mathbf{W}(s)'\right)\right],$$

where $\mathbf{J}_c(s)$ is an Ornstein-Uhlenbeck vector process and $\mathbf{W}$ is a vector of standard Brownian motions. Joint confidence sets for horizon h are obtained by inverting the LR_T test on a grid of possible values of $\Theta_{0,h}$. Pointwise confidence intervals at horizon h are obtained using the projection method (see, e.g., Dufour and Taamouti 2005; Chaudhuri and Zivot 2011).

There is no simulation evidence on the accuracy of Gospodinov's approach for VAR models, but evidence for univariate models suggests consistently high coverage at all but the shortest horizons. Compared with the two-step procedure of Pesavento and Rossi (2006), the inversion of the LR test statistic is more computationally intensive, which is why this method does not appear to have been used for VAR models in practice.

12.7.2 Inference in Levels for Local-to-Unity VAR Models

It can be shown that the results of Sims, Stock, and Watson (1990) and Inoue and Kilian (2002a) on the asymptotic validity of inference in higher-order vector autoregressions with exact unit roots also extend to data generating processes with roots local-to-unity. As long as the delta method can be applied, inference about structural impulse responses remains standard. Likewise, lag augmentation may be used to render asymptotic inference invariant to roots local to unity (see Bauer and Maynard 2012).

One caveat is that these results are conditional on the horizon being fixed with respect to the sample size. In practice, one would expect conventional bootstrap or delta method intervals to be accurate only at reasonably short horizons relative to the sample size. In this sense, the conventional approach is complementary to the nonstandard asymptotic analysis in Pesavento and Rossi (2006) for longer horizons, and one might consider ways of combining these asymptotic thought experiments. We will return to this point later in this chapter.

A second and less obvious caveat is that the robustness results in Inoue and Kilian (2002a) do not extend to all VAR models based on long-run identifying restrictions. Consider the bivariate model of Blanchard and Quah (1989), for example. Their identifying assumptions require that real GDP be treated as $I(1)$ and hence expressed in differences, whereas the unemployment rate must be $I(0)$ and hence expressed in levels. In practice, an obvious concern is that the unemployment rate itself is quite persistent. Gospodinov (2010) proves that

if the unemployment rate is parameterized as a local-to-unity process rather than an $I(0)$ process, reflecting our uncertainty about its true integration status, the structural impulse responses can no longer be consistently estimated, even assuming that the horizon is fixed. Likewise, more conventional asymptotic and bootstrap confidence intervals for the structural impulse responses are invalid. This result suggests caution in interpreting estimates of Blanchard-Quah style models, unless the second variable is well inside the stationary region.

12.7.3 The Grid Bootstrap Method

So far we have considered the local-to-unity framework from an asymptotic point of view. We now turn to bootstrap methods for local-to-unity processes. While conventional bootstrap methods remain valid asymptotically for approximating the distribution of individual slope parameters in AR(p) models as long as $p > 1$, they are unable to capture the distribution of the sum of the autoregressive slope coefficients (see Inoue and Kilian 2002a). Indeed, that problem is formally equivalent to the problem of bootstrapping the slope parameter of the AR(1) model with a unit root. As mentioned earlier, Basawa, Mallik, Cormick, Reeves, and Taylor (1991) proved that the conventional bootstrap for autoregressions fails in that situation.[12] While this result also extends to the local-to-unity AR(1) model, such proofs do not rule out the possibility that nonstandard bootstrap methods remain valid in this case.

Hansen (1999) proposed such a nonstandard bootstrap approach, called the grid bootstrap, for univariate time series. His objective was to construct confidence intervals for the sum of the autoregressive roots in the AR(p) local-to-unity model. He showed that confidence intervals constructed with this method achieve asymptotically first-order correct coverage for both stationary and local-to-unity autoregressive models.

The sum of the autoregressive coefficients corresponds to the parameter ρ in the augmented Dickey-Fuller (ADF) representation of y_t,

$$y_t = \rho y_{t-1} + \zeta_1 \Delta y_{t-1} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + u_t.$$

The grid bootstrap exploits the fact that we can estimate the remaining parameters of this ADF model conditional on the null hypothesis about ρ. The critical values for the t-statistic for ρ depend on these restricted parameter estimates and will change with the specification of the null hypothesis. Because there are many possible null hypotheses, Hansen considers a grid of null values for the statistic of interest and composes a confidence set for ρ from the ρ values that are not rejected.

[12] For further discussion of the causes of that bootstrap failure, see also Inoue and Kilian (2003).

In Section 12.4, we already alluded to the fact that pointwise asymptotic approximations do not guarantee confidence intervals that have correct coverage uniformly across the parameter space. Such problems often arise when the convergence rate of the estimators to their asymptotic distribution changes across the parameter space. This problem has also been studied extensively in the context of processes with roots near unity. Mikusheva (2007) proves that the Hansen (1999) grid bootstrap procedure produces uniformly asymptotically correct confidence intervals for $|\rho| \leq 1$. Simulation evidence in Hansen (1999) and Kilian (1999b) confirms that this grid bootstrap is quite accurate in finite samples for a wide range of degrees of persistence including the unit root limit. It remains highly accurate even for values of ρ close to zero.

The original grid bootstrap is not designed for structural impulse response inference in VAR models, however. A substantial generalization of the grid bootstrap is provided in Mikusheva (2012). Mikusheva observes that under the assumption that the impulse response horizon is fixed, one indeed obtains impulse response estimates that typically converge to a normal distribution, allowing the use of the standard delta method and of conventional bootstrap methods as shown in Inoue and Kilian (2002a). Under the assumption that the horizon increases with the sample size and that the roots are local to unity, however, the asymptotic distribution will be nonstandard, calling for alternative methods such as the method of Pesavento and Rossi (2006) that we discussed earlier. The incompatibility of these two approaches requires practitioners to choose one or the other. Mikusheva's objective is to provide an approach that works uniformly well for both short horizons and long horizons independently of the presence of a unit root. She accomplishes this objective by applying the grid bootstrap not to the sum of the autoregressive coefficients as in Hansen (1999), but to the LR test statistic of Gospodinov (2004). Her grid bootstrap approach delivers a uniform approximation of the distribution of the reduced-form impulse response statistic over the entire parameter space. For short horizons the approximating distribution is close to normal, while for long horizons it approaches the local-to-unity limit.

These theoretical results help explain the performance of currently used methods of impulse response inference in simulation studies. They are not without important limitations, however. First, like Gospodinov's (2004) analysis, Mikusheva's results apply to the reduced-form impulse responses Φ_h, rather than the structural VAR responses Θ_h, and hence are not immediately relevant for applied work using structural VAR models. While one should be able to relax this simplifying assumption following Gospodinov (2003), this has not yet been done.

Second, unlike the results in Pesavento and Rossi (2006), Mikusheva's theoretical analysis only covers VAR(p) models with at most one large root. It is not uncommon in applied work to deal with multiple large roots. For example, in a

VAR model of the global oil market, both the real price of oil and oil production are highly persistent time series, but there is no reason for these variables to be near-cointegrated. Extending Mikusheva's results to allow for additional large roots, while conceptually straightforward, in practice would require evaluating the LR test statistic on a high-dimensional grid, which would render the construction of the grid bootstrap intervals even more computationally demanding. This is a problem because we have to evaluate this grid repeatedly, as we bootstrap the model, compounding the computational cost of Gospodinov's (2003) method, which itself is already computationally infeasible in realistic VAR models. As a result, Mikusheva's method will remain computationally infeasible in applied work for some time to come. Indeed, she does not demonstrate the feasibility of her grid bootstrap method for models more complex than a univariate AR(2) model.

Finally, we have to ask how well this grid bootstrap works in finite samples. Simulation results for the AR(2) DGPs in Mikusheva (2012) show that the LR grid bootstrap appears to control the coverage accuracy at short and long horizons, as intended, but fails to control coverage accuracy at intermediate horizons. The simulation examples suggest coverage deficiencies of up to 10% for a nominal 95% confidence interval. These coverage deficiencies appear modest, but additional evidence for more realistic models would be useful, as would be a direct comparison to the Bonferroni-based method of Pesavento and Rossi (2006) discussed next.

12.7.4 A Hybrid Method

In closely related work, Pesavento and Rossi (2006) proposed another method of constructing confidence intervals for structural impulse responses in VAR models based on local-to-unity asymptotics. Building on insights in Wright (2000), this hybrid method combines elements of local-to-unity long-run asymptotics and of conventional Gaussian short-run asymptotics based on Bonferroni bounds. There are three steps involved. Under the same assumptions as before, Pesavento and Rossi first compute a $(1-\gamma_1)100\%$ confidence interval $(\lambda_{L,j}, \lambda_{U,j})$ for λ_j, $j=1,\ldots,K$, by inverting augmented Dickey-Fuller (ADF) tests. They use these confidence intervals to compute $e^{\delta\Lambda_L}$ and $e^{\delta\Lambda_U}$, where $\delta > 0$ is implied by $h = [\delta T]$. Second, they estimate the VAR model (12.7.1) in quasi-differences. This may be accomplished either by approximating Ψ_i at horizons $i=1,\ldots,H$, based on the cumulative impulse responses from a VAR in first differences or from the residuals of a VAR(1) in levels. Pesavento and Rossi then construct $(1-\gamma_2)100\%$ confidence intervals for the elements of Ψ_i at horizons $i=1,\ldots,H$, using Lütkepohl's (1990) delta method. Third, for the lower and upper bound of the confidence interval for the elements of Ψ_i, Pesavento and Rossi compute $e^{\delta\lambda_{L,j}}\psi_{jk,i}$ and $e^{\delta\lambda_{U,j}}\psi_{jk,i}$,

where $\psi_{jk,i}$ denotes the jk^{th} element of Ψ_i. Then the confidence intervals for the elements of Ψ_i are given by

$$\left(\min_i e^{\delta\lambda_{L,j}}\psi_{jk,i},\ \max_i e^{\delta\lambda_{U,j}}\psi_{jk,i}\right).$$

By the Bonferroni inequality, the coverage of the pointwise intervals for the elements of Ψ_i should be at least $(1-\gamma_1-\gamma_2)100\%$ at each horizon $i=1,\ldots,H$.

A simulation study in Pesavento and Rossi (2006) suggests that this hybrid method is conservative in that its coverage probability tends to exceed the nominal coverage rate at intermediate horizons. These intervals tend to be substantially more accurate at short horizons than the intervals implied by Pesavento and Rossi's baseline method discussed in Section 12.7.1.

Pesavento and Rossi's hybrid method was explicitly intended for inference on responses of persistent variables. Extensions to VAR models with multiple large roots are straightforward, given the assumption that Λ is diagonal. An open question is how well the proposed approximation works when conducting inference about responses of variables that are not very persistent. If the roots are known to be small, it may make sense to rely on conventional asymptotics for these responses instead. Alternatively, it may be helpful to express these model variables in log-levels rather than differences. Another open issue is how to deal with near-cointegration in the VAR model. In their appendix, Pesavento and Rossi (2006) consider the case of a known cointegration vector but do not discuss how to deal with a setting in which two (or more) model variables may or may not share a common root.

12.7.5 Implications for Second-Stage Inference after Pretesting

It is common among applied users of VAR models to rely on pretests for unit roots and cointegration to determine the appropriate transformation of the time series to be included in the VAR model. Inference is then conducted conditional on the model specification chosen. This approach is questionable. It has been shown that, when applying this procedure to data from a local-to-unity model, even arbitrarily small deviations from exact unit roots may produce impulse response estimates that are severely distorted and confidence intervals that are highly inaccurate. In contrast, the levels specification of the VAR model is robust to these pretest biases.

The Elliott Critique of Pretesting for Unit Roots and Cointegration. Exact unit roots are not an implication of economic theory (see, e.g., Sims 1988). The common perception among economists that there are exact unit roots in

many economic time series can be attributed in no small measure to the inability of unit root tests to reject the unit root null hypothesis (see Nelson and Plosser 1982). In fact, it is common in applied work to impose unit roots even on variables such as interest rates that in conventional economic theory do not exhibit unit root behavior. The problem is that a failure to reject the null hypothesis does not necessarily imply that this null hypothesis is true because unit root tests lack power against local alternatives. Imposing unit roots on the specification of the VAR model based on the results of pretests therefore may do more harm than good.

This problem was first discussed in Elliott (1998), who investigated the effect of pretest for unit roots and rank tests for cointegration in a local-to-unity VAR framework. As before, this framework serves as a device to model situations in which we are unable to determine from the data with any degree of reliability the presence of unit roots or cointegration. Elliott's concern is that in vector autoregressive analysis we are not interested in the presence of unit roots or cointegration for their own sake. Their presence matters only because it affects the specification of the VAR model and hence the distribution of estimators of smooth functions of the parameters of the VAR model. Elliott showed that commonly applied hypothesis tests for the latter objects will typically no longer have their usual χ^2 asymptotic distribution after pretesting. He also showed that confidence intervals for these statistics will lack coverage accuracy after conditioning on the results of pretests.

Implications for Structural Impulse Responses: Simulation Evidence. Elliott's warnings have been ignored by many applied researchers, perhaps because of the perception that estimates of structural impulse responses based on the levels specification are not without their own drawbacks. A systematic simulation analysis of this question is presented in Gospodinov, Herrera, and Pesavento (2013) who focus on the robustness of structural impulse response estimates to pretesting. Their evidence is based on a range of vector autoregressive DGPs with roots of 0.92, 0.95, 0.98, and 1. They show that structural impulse response estimators based on the levels specification have systematically (and often substantially) lower mean-squared error than estimators based on the pretest model, except when the dominant autoregressive root in the DGP is exactly unity. In the latter case, the results are mixed.

As to the coverage accuracy of confidence bands for structural impulse responses, Gospodinov et al. distinguish between models based on short-run identifying restrictions and long-run identifying restrictions. In both cases, intervals based on the bias-adjusted bootstrap method of Kilian (1999b) applied to the levels specification are substantially more accurate than for the pretest model. Related simulation results are also presented in Pesavento and Rossi (2006), who show that pretests undermine the coverage accuracy of

conventional delta method intervals for structural impulse responses, especially for processes with roots close to unity.

Implications for Structural Impulse Responses: Theory. Gospodinov, Maynard, and Pesavento (2011) illustrate why pretesting is particularly problematic in VAR models with near unit roots when using long-run identifying restrictions. The problem is best illustrated in the context of the Blanchard and Quah (1989) model. Recall that $z_t = (\Delta gdp_t, ur_t)'$, where by assumption $z_t \sim I(0)$, but $gdp_t \sim I(1)$. Analogously to model (12.7.1) we may express the bivariate reduced-form model as

$$(I_2 - CL)z_t = \Psi(L)u_t.$$

Building on results in Pesavento and Rossi (2006), Gospodinov et al. show that in this model the matrix C may be parameterized as

$$C = \begin{bmatrix} 1 & \delta \\ 0 & \rho \end{bmatrix},$$

where $\rho = 1 + \lambda/T$, $\lambda < 0$, denotes the dominant root of the unemployment process, and $\delta = -\gamma(1-\rho)$ is a parameter that determines the low-frequency comovement between the unemployment rate and real GDP growth. Equivalently, we may write $A(L)z_t = u_t$, where

$$A(L) = \Psi(L)^{-1} \begin{bmatrix} 1 & \gamma(1-\rho)L \\ 0 & 1-\rho L \end{bmatrix}.$$

Recall that the structural representation of the VAR model is $B(L)z_t = w_t$ such that $w_t = B_0 u_t$. If Σ_w is diagonal and

$$B_0 = \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix},$$

then the long-run multiplier matrix $B(1) = B_0 A(1)$ can be expressed as

$$B(1) = \begin{bmatrix} \psi_{11}(1) - b_{12}\psi_{21}(1) & (1-\rho)\left([\gamma\psi_{11}(1) + \psi_{12}(1)]\right. \\ & \left. -b_{12}\left[\gamma\psi_{21}(1) + \psi_{22}(1)\right]\right) \\ \psi_{21}(1) - b_{21}\psi_{11}(1) & (1-\rho)\left([\gamma\psi_{21}(1) + \psi_{22}(1)]\right. \\ & \left. -b_{21}\left[\gamma\psi_{11}(1) + \psi_{12}(1)\right]\right) \end{bmatrix},$$

where $\psi_{ij}(1)$ denotes the ij^{th} element of $\Psi(1)^{-1}$. As discussed in Chapter 10, imposing Blanchard and Quah's long-run identifying restriction renders $B(1)$ lower triangular, which implies that for $\rho < 0$

$$b_{12} = \left[\gamma\psi_{11}(1) + \psi_{12}(1)\right] / \left[\gamma\psi_{21}(1) + \psi_{22}(1)\right].$$

Note the dependence of this solution on the parameter γ, which in turn determines the low-frequency comovement between the unemployment rate and real GDP growth via the parameter δ.

In contrast, if we changed the Blanchard and Quah model by differencing both real GDP and the unemployment rate, the corresponding VAR long-run multiplier matrix would be

$$B(1) = \begin{bmatrix} \psi_{11}(1) - b_{12}\psi_{21}(1) & \psi_{12}(1) - b_{12}\psi_{22}(1) \\ \psi_{21}(1) - b_{21}\psi_{11}(1) & \psi_{22}(1) - b_{21}\psi_{12}(1) \end{bmatrix}$$

with

$$b_{12} = \psi_{12}(1)/\psi_{22}(1)$$

after imposing that the upper right element of $B(1)$ is zero. Clearly, these two solutions for b_{12} differ, and so do the implied structural impulse responses. Moreover, these differences can be large even for $(1 - \rho)$ arbitrarily close to zero because the value of γ does not depend on ρ.

We conclude that the model with both variables in differences by construction ignores the potential low-frequency comovement between unemployment and real GDP growth. This is not a problem if indeed $\rho = 1$, but even for small departures from $\rho = 1$, such comovement in the data could cause the impulse response estimates to become severely biased. It may seem that it would be safer to express the unemployment rate in levels instead, but, while the latter specification preserves the information about the low-frequency comovement, the implied impulse response estimates obtained by conventional methods are known to be inconsistent if the unemployment rate follows a local-to-unity process (see Gospodinov 2010). They would have to be estimated by an alternative estimator proposed by Gospodinov (2010).

The analysis in Gospodinov, Maynard, and Pesavento (2011) also has important implications for the use of unit root pretests. Given that a unit root pretest for the unemployment rate will fail to reject $\rho = 1$ with high probability, even when the process is only near integrated, the use of pretests effectively amounts to ignoring potential low-frequency comovement, resulting in impulse response estimates that may be severely biased.

The root cause of all these problems, of course, is the fact that the unemployment rate is a fairly persistent process. If we replaced the unemployment rate in Blanchard and Quah's model by a variable with low persistence such as the U.S. economy's capacity utilization rate, for example, standard asymptotic approximations would be adequate.

The local-to-unity approach is but one way of allowing for departures from integer orders of integration. An alternative is the use of fractionally integrated processes, as discussed in Chapters 3 and 11. Tschernig, Weber, and Weigand (2013) also provide evidence that assumptions about variables being either $I(0)$ or $I(1)$ in Blanchard-Quah style models may distort structural impulse responses and inference. They propose to conduct inference about the structural impulse responses in a fractionally integrated VAR model identified

based on long-run exclusion restrictions using a bootstrap approach. Nothing is known about the theoretical validity of this bootstrap approach or its finite-sample accuracy at this point.

12.8 Local Projections

Jordà (2005) proposes an alternative method for estimating structural impulse responses based on so-called local projections. His point of departure is a representation of the standard linear VAR model that we already utilized in Section 4.1, where we used the companion form of a VAR(p) model to derive the representation

$$Y_{t+i} = \mathbf{A}^{i+1} Y_{t-1} + \sum_{j=0}^{i} \mathbf{A}^j U_{t+i-j},$$

where

$$Y_t \equiv \begin{pmatrix} y_t \\ \vdots \\ y_{t-p+1} \end{pmatrix}, \quad \mathbf{A} \equiv \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_K & 0 & & 0 & 0 \\ 0 & I_K & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_K & 0 \end{bmatrix}, \quad \text{and}$$

$$U_t \equiv \begin{pmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and, for simplicity, the intercept has been suppressed. Left-multiplying the equation by $J = [I_K, 0_{K \times K(p-1)}]$ yields

$$y_{t+i} = A^{(i+1)} Y_{t-1} + \sum_{j=0}^{i} \Phi_j u_{t+i-j} = A^{(i+1)} Y_{t-1} + v_{t+i}, \tag{12.8.1}$$

where $i = 0, 1, \ldots, H-1$, $A^{(j)} = J\mathbf{A}^j$ is a $K \times Kp$ matrix, $\Phi_0 = I_K$ by definition, $\Phi_j = J\mathbf{A}^j J'$ for $j = 1, \ldots, H$, is the j^{th} $K \times K$ coefficient matrix of the Wold MA representation of y_t, and $v_{t+i} = \sum_{j=0}^{i} \Phi_j u_{t+i-j}$. Note that the first $K \times K$ block of $A^{(j)}$ is just the j^{th} reduced-form impulse response matrix Φ_j from which the structural impulse responses can be computed as $\Theta_j = \Phi_j B_0^{-1}$.

Jordà observes that the errors in the multivariate linear regression model (12.8.1) are uncorrelated with the regressors. Hence, each equation of the system may be estimated individually by LS. Since Φ_{i+1} is part of the parameter matrix $A^{(i+1)}$, a consistent estimator of Φ_{i+1} may be constructed by LS and a

heteroskedasticity and autocorrelation consistent (HAC) estimator of the error covariance matrix may be employed for inference on Φ_{i+1}.

Because the correlation structure of the error term v_{t+i} is known under the maintained assumption of the DGP being a VAR(p) model, GLS estimation may be used to improve the asymptotic efficiency of the estimator. Thus, an alternative approach is to jointly estimate the system of equations

$$\begin{pmatrix} y_t \\ \vdots \\ y_{t+H-1} \end{pmatrix} = \begin{bmatrix} A^{(1)} \\ \vdots \\ A^{(H)} \end{bmatrix} Y_{t-1} + \begin{pmatrix} v_t \\ \vdots \\ v_{t+H-1} \end{pmatrix}, \tag{12.8.2}$$

which contains all impulse response matrices up to horizon H, $\Phi_1, \ldots, \Phi_H$. Because the autocorrelation structure of the error term of model (12.8.2) can be worked out easily if the true DGP is a VAR(p) process, that structure may be taken into account in GLS estimation.

Given estimates of the Φ_j matrices, the construction of the structural impulse responses is straightforward, provided the structural impact multiplier B_0^{-1} is known. The problem is that the local projection (LP) procedure does not yield an estimate of B_0^{-1}. In this sense, the linear LP estimation procedure is less general than standard linear VAR model estimators, unless the structural impulse responses coincide with the reduced-from impulse responses, as would be the case when considering responses to exogenous variables. In his empirical applications, Jordà deals with this problem by estimating B_0^{-1} from a conventional linear VAR model and substituting this estimate for B_0^{-1} when constructing the LP estimator of the structural impulse responses (see Jordà 2009).

Jordà (2005) suggests that the LP estimator of structual impulse responses has three potential advantages relative to the conventional VAR-based estimator:

1. Impulse responses may be estimated directly by linear LS regressions, simplifying the analysis.
2. Pointwise and joint inference for the impulse responses is straightforward and does not require appealing to the delta method.
3. Local projection estimates of the structural impulse responses are more robust to model misspecification than conventional VAR estimates.

Upon closer examination, none of these alleged advantages hold up to scrutiny. First, using linear LS estimation is indeed simple, but it is also inefficient if the DGP corresponds to the assumed VAR(p). To improve the efficiency of the estimator, one could account for the dependence structure in the error term and use feasible GLS instead of LS. Even feasible GLS

estimation, however, cannot compensate for the much larger number of parameters in the model. Note that there are HpK^2 parameters in the model (12.8.2), many of which are not of direct interest. Only the H impulse response matrices $\Phi_1, \ldots, \Phi_H$, which involve only HK^2 of the HpK^2 estimated parameters, are of direct interest for impulse response analysis. When long horizons are considered, the number HpK^2 can be quite large. In contrast, a VAR(p) model has pK^2 slope parameters only, and the number of parameters to be estimated does not increase with the impulse response horizon. Thus, ultimately, the relative accuracy of the LP estimator and the VAR estimator is an empirical question. Kilian and Kim (2011) show by Monte Carlo simulation that, when the DGP is a linear VAR model, the LP estimator of the structural impulse responses tends to have higher bias as well as higher variance in small samples than the corresponding conventional VAR estimator. Thus, the simplicity of the LP estimator comes at a considerable cost.

Second, it is correct that LP regressions facilitate pointwise inference on the reduced-form impulse responses in that the LS estimates will be asymptotically normal under mild conditions (see Jordà 2009). This result is of no help for conducting inference about structural impulse responses, however. Valid inference in the latter case relies on the same assumptions as required for the conventional VAR-based estimators because of the dependence of the LP impulse response estimator on the VAR estimate of the structural impact multiplier matrix, and it requires appealing to the delta method for inference on structural impulse responses (see also Jordà 2009). Thus, the LP estimator does not relax any of the assumptions of the VAR-based estimator. Nor is the practical performance of the LP estimator of the structural impulse responses encouraging. Simulation evidence in Kilian and Kim (2011) shows that the coverage accuracy of pointwise LP confidence intervals for structural impulse responses tends to be lower than the accuracy of the corresponding VAR confidence intervals, regardless of whether we appeal to first-order asymptotics for the LP estimator or construct LP bootstrap confidence intervals.

Jordà also observes that the joint asymptotic distribution of the reduced-form impulse responses is easily obtained from estimating the LP model (12.8.2). This idea is developed in full in Jordà (2009). The problem, of course, is not deriving the joint asymptotic distribution. In fact, that distribution is also easily derived for the conventional VAR-based estimator under exactly the same assumptions, as we saw earlier in this chapter. The real challenge instead is how to translate this joint asymptotic distribution into valid confidence bands around the structural impulse responses. Jordà (2009) proposes constructing joint confidence bands for a given impulse response function by Scheffé's S-method of simultaneous inference, building on the asymptotic normality of the joint distribution of the structural impulse response coefficients in stationary VAR models (see Scheffé 1953).

It can be shown, however, that the inequalities underlying the construction of this joint confidence band are only approximate and that this confidence band may not have correct coverage probabilities even under ideal conditions, as was pointed out in Wolf and Wunderli (2015) in the closely related context of constructing joint forecast bands. The same reasoning also applies to impulse responses. Thus, it is perhaps not surprising that the simulation evidence in Kilian and Kim (2011) and Lütkepohl, Staszewska-Bystrova, and Winker (2015a) indicates that the joint LP confidence bands proposed by Jordà (2009) and Jordà and Marcellino (2010) tend to have inadequate coverage accuracy.

Third, it is not the case in general that LP estimators are more robust to misspecification than conventional VAR-based estimators. Recall that in linear VAR models reduced-form impulse responses are nonlinear transformations of the VAR slope parameters. Local projections may be viewed as linear approximations to these nonlinear functions. Given that the LP estimator can be constructed from single-equation LS estimation, it might seem at first sight that the LP estimator might be more robust to possible misspecifications of the reduced-form VAR model. This is not necessarily the case. The population impulse response functions by construction depend on the full system of equations because they reflect the interdependencies between all the variables contained in a given model. If that VAR model is misspecified because there are omitted variables, for example, then the impulse responses of the misspecified model do not properly reflect the actual responses of the system. Because the LP estimator is constructed as an appoximation to these misspecified impulse responses, it suffers from the same omitted variable problem.

Nor is it necessarily the case that LP estimators are better suited for approximating VARMA models, because VAR models may be viewed as autoregressive sieves, as discussed in Chapter 2. Indeed, the simulation results in Kilian and Kim (2011) in favor of the conventional VAR-based estimator extend to both VAR and VARMA DGPs.

Finally, it is theoretically possible that the LP estimator of the reduced-form response might be more accurate than a linear VAR model when the DGP is nonlinear, as conjectured by Jordà (2005), but a number of related studies including Marcellino, Stock, and Watson (2006) and Pesaran, Pick, and Timmermann (2011) have provided evidence that the predictive accuracy of linear VAR models for macroeconomic data is higher than that of direct forecasts, provided the VAR model includes enough lags, casting doubt on this conjecture. In any case, if the linearity assumption underlying the VAR model were seriously violated, the LP estimator of the structural impulse response would be invalid by construction because it relies on the linear VAR based estimator of B_0^{-1}. Hence, there is no theoretical basis for Jordà's conjecture that the LP approach would be more robust in this case.

We conclude that there is no evidence to suggest that local linear projections should replace conventional linear VAR models. Given the practical and conceptual limitations of local projections, we do not consider this approach in the remainder of this book.

Although the focus in this chapter is on linear models, it is worth noting that the idea of local projections may be extended to nonlinear models (see Chapter 18). Jordà (2005) discusses a stylized example of a nonlinear local projection. One complication in this case is that accurate nonlinear approximations may require a large number of terms, which tends to undermine estimation efficiency. Another complication is that computing impulse responses from nonlinear local projections requires Monte Carlo integration, because the responses are dependent on the history of the data and the magnitude of the structural shock (see Chapter 18). A third complication is that it is not clear how to obtain an external estimate of the structural impact multiplier matrix in that case.

12.9 Synthesis

We conclude that there is no method of pointwise inference on structural impulse responses (and related objects such as structural forecast error variance decompositions) that works well in all circumstances. It matters how the VAR model is specified and how confidence intervals are constructed. As to the first point, unless we are sure of the nature of the integration and cointegration constraints, a good case can be made for working with a VAR(p) model in levels, provided the model includes an intercept and $p > 1$. Based on recent simulation studies, working with the levels specification tends to produce more accurate results than relying on pretests. In some cases, it will be necessary to use the lag-augmentation approach to ensure asymptotically valid inference. There is at present no simulation evidence on the finite-sample accuracy of structural impulse response inference based on the lag-augmented levels models, but in models with reasonably high lag orders, the inclusion of an extra lag is unlikely to make much of a difference.

As to the second point, bootstrap confidence intervals may improve on the accuracy of the delta method interval for structural impulse responses, but much depends on how these intervals are constructed. Unless the degree of persistence in the process is low, Efron's percentile interval cannot be recommended. For persistent data, Efron's percentile interval based on the bias-adjusted bootstrap method of Kilian (1998c, 1999b) stands out as providing reasonably accurate confidence intervals in most circumstances, at least for VAR models without deterministic time trends. An alternative is the use of the symmetric percentile-t bootstrap interval. The latter is typically preferred over

Hall's percentile interval, provided the statistic of interest is asymptotically normal. It also tends to be more accurate than the equal-tailed percentile-*t* interval in the VAR setting, to the extent that these intervals have been compared by simulation. None of these intervals can be expected to remain accurate at longer horizons, however.

One alternative is the hybrid method of Pesavento and Rossi (2006), which was derived under the assumption of one or more autoregressive roots local to unity. Simulation evidence for this method in the VAR context, however, is limited, and the applicability of this method in the near-cointegrated case remains unclear. Alternative bootstrap methods that are uniformly valid in the VAR parameter space such as the method of Mikusheva (2012) are likely to remain computationally infeasible for some time to come. Local projection estimators, in contrast, although increasingly popular in empirical macroeconomics, are not a practical alternative to VAR-based estimators of structural impulse responses even under stationarity.

12.10 Bayesian Regions of Highest Posterior Density

All methods discussed so far have been motivated from a frequentist point of view. There is a parallel literature on inference about structural impulse responses from a Bayesian point of view (see, e.g., Uhlig 1994, 2005; Sims and Zha 1999). We already reviewed Bayesian inference in some detail in Chapter 5. It is straightforward to extend the algorithms provided there to the problem of constructing the posterior distribution of structural impulse responses and other smooth functions of VAR model parameters. Our discussion, for the time being, assumes that the structural VAR model is exactly identified based on short-run or long-run restrictions. Inference in overidentified models is discussed in Section 12.13.

The Bayesian approach to inference is not only of interest to Bayesian users of VAR models, but Bayesian error bands have also been shown to work reasonably well when evaluated by their coverage properties, although not as well as the bias-adjusted bootstrap (see Kilian 1998c). In fact, under conventional priors, they produce pointwise intervals that asymptotically coincide with frequentist confidence intervals for structural impulse responses in stationary VAR models. This numerical equivalence breaks down, however, in the presence of unit roots in the data.

Our focus in this section is not on providing a comprehensive review of the Bayesian approach to evaluating structural VAR models, but on illustrating – in the context of a widely used Bayesian algorithm – how the Bayesian approach deals with issues such as trends, unit roots, and cointegration that affect frequentist inference about structural impulse responses.

12.10.1 Pointwise Inference on Structural Impulse Responses

It is useful to start with the Bayesian approach most commonly used in applied work. Following Uhlig (2005), consider the VAR(p) model

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

where $u_t \overset{iid}{\sim} \mathcal{N}(0, \Sigma_u)$. Let $Y \equiv [y_1, \ldots, y_T]$, $A \equiv [\nu, A_1, \ldots, A_p]$, $Z \equiv [Z_0, \ldots, Z_{T-1}]$, where $Z_t \equiv (1, y_t', \ldots, y_{t-p+1}')'$. Then the Gaussian ML estimator of (A, Σ_u) is $\widehat{A} = YZ'(ZZ')^{-1}$ and $\widetilde{\Sigma}_u = (Y - \widehat{A}Z)(Y - \widehat{A}Z)'/T$.

Let the prior of the model parameters (A, Σ_u) follow a Gaussian-inverse Wishart distribution of the form

$$\Sigma_u \sim \mathcal{IW}_K(S_*, n), \quad \text{vec}(A)|\Sigma_u \sim \mathcal{N}(\text{vec}(A^*), V_A = V \otimes \Sigma_u),$$

where S_* is a $K \times K$ positive definite covariance matrix and the scalar $n \geq 0$ is a degrees-of-freedom parameter, A^* is the $K \times (Kp + 1)$ prior mean of A, and V is a $(Kp + 1) \times (Kp + 1)$ positive definite matrix.

As derived in Chapter 5, the conjugate posterior distribution is described by

$$\text{vec}(A)|\Sigma_u, Y \sim \mathcal{N}(\text{vec}(\bar{A}), \bar{\Sigma}_A), \quad \Sigma_u|Y \sim \mathcal{IW}_K(S, \tau),$$

where

$$\begin{aligned}
\bar{A} &= (A^* V^{-1} + YZ')(V^{-1} + ZZ')^{-1}, \\
\bar{\Sigma}_A &= (V^{-1} + ZZ')^{-1} \otimes \Sigma_u, \\
S &= T\widetilde{\Sigma}_u + S_* + \widehat{A}ZZ'\widehat{A}' + A^* V^{-1} A^{*\prime} - \bar{A}(V^{-1} + ZZ')\bar{A}', \\
\widehat{A} &= YZ'(ZZ')^{-1},
\end{aligned}$$

and

$$\tau = T + n.$$

Within this framework, a diffuse prior would involve setting

$$\begin{aligned}
n &= 0, \\
V^{-1} &= 0, \\
A^* &= 0, \\
S_* &= 0.
\end{aligned}$$

Substituting these prior parameters into the expressions for the posterior distribution yields

$$\begin{aligned}
\tau &= T, \\
\bar{\Sigma}_A &= (ZZ')^{-1} \otimes \Sigma_u, \\
\bar{A} &= \widehat{A}, \\
S &= T\widetilde{\Sigma}_u.
\end{aligned}$$

Note that this derivation technically violates the maintained assumption of a positive definite S_* and a positive definite V. It is understood, however, that virtually the same results would be obtained if V were chosen to be very large but finite and if S_* were chosen very close to zero, which helps explain why the shortcut of setting $V^{-1} = 0$ and $S_* = 0$ is commonly used in applied work.

Posterior draws from the inverse Wishart distribution are generated as

$$\Sigma_u^{*r}|Y = (\mathcal{R}\mathcal{R}')^{-1},$$

where $\mathcal{R}$ is a $K \times \tau$ matrix consisting of column vectors independently drawn from $\mathcal{N}(0, S^{-1})$. We use * to denote posterior draws. The draws for Σ_u^{*r} are positive definite by construction. Let $\boldsymbol{\alpha}^* = \text{vec}(A^*)$. Given Σ_u^{*r}, we draw

$$\boldsymbol{\alpha}^{*r} \left| \Sigma_u^{*r}, Y \sim \mathcal{N}\left(\text{vec}(\bar{A}), (V^{-1} + ZZ')^{-1} \otimes \Sigma_u^{*r}\right).\right.$$

Given posterior draws $\boldsymbol{\sigma}^{*r} = \text{vech}(\Sigma_u^{*r})$ and $\boldsymbol{\alpha}^{*r}$ for $r = 1, \ldots, R$, it is straightforward to simulate the posterior distribution of the structural impulse response estimator $\widehat{\theta}_{ik,h}^{*r} = g(\boldsymbol{\alpha}^{*r}, \boldsymbol{\sigma}^{*r})$. A credible set is often constructed as the interval

$$\left[\theta_{ik,h,\gamma/2}^*, \quad \theta_{ik,h,1-\gamma/2}^*\right],$$

where $\theta_{ik,h,\gamma/2}^*$ and $\theta_{ik,h,1-\gamma/2}^*$ are the critical points defined by the $\gamma/2$ and $1 - \gamma/2$ quantiles of the posterior distribution of $\widehat{\theta}_{ik,h}^{*r}$. Alternatively, one could form posterior standard error bands

$$\widehat{\theta}_{ik,h}^{*r} \pm z_{\gamma/2}\widehat{\sigma}(\widehat{\theta}_{ik,h}^{*r}), \quad i = 0, 1, 2, \ldots,$$

based on the standard deviation of the simulated $\widehat{\theta}_{ik,h}^{*r}$ draws. The latter approach, however, would at best be valid asymptotically because the finite-sample distribution of the structural impulse responses tends to be highly non-Gaussian (see, e.g, Phillips 1991; Kilian 1999b). Credible sets for structural forecast error variance decompositions may be constructed along the same lines.

Pointwise Inference on Historical Decompositions. An important difference between frequentist and Bayesian methods of inference is that the latter may also be used to construct pointwise credible sets for historical decompositions if the model is estimated in its structural form (see Baumeister and Hamilton 2015b). Recall the structural VAR representation

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

where w_t is the vector of mutually uncorrelated structural shocks. Because the posterior of the structural model parameters reflects our subjective uncertainty about these parameters conditional on the data, it is straightforward to assess the posterior uncertainty about the historical decompositions by drawing from the posterior of the structural model parameters, $[B_0, \ldots, B_p]$, recomputing the

structural shocks for each draw as

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

where $y_1, \ldots, y_T$ denotes the original set of data, and reconstructing the historical decomposition as discussed in Chapter 4. By construction, the cumulative effects of each shock on a given variable will differ across posterior draws, but will add up to the actual demeaned data, allowing us to construct pointwise error bands by simulation.

A similar approach does not work in the frequentist setting because in that setting the only source of uncertainty in $[B_0, \ldots, B_p]$ is sampling uncertainty. A historical decomposition conditions on the historical path of the data. Allowing the data to vary in repeated sampling would defeat the purpose of constructing a historical decomposition. If we wish to understand the determinants of the 1982 recession, for example, considering alternative hypothetical histories, in which there is no recession in 1982, is not helpful. Thus, the Bayesian algorithm discussed in Baumeister and Hamilton (2015b) cannot be adapted to a frequentist setting. There are at present no frequentist methods for conducting inference on historical decompositions.

Limitations of Conventional Methods of Bayesian Inference. The widely used Bayesian algorithm based on a conjugate Gaussian-inverse Wishart prior is not without limitations. While we may relax the assumption of a diffuse prior, the derivation of the posterior remains valid only as long as the prior is symmetric across the equations of the VAR system. This requirement is quite restrictive and violated even when using standard Minnesota priors. As a result, prior to the work of Kadiyala and Karlsson (1997), impulse responses from Bayesian models estimated subject to informative priors have typically not been presented with posterior error bands. Kadiyala and Karlsson's work in turn was extended and generalized to identified VAR models subject to informative priors in Sims and Zha (1998) and Baumeister and Hamilton (2015a). The latter methods specify priors directly on the parameters of the structural VAR representation.

Another limitation of standard Bayesian methods of inference is that the Gaussian model underlying the analysis does not allow for conditional heteroskedasticity. While it is possible to relax this assumption, the standard closed-form solutions for the posterior distribution of the reduced-form VAR model parameters no longer apply in that case. Finding the posterior distribution requires simulation methods (see Uhlig 1997). In practice, most applied users of VAR models are content to rely on methods derived for the Gaussian model.

The Role of Unit Roots and Cointegration in Bayesian Inference. Notwithstanding the obvious differences in motivation and interpretation, exact

finite-sample Bayesian error bands or regions of highest posterior density numerically coincide with frequentist confidence intervals in the asymptotic limit, as long as the model is exactly identified based on short-run or long-run restrictions. This result allows even frequentists to appeal to these intervals with suitable changes in interpretation. This result holds only for stationary models, however. In the unit root case, frequentist and Bayesian intervals differ, as illustrated by Sims and Uhlig (1991).

Formally, none of the Bayesian results we discussed thus far hinges on the presence or absence of a unit root, as the unit root model is measure zero from a Bayesian point of view. This does not mean that Bayesians do not have to worry about the possible presence of unit roots or cointegration. The mirror image of short samples being uninformative about long-run behavior is that priors regarding long-run behavior including the presence of unit roots become important. For example, it has been shown that evidence for trends and for autoregressive unit roots can be quite sensitive to the specification of the prior. The implications of VAR model priors for the long-run behavior of the model can be subtle, inadvertent, and not readily apparent.

Strong beliefs in unit root and cointegration restrictions among applied researchers have motivated the development of alternative Bayesian methods of inference that allow the user to shrink the VAR model parameters toward unit root models (see also Chapter 5). The central insight exploited in that literature is that prior information about a regression model can be introduced in the form of extra dummy observations in the data matrix. One way of allowing for the presence of unit roots is known as the sum-of-coefficients prior (see Doan, Litterman, and Sims 1984). This prior may be implemented with suitably chosen dummy observations (see Chapter 5). In the limit, as the prior restrictions are tightened, this prior implies the presence of a unit root with drift in each VAR equation, which rules out cointegration. An alternative is a prior involving a dummy initial observation proposed in Sims (1993), which allows in the limit for an unspecified number of unit roots without drift, possibly smaller than K, accommodating cointegration. For further discussion and references, see Sims and Zha (1998), who show how to construct the posterior of structural impulse responses within this alternative Bayesian framework.

The main point to take away from this discussion is that Bayesian analysis is not immune from the typical concerns that affect frequentist inference in structural VAR models. The difference is that rather than choosing between alternative asymptotic approximations and model specifications, Bayesians need to keep in mind the implications of their choice of prior.

12.11 Joint Inference on Structural Impulse Responses

As in frequentist inference about structural impulse responses, the focus among Bayesian researchers traditionally has been on pointwise inference. Concerns about this practice were first voiced in Sims and Zha (1999).

Confidence bands obtained by connecting pointwise confidence intervals tend to be too narrow and lack coverage accuracy, resulting in spurious findings of statistical significance.

To the extent that the problem of joint impulse response confidence sets has been discussed in the frequentist VAR literature, it has often been reduced to a problem of conducting joint inference across a range of horizons for a given impulse response function. For example, Jordà (2009) proposes one solution to this problem and Staszewska (2007) and Lütkepohl, Staszewska-Bystrova, and Winker (2015a) propose several alternatives. Simulation evidence on the finite-sample accuracy of these confidence bands is discussed in Kilian and Kim (2009) and Lütkepohl, Staszewska-Bystrova, and Winker (2015a).

It is important to stress that these approaches, while representing an important step forward, are too restrictive for applied work. Most users of structural VAR models are interested in conducting inference about multiple impulse response functions at the same time. Sometimes the economic question of interest by construction involves multiple impulse response functions. For example, the question of whether an oil price shock creates stagflation in the domestic economy by necessity involves studying the responses of inflation as well as real output. Likewise, researchers interested in the effects of a U.S. monetary policy shock care about the responses of both real output and inflation because the loss function of the Federal Reserve depends on both real output and inflation.

It is also common for researchers to be interested in assessing the implications of economic theory for a range of different impulse response functions simultaneously. For example, Blanchard (1989) uses a macroeconomic VAR model to evaluate the implication of standard Keynesian models that (1) positive demand innovations increase output and decrease unemployment persistently, and (2) a favorable supply shock triggers an increase in unemployment without a decrease in output. This example involves inference about four impulse response functions simultaneously. There are even cases in which users of structural VAR models are interested in studying the responses of all model variables to all structural shocks simultaneously. A good example is recent structural VAR models of the global market for industrial commodities such as crude oil (e.g., Kilian 2009). A proper solution to this problem requires taking account of the dependence of all structural impulse responses of interest, not just of the responses in a given impulse response function. Next, we review several methods designed for this task.

12.11.1 Joint Confidence Sets for Structural Impulse Responses

The Bonferroni Method. The simplest method of constructing a joint confidence set for structural impulse responses involves constructing Bonferroni bounds. The construction of these bounds is straightforward because it does

not utilize measures of the dependence across the structural impulse response coefficients (see Lütkepohl, Staszewska-Bystrova, and Winker 2015b). Let θ denote the $M \times 1$ vector of structural impulse responses of interest. Recall that $\theta = g(\boldsymbol{\alpha}, \boldsymbol{\sigma})$, where $g(\cdot)$ denotes a nonlinear transformation of $\boldsymbol{\alpha} = \text{vec}([A_1, \ldots, A_p])$ and $\boldsymbol{\sigma} = \text{vech}(\Sigma_u)$. Let $\widehat{\eta}$ denote any estimator of $\eta \equiv (\boldsymbol{\alpha}, \boldsymbol{\sigma})$ such that

$$\sqrt{T}(\widehat{\eta} - \eta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_\eta),$$

where Σ_η is an unknown nonsingular covariance matrix.

The $(1-\gamma)100\%$ Bonferroni band is

$$B_{1-\gamma} = [l_1, u_1] \times \cdots \times [l_M, u_M],$$

where for $m = 1, \ldots, M$, the lower bound l_m and the upper bound u_m are the $\gamma/(2M)$ and $(1-\gamma/(2M))$ quantiles of the marginal distribution of the m^{th} component of $\widehat{\theta}$. In practice, this distribution may be approximated by bootstrapping the reduced-form VAR model.

A well-known result is that this joint confidence band is conservative in that

$$\mathbb{P}\left[\theta \in B_{1-\gamma}\right] \geq 1 - \gamma.$$

Even for VAR models of moderate dimension, the dimension of θ tends to be large. In a recursively identified model the dimension of $\theta' = [\text{vech}(\Theta_0)', \text{vec}([\Theta_1, \ldots, \Theta_H])']$ is $M = K^2H + K(K+1)/2$. For example, for $K = 3$ and a maximum horizon of $H = 12$, $M = 114$. For a nominal 68% joint confidence band, this would require choosing individual confidence levels of $(1 - \gamma/M) = 1 - 0.32/114 = 0.9972$. Estimating the tail quantiles therefore may require a very large number of bootstrap draws. As Lütkepohl, Staszewska-Bystrova, and Winker (2015b) observe, in some cases, only the responses to specific shocks are of interest, which helps ameliorate this problem somewhat.

The Wald Approach of Lütkepohl, Staszewska-Bystrova, and Winker (2015b). An alternative approach to constructing joint confidence sets for structural impulse responses was suggested by Lütkepohl, Staszewska-Bystrova, and Winker (2015b). Consider the Wald test statistic

$$W(\eta_0) = T(\widehat{\eta} - \eta_0)' \widehat{\Sigma}_\eta^{-1} (\widehat{\eta} - \eta_0),$$

where $\widehat{\Sigma}_\eta$ is an estimator of Σ_η. An approximation of the small-sample distribution of this Wald test statistic may be obtained by bootstrapping the reduced-form vector autoregression. Let $\widehat{\eta}^{*r}$ and $\widehat{\Sigma}_\eta^{*r}$ denote the r^{th} bootstrap realization, where $r = 1, \ldots, R$. Define the bootstrap Wald test statistic

$$W^{*r} = T(\widehat{\eta}^{*r} - \widehat{\eta})' \left(\widehat{\Sigma}_\eta^{*r}\right)^{-1} (\widehat{\eta}^{*r} - \widehat{\eta}).$$

Upon ordering the bootstrap estimates of the test statistic such that $W^{*1} \leq \cdots \leq W^{*R}$, choose the Wald confidence set for η such that $100\gamma\%$ of the bootstrap realizations W^{*r} exceed the Wald test statistic obtained from the original data:

$$W_{1-\gamma}^{\eta} = \left\{\eta | W(\eta) \leq W^{*(1-\gamma)R}\right\}.$$

As T and R go to infinity, under the usual conditions underlying applications of the bootstrap,

$$\mathbb{P}\left[W(\eta) \leq W^{*(1-\gamma)R}\right] \to 1-\gamma,$$

and hence asymptotically, for the true parameters η_0,

$$\mathbb{P}\left[\eta_0 \in W_{1-\gamma}^{\eta}\right] \to 1-\gamma.$$

Given the set of η values in $W_{1-\gamma}^{\eta}$, the implied joint confidence region for the impulse response estimator $\widehat{\theta}$ may be inferred from the mapping $\theta = g(\eta)$. If this function is one-to-one, as would be the case in standard recursively identified models when $H \leq p$, the set

$$W_{1-\gamma}^{\theta=g(\eta)} = \left\{g(\eta) | \eta \in W_{1-\gamma}^{\eta}\right\}$$

is an asymptotically exact $1-\gamma$ joint confidence set. If $\theta = g(\eta)$ is not one-to-one, this region has at least $1-\gamma$ confidence level.

Plotting this joint confidence set is impractical in realistic applications. Lütkepohl, Staszewska-Bystrova, and Winker (2015b) therefore propose to represent this joint confidence set in the form of a confidence band. Let $\widehat{\theta}_m^{*r} = g_m(\widehat{\eta}^{*r})$ denote the r^{th} bootstrap realization of the m^{th} element of θ obtained from $\widehat{\eta}^{*r}$. Then a joint confidence band for θ may be constructed by determining the lower bounds

$$l_m = \min\left\{\widehat{\theta}_m^{*r} | \widehat{\eta}^{*r} \in W_{1-\gamma}^{\eta}\right\}$$

and upper bounds

$$u_m = \max\left\{\widehat{\theta}_m^{*r} | \widehat{\eta}^{*r} \in W_{1-\gamma}^{\eta}\right\},$$

for $m = 1, \ldots, M$, and defining

$$W_{1-\gamma}^{\theta=g(\eta),\text{band}} = [l_1, u_1] \times \cdots \times [l_M, u_M].$$

Values of the structural impulse responses outside of this confidence band can be rejected with $100(1-\gamma)\%$ confidence. Lütkepohl, Staszewska-Bystrova, and Winker (2015b) provide evidence based on stationary bivariate VAR(1) models with varying degrees of persistence that these joint confidence bands have reasonably accurate coverage rates. They tend to be more conservative

and wider than the Bonferroni bands. For larger models, one would expect the Bonferroni bands to be conservative.

The Wald Approach of Inoue and Kilian (2016). An alternative Wald approach was proposed by Inoue and Kilian (2016). Their proposal is to invert the Wald test statistic for the structural impulse response parameters of interest, denoted by

$$W = (\widehat{\theta} - \theta_0)' \widehat{\Sigma}_{\widehat{\theta}}^{-1} (\widehat{\theta} - \theta_0),$$

where $\widehat{\Sigma}_{\widehat{\theta}}$ is an estimate of the variance-covariance matrix of $\widehat{\theta}$. One difficulty in this context is that the asymptotic distribution of $\sqrt{T}(\widehat{\theta} - \theta_0)$ is degenerate when the number of structural responses exceeds the number of model parameters. Thus, even in stationary vector autoregressions, the joint asymptotic distribution of $\widehat{\theta}$ may be degenerate and hence the distribution of W nonstandard (see, e.g., Lütkepohl and Poskitt 1991). This degeneracy may be overcome by transforming the estimator appropriately. Inoue and Kilian show that the joint Wald test statistic is invariant to this transformation and can be approximated by bootstrap methods, providing a theoretical justification for constructing joint confidence sets for $\widehat{\theta}$ from the bootstrap realizations W^{*r}, $r = 1, \ldots, R$.

To test a given null hypothesis, $\mathbb{H}_0 : \theta = \theta_0$, the value of the original Wald statistic

$$W = (\widehat{\theta} - \theta_0)' \widehat{\Sigma}_{\widehat{\theta}^*}^{-1} (\widehat{\theta} - \theta_0), \tag{12.11.1}$$

where

$$\widehat{\Sigma}_{\widehat{\theta}^*} = \frac{1}{R} \sum_{r=1}^{R} (\widehat{\theta}^* - \widehat{\theta})(\widehat{\theta}^* - \widehat{\theta})', \tag{12.11.2}$$

must be compared with the $100(1 - \gamma)$ percentile of the empirical distribution of the bootstrap Wald test statistics

$$W^{*r} = (\widehat{\theta}^{*r} - \widehat{\theta})' \widehat{\Sigma}_{\widehat{\theta}^{*r}}^{-1} (\widehat{\theta}^{*r} - \widehat{\theta}), \tag{12.11.3}$$

for $r = 1, 2, \ldots, R$. Generating the bootstrap critical values requires a nested bootstrap loop, because for each bootstrap realization of the Wald statistic the term $\widehat{\Sigma}_{\widehat{\theta}^{*r}}$ in turn must be evaluated by bootstrap simulation.

In the absence of a specific null value, θ_0, a joint confidence set $W_{1-\gamma}^{\theta}$ may be constructed by comparing the value of each of the bootstrap Wald test statistics

$$\widetilde{W}^{*r} = (\widehat{\theta} - \widehat{\theta}^{*r})' \widehat{\Sigma}_{\widehat{\theta}^*}^{-1} (\widehat{\theta} - \widehat{\theta}^{*r}), \tag{12.11.4}$$

against the $100(1 - \gamma)$ percentile of the distribution of the bootstrap statistic (12.11.3). If the value of the statistic (12.11.4) is less than this critical value, the

bootstrap draw $\widehat{\theta}^{*r}$ is retained and becomes a member of the joint confidence set $W_{1-\gamma}^{\theta}$. As T and R go to infinity,

$$\mathbb{P}\left[\theta_{\diamond} \in W_{1-\gamma}^{\theta}\right] \to 1-\gamma,$$

where $\theta_{\diamond}$ denotes the true value of θ. Each realization of θ contained in $W_{1-\gamma}^{\theta}$ may be viewed as one structural model.

Inoue and Kilian (2016) provide simulation evidence that this bootstrap Wald joint confidence set is reasonably accurate even in large-dimensional and highly persistent VAR models, whereas the Bonferroni method is conservative. The latter result is not unexpected when the number of structural impulse responses is large. Although $W_{1-\gamma}^{\theta=g(\eta)}$ has theoretical advantages compared with $W_{1-\gamma}^{\theta}$ because it restricts estimation to the parameters in η, in practice the differences in coverage accuracy tend to be small. Simulation evidence suggests that in the rare cases, in which there is a larger difference in accuracy, the joint confidence region $W_{1-\gamma}^{\theta}$ is more accurate.

Plotting the joint confidence region $W_{1-\gamma}^{\theta}$ is impractical in realistic applications. One way of representing the joint confidence set is to form a joint confidence band. Let $\widehat{\theta}_m^{*r}$ denote the r^{th} bootstrap realization of the m^{th} element of θ. Then

$$W_{1-\gamma}^{\theta,\text{band}} = [l_1, u_1] \times \cdots \times [l_M, u_M],$$

where

$$l_m = \min_r \left\{\widehat{\theta}_m^{*r} | \widehat{\theta}^{*r} \in W_{1-\gamma}^{\theta}\right\}$$

and

$$u_m = \max_r \left\{\widehat{\theta}_m^{*r} | \widehat{\theta}^{*r} \in W_{1-\gamma}^{\theta}\right\}$$

for $m = 1, \ldots, M$.

This approach is sufficient to evaluate whether zero responses can be rejected at conventional significance levels, but results in a loss of information because one is no longer able to discern the evolution over time of the impulse response functions associated with any one structural model estimate in the joint confidence set. It is precisely this evolution of the response function that users of structural VAR models typically are interested in (see, e.g., Cochrane 1994). For example, many macroeconomists have abandoned frictionless neoclassical models and adopted models with nominal or real rigidities based on VAR evidence of sluggish or delayed responses of inflation and output (see, e.g., Woodford 2003). This is also true in other applications. Whereas macroeconomists may be interested in whether a response function for real output is hump-shaped or not, users of structural VAR models in international economics may be interested in whether there is delayed overshooting in the response of the exchange rate to monetary policy shocks. It is difficult

to answer such questions about the shape of a given impulse response function based on two-dimensional joint confidence bands, because such bands typically are consistent with a multitude of different response patterns. These difficulties are compounded when considering the analysis of more than one impulse response function at a time, as is common in applied work.

Inoue and Kilian (2016) therefore propose a complementary approach to evaluating joint confidence sets of the form $W_{1-\gamma}^{\theta}$ (or $W_{1-\gamma}^{\theta=g(\eta)}$) based on plotting all sets of impulse responses associated with the structural models contained in the joint Wald confidence set. As a result, each impulse response function viewed in isolation looks like a shotgun trajectory chart. Unlike conventional representations of confidence sets, this "shotgun plot" may be frayed around the edges. The joint confidence bands discussed earlier can be viewed as an envelope fit around this shotgun trajectory plot. This envelope may be interpreted as a conservative two-dimensional joint confidence band, but that band is only a side product of this approach. The main objective instead is to characterize the shapes and patterns of the set of impulse responses. Although the shotgun trajectory plot does not lend itself to learning about the responses associated with individual structural models in the joint confidence set, it provides important additional information not conveyed by joint confidence bands, as the following two examples from Inoue and Kilian (2016) illustrate.

A VAR Model of Monetary Policy. Consider a quarterly VAR model of monetary policy for inflation, growth in real output, and the short-term interest rate. The model is identified recursively with the interest rate ordered last. One question of economic interest is whether the implied response of real output to an unexpected monetary policy expansion is hump-shaped, as suggested by standard business cycle models. The conventional joint confidence band cannot answer this question, because this band is wide enough to accommodate any number of shapes of this impulse response function, some consistent with economic theory and some not (see Figure 12.2).

Having formally defined a hump-shaped response function (for example, as the response function reaching its maximum between horizon 1 and $H-1$, possibly strengthened by additional monotonicity restrictions), Inoue and Kilian's proposal is to run an iterative search on the response functions contained in the 68% joint confidence set to determine whether the response functions in the set are consistent or inconsistent with this hump shape. Evidence that all response functions in the 68% confidence set are hump shaped would lead us to reject the hypothesis that there is no hump in the response function. Evidence that none of the response functions in the set is hump shaped, in contrast, would imply a rejection of the hypothesis of a hump-shaped response. Evidence that some response functions in the set are hump shaped and some are not means that the data are not dispositive about the presence or absence

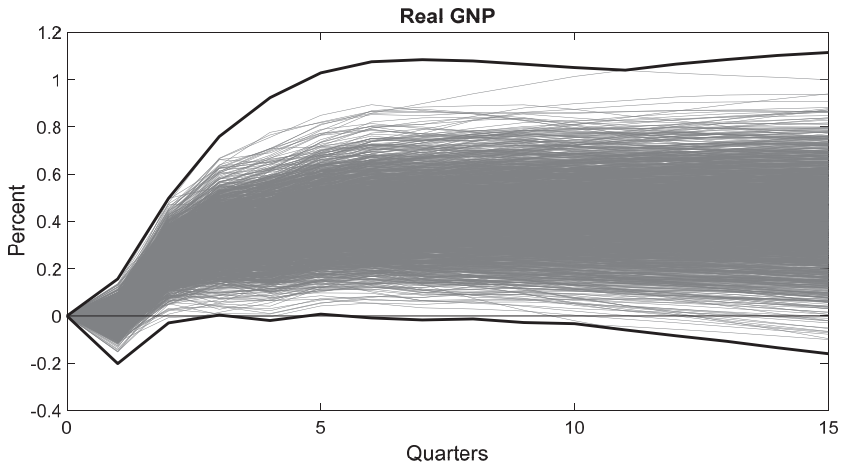

Figure 12.2. Responses of U.S. real GNP to an unexpected loosening of monetary policy: Shotgun plot implied by joint 68% Wald confidence set.
Notes: Obtained by cumulating the responses of the growth rate of real GNP and flipping the sign of the monetary policy shock in the estimated VAR model. The joint confidence band obtained by constructing an envelope around the shotgun plot is shown as bold lines.
Source: Inoue and Kilian (2016).

of a hump shape at the 68% confidence level. This exercise provides additional information about the properties of the estimated model that could not have been inferred from conventional joint confidence bands. As Table 12.1 shows, in this example, the hump shape cannot be ruled out at conventional confidence levels, nor can the hypothesis be rejected that there is no hump shape in the response of real GNP.

A VAR Model of the Stagflationary Effects of Oil Price Shocks. Often the hypotheses of interest to economists involve more than one response function in the same structural model. This situation can be illustrated using the example of a VAR model for the U.S. economy including the real price of oil, inflation (π), and growth in real output (Δgdp). The question of interest is whether oil price shocks have stagflationary effects on the U.S. economy. Answering this question requires us to consider simultaneously two response functions for the same structural model. Specifically, we need to verify whether the responses to an oil price shock of Δgdp_{t+h} and $\Delta\pi_{t+h}$ (the latter having been obtained by differencing the responses of π_{t+h}) are of opposite sign for all horizons of interest for the models in the joint confidence set $W^{\theta}_{1-\gamma}$ (see, e.g., Kilian 2008c). It is immediately obvious that this type of information cannot be inferred from joint confidence bands, but may be computed based on the shotgun plot. Moreover,

Table 12.1. *Percentage of Models in Joint Wald Confidence Set Consistent with a Hump-Shaped Response Function of Real GNP to an Unexpected Loosening of Monetary Policy*

	Joint Confidence Level	
Definition of hump shape	68%	95%
No piecewise monotonicity imposed	69.0%	70.1%
Piecewise monotonicity imposed	15.6%	15.3%

Notes: Based on a minimal definition of a hump-shaped response as a response reaching its maximum between horizon 1 and 15. The optional piecewise monotonicity constraint restricts the slope of the impulse response function to be positive to the left of the peak response and negative to the right of the peak. Source: Inoue and Kilian (2016).

to the extent that there is evidence of some models in $W_{1-\gamma}^{\theta}$ exhibiting stagflationary responses, the shotgun plot allows us to assess graphically the extent to which the subset of these models is associated with cumulative reductions in real GDP growth, which is the ultimate question of interest to users of this structural model.

It would be tempting to try to answer this question instead by observing that the data must be consistent with stagflationary responses as long as both joint confidence bands include positive as well as negative values, so hypothetically a stagflationary response fits within these bands. Inoue and Kilian (2016) provide an empirical example, in which nevertheless none of the structural models in question exhibit stagflationary behavior, reinforcing the point that joint confidence bands may obscure important information about the shapes of impulse response functions. Figure 12.3 shows that even when applying the definition of stagflation only to horizons up to one year, there are few stagflationary responses in the 68% confidence set and none that imply a large recessionary effect four years later.

12.11.2 Joint Credible Sets

The same problems of joint inference arise when using Bayesian methods of estimation and inference. A solution to this problem within the conventional Bayesian framework discussed earlier has been proposed by Inoue and Kilian (2013).

Highest-Posterior Density Joint Credible Sets. Consider a structural VAR model that is fully identified by imposing a recursive structure on B_0^{-1}. An example would be the global oil market model of Kilian (2009). When the results apply specifically to recursively identified models, we denote this specific B_0^{-1} by P, where P is the lower-triangular Cholesky decomposition of

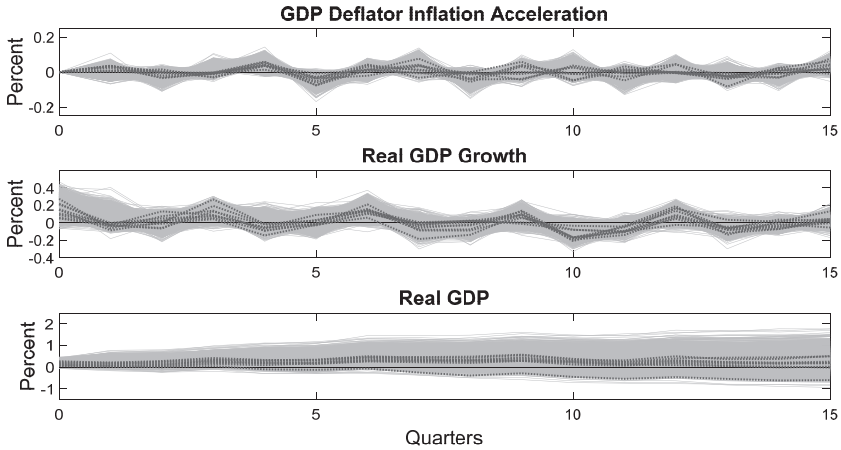

Figure 12.3. Responses to an unexpected increase in the real price of oil: Shotgun plot implied by joint 68% Wald confidence set with subset of stagflationary responses highlighted.
Notes: The pairs of response functions associated with the models in the 68% joint confidence set that exhibit stagflationary responses at horizons $1, \ldots, 4$ and a positive inflation acceleration response at horizon 1 are highlighted as dark dotted lines. There are no models with stagflationary responses at all horizons.
Source: Inoue and Kilian (2016).

Σ_u such that $PP' = \Sigma_u$. Any such structural VAR model can be defined in one of two ways: (a) as the set of the first $p+1$ structural impulse responses $\theta = [\text{vech}(P)', \text{vec}(\Phi_1 P)', \text{vec}(\Phi_2 P)', \ldots, \text{vec}(\Phi_p P)']'$, or (b) as the vector $(\text{vec}(A)', \text{vech}(P)')'$, where $A \equiv [A_1, \ldots, A_p]$. In fact, there is a one-to-one mapping between these representations, given that the nonlinear function $\theta = h(\text{vec}(A), \text{vech}(P))$ is known and Σ_u is nonsingular. This fact allows the derivation of the joint distribution of θ by the change-of-variable method, if the joint distribution of $(\text{vec}(A)', \text{vech}(P)')$ is Gaussian-inverse Wishart, as postulated earlier. The posterior density takes the form

$$f(\theta) \propto \left(\left| \frac{\partial \theta}{\partial [\text{vec}(A)', \text{vech}(P)']} \right| \right)^{-1} \left| \frac{\partial \text{vech}(\Sigma_u)}{\partial [\text{vech}(P)']} \right| f(A|\Sigma_u) f(\Sigma_u),$$

where the determinant of the Jacobian has a closed form. This result allows us to rank each structural model θ by its posterior density value. Given this ranking, a $100(1-\gamma)\%$ highest posterior density (HPD) credible set can be defined as

$$S = \{\theta | \, f(\theta) \geq c_\gamma\}, \tag{12.11.5}$$

where $f(\theta)$ is the posterior density of θ and c_γ is the largest constant such that

$$\mathbb{P}(S) \geq 1 - \gamma.$$

In practice, we rank each structural model θ^{*r}, $r = 1, \ldots, R$, obtained by drawing from the Gaussian-inverse Wishart posterior by its posterior density value $f(\theta^{*r})$, which can be readily computed given $(\boldsymbol{\alpha}^{*r}, \boldsymbol{\sigma}^{*r})$, where $\boldsymbol{\alpha}^{*r} = \text{vec}(A^{*r})$ and $\boldsymbol{\sigma}^{*r} = \text{vech}(\Sigma_u^{*r})$. Then a 68% joint credible set, for example, would be constructed by simply retaining the 68% of the posterior draws θ^{*r} with the highest $f(\theta^{*r})$ values. These credible sets differ from conventional error bands for structural impulse responses in that the elements of the credible set consist of all K^2 structural impulse response functions up to some prespecified horizon, generated by a given structural model draw. When plotting the K^2 structural impulse response functions associated with each member of the credible set, the resulting plot of the credible set will exhibit a shotgun trajectory pattern, as discussed earlier. The evaluation of this credible set may proceed analogously to that of the joint confidence set in Inoue and Kilian (2016).

This baseline result applies to models that are fully identified, as opposed to partially identified. Partial identification here relates to a situation in which only a subset of the structural shocks is identified. For example, in many VAR models of monetary policy, we are only interested in identifying responses to monetary policy shocks with the other shocks remaining unidentified. Ranking structural models that are only partially identified requires the marginalization of the joint posterior $f(\theta)$ with respect to the model parameters that do not enter the posterior distribution of the structural impulse responses of interest, before joint credible sets may be constructed (see Inoue and Kilian 2013).

Although our discussion focused on recursively identified structural VAR models, with suitable changes in notation this approach can be generalized to non-recursive identification schemes. The latter results, of course, must be worked out on a case-by-case basis, depending on the nature of the identifying restrictions.

The Sims-Zha Method. An alternative Bayesian solution to the problem of joint inference was proposed by Sims and Zha (1999). Their baseline method involves draws from the first and second moment of the (unspecified) joint distribution of the posterior impulse responses, which will be a good approximation only if the joint distribution is close to Gaussian. One concern is that it is well known that the finite-sample joint posterior distribution of impulse responses is far from Gaussian, so this method will be at best an approximation, whereas the method in Inoue and Kilian (2013) is exact by construction. The other concern is that the marginalized posterior is Gaussian only if the joint posterior is Gaussian, making it problematic to conduct such joint inference in partially identified models without doing Monte Carlo integration by numerical methods.

In recognition of the first concern, Sims and Zha suggest an alternative method designed to account for asymmetry in the marginal distribution of the

impulse responses. This method is based on quantiles of the marginal distribution constructed from the same draws used for the baseline method. This proposal indeed accounts for asymmetry in the approximate marginal distribution, but it does not address the second concern regarding the non-Gaussianity of the underlying joint distribution.

Optimal Pointwise versus Optimal Joint Inference. Baumeister and Hamilton (2015b) recently argued that pointwise inference on the structural impulse responses is preferred in applied work because it is optimal from a Bayesian point of view. They made the case for reporting the vector of pointwise medians or means of the posterior of the structural responses. This approach, however, is optimal only under loss functions that imply that the user is not concerned with the dependence of the structural responses across horizons and across response functions. In other words, the user does not care about the shapes of the impulse response functions or the comovement across response functions.

Specifically, Baumeister and Hamilton (2015b) stress that under the loss function

$$\omega_0|\theta_{jk,0} - \widehat{\theta}_{jk,0}| + \cdots + \omega_h|\theta_{jk,h} - \widehat{\theta}_{jk,h}|,$$

the optimal estimate, for any set of positive weights $\{\omega_i\}_{i=0}^{h}$, is the vector of pointwise posterior medians. Although this solution is optimal conditional on this specific choice of loss function, of course, so are the solutions obtained under alternative loss functions. What is at issue therefore is not the optimality of the solution to a given loss function, but the appropriate choice of the loss function.

The solution to the loss function underlying the construction of the modal model in Inoue and Kilian (2013), for example, is as optimal as the solution proposed in Baumeister and Hamilton (2015b), but it is derived under a different loss function that arguably is more in line with the interests of applied researchers because it takes account of the dependence across structural responses. Thus, it avoids the problems of interpreting median response functions documented in Fry and Pagan (2011), Kilian and Murphy (2012), and Inoue and Kilian (2013).

Summary. It is rare for researchers to be interested in pointwise inference on structural impulse responses or related statistical objects. The conventional approach of constructing confidence bands for structural impulse responses from pointwise confidence intervals is known to be potentially misleading. With the recent introduction of methods of constructing truly joint confidence regions one would expect pointwise methods to be gradually replaced

in applied work. If we are interested in constructing joint confidence sets, both the Bonferroni method and the approach based on the Wald test statistic may be used. Simulation evidence in Inoue and Kilian (2016) suggests that both types of joint confidence bands have reasonably accurate coverage rates in realistic settings, even when the DGP is highly persistent and the maximum horizon considered is high. The coverage of Bonferroni-based methods of joint inference, however, tends to be conservative when the number of structural impulse responses of interest is large.

For a given confidence level, joint confidence sets tend to be wider than pointwise intervals if the same confidence level is used, making it more difficult for applied users to produce statistically significant results. It is, however, important to account for the true uncertainty in these estimates. In addition, joint inference may be used to investigate the shapes and patterns of impulse response functions, as emphasized in the work of Inoue and Kilian (2016), allowing researchers to address more directly the implications of economic theory.

12.12 Other Bootstrap Applications

It should be noted that the bootstrap methods discussed in this chapter are intended for approximating the unconditional distribution of statistics such as structural impulse responses or structural forecast error variance decompositions. These methods are not suitable, for example, for generating bootstrap prediction intervals because in prediction we must condition on the same last p observations as in the actual data (see Chapter 2).

12.12.1 Bootstrap Prediction

There are at present no bootstrap methods for approximating the distribution of the data conditional on the p most recent observations that do not effectively require Gaussian model innovations (see, e.g., Berkowitz and Kilian 2000). One practical solution to this problem is to bootstrap the unconditional distribution of the slope parameters, as discussed earlier, exploiting the fact that their conditional distribution under stationarity will converge to the unconditional distribution asymptotically. Consider, for example, the problem of constructing the predictive density of the reduced-form VAR(p) model

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

where $u_t \overset{iid}{\sim} F$. In that case, we can approximate the DGP by

$$y_t^* = \widehat{\nu} + \widehat{A}_1 y_{t-1}^* + \cdots + \widehat{A}_p y_{t-p}^* + u_t^*,$$

where u_t^* is generated by drawing with replacement from the estimated residuals. Having generated a draw for $[\widehat{\nu}^*, \widehat{A}_1^*, \ldots, \widehat{A}_p^*]$ from this bootstrap DGP, for each such draw we construct predictions by iterating forward the model conditional on the p most recent observations:

$$\begin{aligned} y_{t+1}^* &= \widehat{\nu}^* + \widehat{A}_1^* y_t + \widehat{A}_2^* y_{t-1} + \cdots + \widehat{A}_1^* y_{t-p+1} + u_{t+1}^* \\ y_{t+2}^* &= \widehat{\nu}^* + \widehat{A}_1^* y_{t+1}^* + \widehat{A}_2^* y_t + \cdots + \widehat{A}_1^* y_{t-p+2} + u_{t+2}^* \\ &\vdots \end{aligned}$$

Once the predictive disribution has been simulated by bootstrap, the construction of prediction intervals based on the percentiles of the predictive distribution is straightforward. It is useful to keep in mind that prediction intervals differ from confidence intervals in that they relate to a random variable (rather than a model parameter) and hence are intended to include future realizations with a prespecified probability. Pascual, Romo, and Ruiz (2004) provide a proof of the asymptotic validity of this approach in the univariate AR(p) model. Kim (1999, 2004) provides evidence that the accuracy of bootstrap prediction intervals may be improved by relying on the bias adjustments proposed by Kilian (1998c) for structural impulse response analysis. As in the case of impulse response analysis, the case can be made that inference should be joint across the horizons of the forecasts rather than pointwise. Wolf and Wunderli (2015) discuss the construction of joint bootstrap prediction regions for VAR models and VECMs and prove their asymptotic validity.

12.12.2 Bootstrapping the Critical Values of Test Statistics

The bootstrap algorithms discussed so far must also be modified if we are interested in constructing critical values for test statistics. In that case, the bootstrap DGP is obtained from the best-fitting model under the null hypothesis rather than from the best-fitting unrestricted model. Consider the example of testing for Granger causality in a bivariate stationary VAR(p) model of the form

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} + \sum_{i=1}^{p} \begin{bmatrix} a_{11,i} & a_{12,i} \\ a_{21,i} & a_{22,i} \end{bmatrix} \begin{pmatrix} y_{1,t-i} \\ y_{2,t-i} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}.$$

Under the null hypothesis that y_{2t} is not Granger-causal for y_{1t}, we have $a_{12,i} = 0$, $i = 1, 2, \ldots, p$. Thus, the bootstrap DGP under the null hypothesis reduces to

$$\begin{pmatrix} y_{1t}^* \\ y_{2t}^* \end{pmatrix} = \begin{pmatrix} \widehat{\nu}_1 \\ \widehat{\nu}_2 \end{pmatrix} + \sum_{i=1}^{p} \begin{bmatrix} \widehat{a}_{11,i} & 0 \\ \widehat{a}_{21,i} & \widehat{a}_{22,i} \end{bmatrix} \begin{pmatrix} y_{1,t-i}^* \\ y_{2,t-i}^* \end{pmatrix} + \begin{pmatrix} u_{1t}^* \\ u_{2t}^* \end{pmatrix},$$

where the coefficient estimates are obtained by restricted ML or restricted GLS methods (see Chapter 2). The bootstrap innovations are generated by drawing

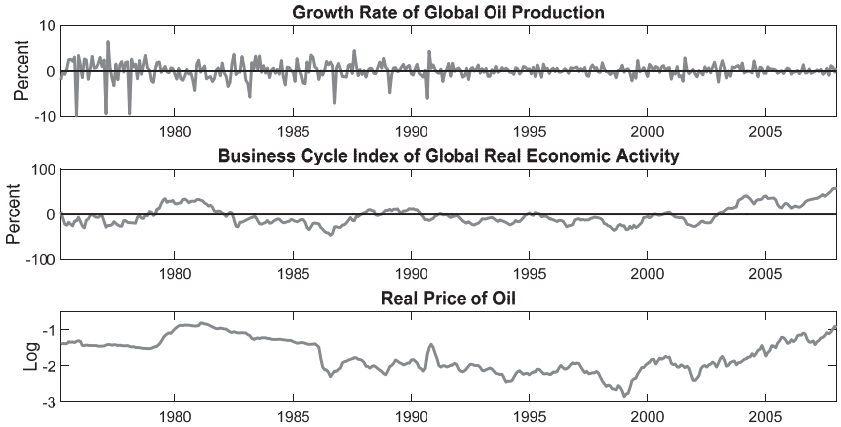

Figure 12.4. Global oil market data.
Notes: The sample period is 1973m2–2007m12. For a description of the data sources, see Kilian (2009).

with replacement from the estimated residuals. The power of the test may be improved by utilizing residuals from the unrestricted model rather than the restricted model, because the latter are consistent estimates under both the null and under the alternative hypothesis, whereas the former are consistent estimates only under the null hypothesis.

12.13 Examples of Impulse Response Confidence Intervals

12.13.1 An Exactly Identified Model

For expository purposes we consider the recursively identified monthly structural VAR(24) model of the global market for crude oil of Kilian (2009), as already discussed in Chapter 8. Recall that $y_t = (\Delta prod_t, rea_t, rpoil_t)$, where $\Delta prod_t$ denotes the percent change in world crude oil production, rea_t is a business cycle index measuring global real economic activity, and $rpoil_t$ is the log of the real price of oil. The model is estimated by LS with an intercept included. The estimation period is 1973m2–2007m12. Figure 12.4 shows the $T - 24$ net sample observations of the transformed time series. The dominant root of the estimated process is 0.989. The identifying assumptions imply that

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 & 0 \\ b_0^{21} & b_0^{22} & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} \end{bmatrix} \begin{pmatrix} w_t^{\text{oil supply}} \\ w_t^{\text{aggregate demand}} \\ w_t^{\text{oil-specific demand}} \end{pmatrix},$$

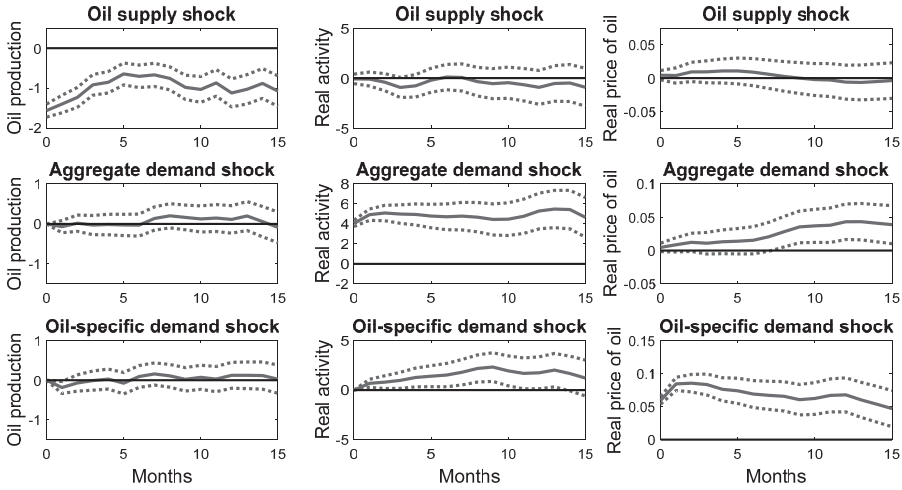

Figure 12.5. 95% delta method confidence intervals based on bootstrap standard error estimates.

Notes: All estimates were generated using the data in Figure 12.4 and the recursively identified structural VAR model in Kilian (2009).

allowing us to estimate B_0^{-1} as the lower-triangular Cholesky decomposition of $\widehat{\Sigma}_u$.

Pointwise confidence intervals for the structural impulse responses can be generated by a wide range of alternative methods, Figure 12.5 plots the point estimates of the structural impulse responses together with 95% pointwise confidence intervals constructed by the delta method. Rather than relying on asymptotic closed-form solutions for the standard errors of the structural impulse responses, we generated estimates of the standard errors from 500 bootstrap replications of the estimated VAR(24) process under the assumption that the innovations are iid with unknown distribution.

The oil supply shocks in Figure 12.5 have been normalized to represent responses to unexpected oil supply disruptions. Because the lower-triangular Cholesky decomposition generates responses to positive one-standard-deviation shocks, this normalization involves multiplying all responses to oil supply shocks by -1. The responses of global oil production are obtained by cumulating the responses of the growth rate of global oil production.

Figure 12.5 shows that negative oil supply shocks are associated with a statistically significant decline in oil production at all horizons, but the positive responses of the real price of oil and the negative responses of global real activity are not individually statistically significant. In contrast, positive aggregate demand shocks (defined as demand shocks that affect both the market for crude oil and the market for other globally traded industrial commodities) have a statistically significant positive effect on global real activity at all horizons

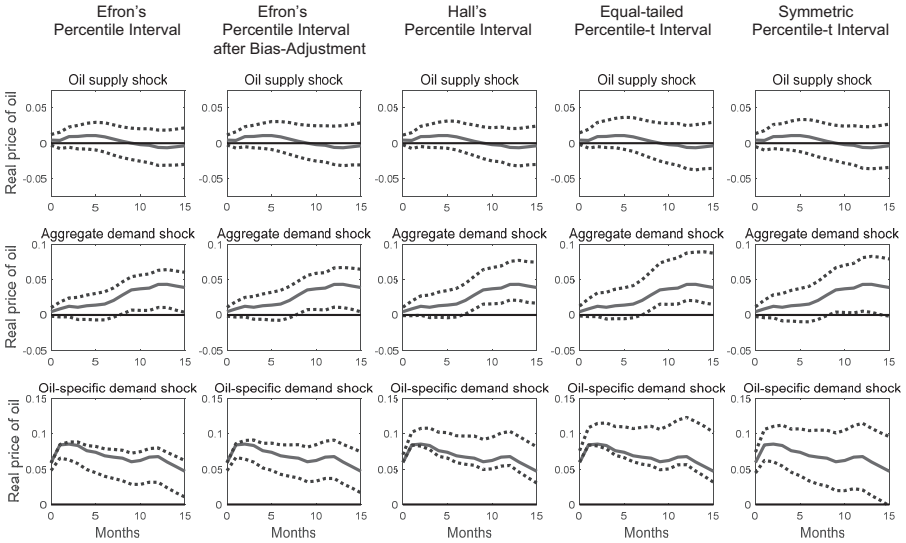

Figure 12.6. Alternative 95% bootstrap confidence intervals.
Notes: All estimates were generated using the data in Figure 12.4 and the recursively identified structural VAR model in Kilian (2009).

and cause a statistically significant increase in the real price of oil at horizons of eight or more months. They do not have a statistically significant effect on global oil production at any horizon, however. Finally, positive oil-market-specific demand shocks trigger a highly statistically significant increase in the real price of oil at all horizons. They are also associated with statistically significant increases in global real activity at horizons up to 13 months and a statistically significant decline in global oil production at horizon 1.

For comparison, Figure 12.6 presents the corresponding pointwise 95% bootstrap confidence intervals obtained by applying a recursive-design bootstrap algorithm under the premise that the VAR model innovations are iid. Estimates of the bootstrap interval endpoints are generated from 2,000 bootstrap replications. To conserve space, only the responses of the real price of oil are shown.

The first column of Figure 12.6 shows confidence intervals constructed from Efron's percentile interval. These interval endpoints tend to be downward biased relative to the point estimate, as is evident especially in the response of the real price of oil to the oil-market specific demand shock. The second column shows results for Efron's percentile interval obtained after applying the first-order mean bias correction to the slope parameters, as proposed in Kilian (1999b). The bias estimates were constructed using the closed-form solution of Pope (1990), as discussed in Section 2.3.3. As a result of the bias adjustment,

the interval endpoints tend to be shifted upward, as is especially evident for the response of the real price of oil to the oil-specific demand shock.[13]

Hall's percentile interval applied to the original LS estimate of the slope parameters embodies an implicit bias adjustment for the structural impulse response estimator, which causes an even more dramatic upward shift in the bootstrap confidence interval endpoint, as shown in the third column. These adjustments are further compounded when using the equal-tailed percentile-*t* interval in the fourth column. In contrast, the lower interval endpoints of the symmetric percentile-*t* interval in the last column are closer to those in the second column, while the upper interval endpoints tend to be considerably higher. The standard error estimates for the symmetric percentile-*t* intervals were obtained by bootstrapping each bootstrap estimate of the VAR model.

Notwithstanding these differences, from an economic point of view, all methods agree that a positive aggregate demand shock causes a statistically significant increase in the real price of oil at horizons beyond about half a year, that negative oil supply shocks do not have a statistically significant effect on the real price of oil, and that positive oil-market-specific demand shocks cause a highly statistically significant increase in the real price of oil at all horizons.

12.13.2 Guarding against Conditional Heteroskedasticity

The case can be made that at least some of the innovations in the Kilian (2009) oil market model are conditionally heteroskedastic. This fact suggests that allowing for conditional heteroskedasticity may improve the accuracy of inference. Figure 12.7 compares three alternative versions of Hall's percentile interval constructed using a recursive design bootstrap under the assumption of iid errors (as in Figure 12.6), constructed using the multivariate generalization of the recursive-design residual wild bootstrap of Gonçalves and Kilian (2004), and constructed using the recursive design residual block bootstrap of Brüggemann, Jentsch, and Trenkler (2016). We use a block length $l = 36$. The results are not very sensitive to the choice of l. Figure 12.7 illustrates that often the wild bootstrap implies intervals that are wider than the iid bootstrap, as expected. The residual-based block bootstrap, which is the only method designed to fully protect against conditional heteroskedasticity at all horizons, tends to imply even wider intervals. Overall, nevertheless, the results are quite similar.

[13] It should be noted that it is possible for the least-squares point estimate to be above the upper interval endpoint. Likewise, it is possible for the LS estimator to lie below the lower interval endpoint. Only the delta method interval and the symmetric percentile-*t* interval impose symmetry on the interval, preventing such situations from arising. This does not mean that there is anything wrong with confidence intervals not imposing symmetry. Rather, evidence of the point estimate lying outside the confidence band could be an indication of the LS point estimate being biased.

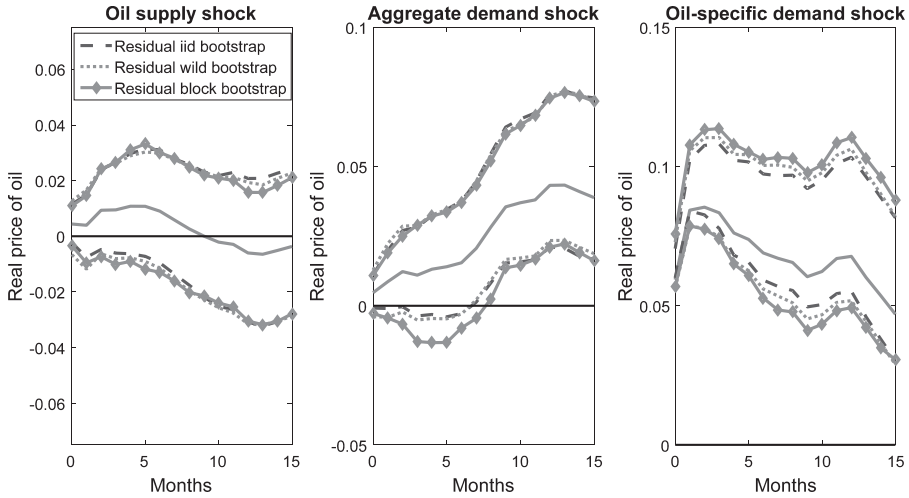

Figure 12.7. Responses of the real price of oil to oil demand and supply shocks with alternative 95% confidence intervals.
Notes: All results are based on Hall's percentile interval. The residual block bootstrap relies on block length $l = 36$. Other block sizes produced similar results.

12.13.3 Extensions to Overidentified Models

Sometimes, a researcher is interested in imposing one or more overidentifying restrictions on an already exactly identified structural VAR model. A case in point is the Kilian (2009) model. Recall that

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 & 0 \\ b_0^{21} & b_0^{22} & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} \end{bmatrix} \begin{pmatrix} w_t^{\text{oil supply}} \\ w_t^{\text{aggregate demand}} \\ w_t^{\text{oil-specific demand}} \end{pmatrix}.$$

Given that we already imposed the identifying restriction that an oil-specific demand shock does not lower global real economic activity on impact, it would seem natural to impose a similar restriction on the impact effect of oil supply shocks on global real economic activity. This reasoning suggests the presence of an overidentifying restriction such that $b_0^{21} = 0$:

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 & 0 \\ 0 & b_0^{22} & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} \end{bmatrix} \begin{pmatrix} w_t^{\text{oil supply}} \\ w_t^{\text{aggregate demand}} \\ w_t^{\text{oil-specific demand}} \end{pmatrix}.$$

After imposing this extra restriction, B_0^{-1} is no longer recursive and standard method-of-moments estimators no longer apply because there are more identifying restrictions than unknowns.

Estimation Subject to Overidentifying Restrictions. As discussed in Chapter 9, constructing the GMM estimator of overidentified models requires choosing the unknown elements of B_0^{-1} to minimize

$$J = T(\text{vech}(\widehat{\Sigma}_u) - \text{vech}(B_0^{-1}B_0^{-1\prime}))'\widehat{W}(\text{vech}(\widehat{\Sigma}_u) - \text{vech}(B_0^{-1}B_0^{-1\prime})),$$

where all identifying and overidentifying restrictions have been imposed on B_0^{-1} and $\widehat{W}$ is the inverse of the estimate of the variance-covariance matrix of the sample moment restrictions.

Recall that we wish to impose four exclusion restrictions such that

$$B_0^{-1} = \begin{bmatrix} b_0^{11} & 0 & 0 \\ 0 & b_0^{22} & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} \end{bmatrix}.$$

There are five unknown elements in B_0^{-1}. In the absence of the overidentifying restriction

$$\widehat{P} = \text{chol}(\widehat{\Sigma}_u) = \begin{bmatrix} 1.5617 & 0 & 0 \\ 0.0735 & 4.0588 & 0 \\ -0.0044 & 0.0047 & 0.0594 \end{bmatrix}.$$

A reasonable initial guess for minimizing J therefore is

$$\begin{bmatrix} 1.5617 & 0 & 0 \\ 0 & 4.0588 & 0 \\ -0.0044 & 0.0047 & 0.0594 \end{bmatrix},$$

where the overidentifying restriction has simply been imposed on the $(1, 2)$ element of $\widehat{P}$.

Next we need to construct the weighting matrix. Let $u_t = (u_{1,t}, u_{2,t}, u_{3,t})'$. Then $\widehat{W} = (\overline{W}'\overline{W}/T)^{-1}$, where $\overline{W}$ is obtained by demeaning the elements of the matrix

$$\begin{bmatrix} \widehat{u}_{1,1}^2 & \widehat{u}_{1,1}\widehat{u}_{2,1} & \widehat{u}_{1,1}\widehat{u}_{3,1} & \widehat{u}_{2,1}^2 & \widehat{u}_{2,1}\widehat{u}_{3,1} & \widehat{u}_{3,1}^2 \\ \widehat{u}_{1,2}^2 & \widehat{u}_{1,2}\widehat{u}_{2,2} & \widehat{u}_{1,2}\widehat{u}_{3,2} & \widehat{u}_{2,2}^2 & \widehat{u}_{2,2}\widehat{u}_{3,2} & \widehat{u}_{3,2}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{u}_{1,T}^2 & \widehat{u}_{1,T}\widehat{u}_{2,T} & \widehat{u}_{1,T}\widehat{u}_{3,T} & \widehat{u}_{2,T}^2 & \widehat{u}_{2,T}\widehat{u}_{3,T} & \widehat{u}_{3,T}^2 \end{bmatrix}$$

which is constructed by stacking vectors of observations for the $K(K+1)/2$ elements of $\text{vech}(\widehat{u}_t\widehat{u}_t')$. $\overline{W}$ in this example is of dimension $T \times 6$. Thus

$$\widehat{W} = \begin{bmatrix} 0.0000 & 0.0000 & 0.0001 & -0.0000 & 0.0000 & -0.0003 \\ 0.0000 & 0.0000 & 0.0000 & -0.0000 & 0.0000 & 0.0001 \\ 0.0001 & 0.0000 & 0.0167 & -0.0000 & -0.0002 & 0.1078 \\ -0.0000 & -0.0000 & -0.0000 & 0.0000 & -0.0000 & -0.0002 \\ 0.0000 & 0.0000 & -0.0002 & -0.0000 & 0.0027 & -0.0291 \\ -0.0003 & 0.0001 & 0.1078 & -0.0002 & -0.0291 & 3.5177 \end{bmatrix} \times 10^4.$$

Given $\widehat{W}$ and the initial guess for B_0^{-1}, the expression for J may be iterated to convergence using a suitable numerical optimizing routine.[14] We obtain the GMM estimator

$$\widehat{B}_0^{-1} = \begin{bmatrix} 1.5624 & 0 & 0 \\ 0 & 4.0604 & 0 \\ -0.0044 & 0.0047 & 0.0594 \end{bmatrix}.$$

The J statistic is minimized with respect to the unknown elements of B_0^{-1} at

$$\widehat{J} = 0.1311.$$

The point estimates in Figure 12.8 show that not only the estimate of the structural impact multiplier matrix B_0^{-1} but also the implied structural impulse response functions are quite similar to those in the recursively identified model.

Testing for Overidentifying Restrictions. Before imposing an overidentifying restriction, it is useful to test its validity. It has to be stressed that this test is conditional on the remaining identifying assumptions being valid. It is not a test of the validity of the set of identifying restrictions more generally. As discussed in Chapter 9, under the null hypothesis that the overidentifying restrictions are correct,

$$\widehat{J} \xrightarrow{d} \chi^2(n),$$

where n is the number of overidentifying restrictions. In our empirical example, $n = 1$ and we are unable to reject the overidentifying restriction, given the asymptotic p-value of 0.7173.

Inference Subject to Overidentifying Restrictions. Bootstrap inference for the GMM estimator of the structural impulse responses in overidentified structural VAR models may be conducted based on the results in Hall and Horowitz (1996). The key difference compared with the bootstrap methods already

[14] For example, the function *fminsearch* in the MATLAB optimization toolbox finds a vector x such that a given function $f(x)$ is minimized.

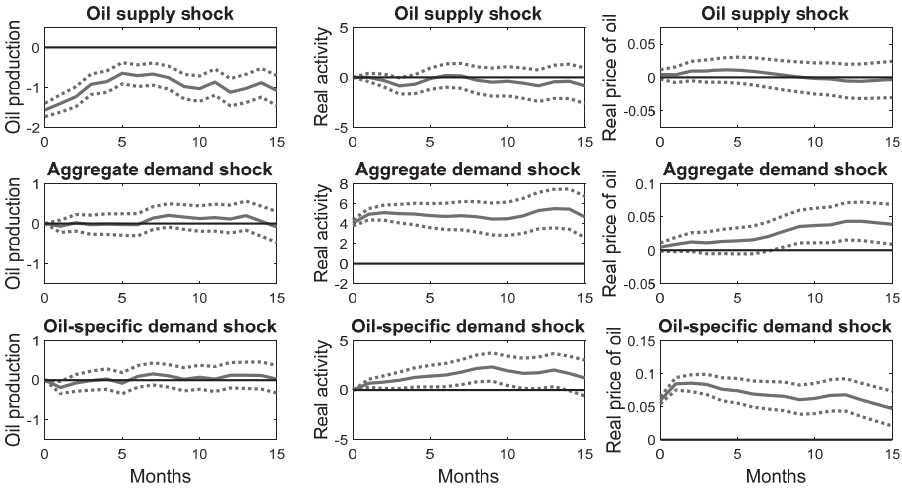

Figure 12.8. 95% delta method confidence intervals based on bootstrap standard error estimates in the overidentified model.
Notes: All estimates were generated using the data in Figure 12.4 and the recursively identified structural VAR model in Kilian (2009) with the overidentifying restriction $b_0^{21} = 0$ imposed.

discussed for exactly identified structural VAR models is that the bootstrap moment conditions must be recentered. This recentering is necessary because in overidentified models the population moment conditions hold with equality, whereas the sample moment conditions evaluated at the GMM estimator by construction do not. Thus, we need to recenter the bootstrap moment conditions and choose the unknown elements of B_0^{*-1} to minimize

$$J^* = T(\text{vech}(\widehat{\Sigma}_u^*) - \text{vech}(B_0^{*-1}B_0^{*-1\prime}) - c)'\widehat{W}^* \times(\text{vech}(\widehat{\Sigma}_u^*) - \text{vech}(B_0^{*-1}B_0^{*-1\prime}) - c),$$

where $c = \text{vech}(\widehat{\Sigma}_u) - \text{vech}(\widehat{B}_0^{-1}\widehat{B}_0^{-1\prime})$ is a vector of constants and $\widehat{B}_0^{-1}$ denotes the GMM estimate of B_0^{-1}. Otherwise, the construction of bootstrap confidence intervals proceeds exactly as discussed earlier. Figure 12.8 shows the GMM point estimates of the structural impulse responses together with pointwise 95% delta method confidence intervals. The standard errors of the impulse response estimates were estimated by bootstrap methods with the bootstrap moment condition suitably recentered. The results are virtually indistinguishable from those in Figure 12.5, which did not impose the overidentifying restriction.

The most common use of overidentifying restrictions in the literature has been to discriminate between alternative models of the transmission of monetary policy shocks by considering a structural VAR model that encompasses

multiple channels of transmission, the existence of which implies additional overidentifying restrictions (see, e.g., Bernanke 1986; Sims 1986; Bernanke and Mihov 1998b; Gordon and Leeper 1994; Leeper, Sims, and Zha 1996). Other applications of overidentified structural VAR models include Galí (1992), Fisher (2006), and Inoue, Kilian, and Kiraz (2009).

Alternative Estimation Methods for Overidentified Models. As discussed in Chapter 9, the model of interest in this empirical example cannot be estimated by IV methods because the identifying restrictions are imposed on B_0^{-1} rather than B_0. An alternative estimation approach would be to employ the FIML estimator for overidentified structural VAR models.

Another alternative to the GMM estimator for overidentified structural VAR models is the use of Bayesian estimation methods. As in the case of bootstrap methods, care must be exercised in evaluating the posterior of the structural impulse responses in the overidentified case. The question of how to conduct inference in overidentified structural VAR models from a Bayesian point of view was first discussed in Sims and Zha (1998, 1999). Sims and Zha showed that standard methods of computing regions of high posterior density for structural impulse responses are not valid when the structural VAR model is overidentified, because overidentifying restrictions imply restrictions on the error covariance matrix. As discussed in Chapter 9, in this case it is necessary to derive the posterior from the structural VAR representation rather than the reduced-form VAR representation. The posterior of the structural impulse responses is no longer Gaussian-inverse Wishart and has to be evaluated using an importance sampler or Metropolis-Gibbs sampler.[15]

These Bayesian methods, however, are designed for structural VAR models with linear identifying restrictions on B_0 and cannot be used in the current context. It is not clear whether the generalized Bayesian estimation method by Canova and Pérez Forero (2015), which allows particular nonlinear restrictions on B_0, could be adapted to the estimation of overidentified structural VAR models with restrictions on B_0^{-1}.

[15] In related work, Waggoner and Zha (2003) propose a Gibbs sampler. The latter approach is applicable to a wider class of overidentified models including models with restrictions on lagged slope coefficients. Such restrictions may arise from exogeneity restrictions on the VAR model, for example, and differ from the class of overidentifying restrictions on the error covariance matrix considered here.