Essays in Macroeconometrics

A dissertation presented

by

Mikkel Plagborg-Moller

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

May 2016

Dissertation Advisor:
Professor James H. Stock

Author:
Mikkel Plagborg-Moller

Essays in Macroeconometrics

Abstract

This dissertation consists of three independent chapters on econometric methods for macroeconomic analysis.

In the first chapter, I propose to estimate structural impulse response functions from macroeconomic time series by doing Bayesian inference on the Structural Vector Moving Average representation of the data. This approach has two advantages over Structural Vector Autoregression analysis: It imposes prior information directly on the impulse responses in a flexible and transparent manner, and it can handle noninvertible impulse response functions.

The second chapter, which is coauthored with B. J. Bates, J. H. Stock, and M. W. Watson, considers the estimation of dynamic factor models when there is temporal instability in the factor loadings. We show that the principal components estimator is robust to empirically large amounts of instability. The robustness carries over to regressions based on estimated factors, but not to estimation of the number of factors.

In the third chapter, I develop shrinkage methods for smoothing an estimated impulse response function. I propose a data-dependent criterion for selecting the degree of smoothing to optimally trade off bias and variance, and I devise novel shrinkage confidence sets with valid frequentist coverage.

Contents

Preface

This dissertation consists of three independent chapters on econometric methods for macroeconomic analysis.

In the first chapter, I propose to estimate structural impulse responses from macroeconomic time series by doing Bayesian inference on the Structural Vector Moving Average representation of the data. This approach has two advantages over Structural Vector Autoregressions. First, it imposes prior information directly on the impulse responses in a flexible and transparent manner. Second, it can handle noninvertible impulse response functions, which are often encountered in applications. To rapidly simulate from the posterior of the impulse responses, I develop an algorithm that exploits the Whittle likelihood. The impulse responses are partially identified, and I derive the frequentist asymptotics of the Bayesian procedure to show which features of the prior information are updated by the data. I demonstrate the usefulness of my method in a simulation study and in an empirical application that estimates the effects of technological news shocks on the U.S. business cycle.

The second chapter, which is joint work with B. J. Bates, J. H. Stock, and M. W. Watson, considers the estimation of approximate dynamic factor models when there is temporal instability in the factor loadings. We characterize the type and magnitude of instabilities under which the principal components estimator of the factors is consistent and find that these instabilities can be larger than earlier theoretical calculations suggest. We also discuss implications of our results for the robustness of regressions based on the estimated factors and

of estimates of the number of factors in the presence of parameter instability. Simulations calibrated to an empirical application indicate that instability in the factor loadings has a limited impact on estimation of the factor space and diffusion index forecasting, whereas estimation of the number of factors is more substantially affected.

In the third chapter, I develop a method for optimally smoothing an estimated impulse response function. The degree of smoothing can be selected based on an unbiased estimate of the mean squared error, thus trading off bias and variance. The smoothing procedure is a member of a flexible and computationally convenient class of shrinkage estimators applicable to both time series and panel data. I give conditions under which the smoothed estimator dominates the initial non-smooth estimator in terms of mean squared error. I develop novel shrinkage confidence sets with valid frequentist coverage in a finite-sample normal location model with arbitrary known covariance structure. The finite-sample results imply uniform asymptotic validity of the confidence sets even when normality fails.

Acknowledgements

I experienced some of the best years of my life during my time as a Ph.D. student, and I have a great many people to thank for that.

I am indebted to my advisors for their help, guidance, and generosity – in every meaning of these words. Jim Stock took me under his wing right from the start and inspired me with his open mind and breadth of interests. I could always count on Gita Gopinath's support and sharp advice on applied and macro theory projects. Neil Shephard taught me so many things in just three years that I envy his current students. Gary Chamberlain challenged me to expand my horizons and never failed to provide deep insights.

A special thanks goes to Isaiah Andrews and Anna Mikusheva, who time and again gave superb feedback on my half-baked ideas. I am fortunate to have coauthored with and learned a lot from Sophocles Mavroeidis and Mark Watson.

My graduate school experience could never have been the same without my fellow students and occasional coauthors in the department. I thank my lucky stars that I was in the same cohort as David Rezza Baqaee, Aubrey Clark, Peter Ganong, Duncan Gilchrist, Ben Hébert, Simon Jäger, Rohan Kekre, Guillaume Pouliot, Martin Rotemberg, and Fernando Yu. I feel equally privileged to be friends with Anand Krishnamurthy, Eben Lazarus, Daniel Lewis, Pepe Montiel Olea, Alex Peysakhovich, Daniel Pollmann, Jesse Schreger, and Jann Spiess. I cannot wait to see what the future holds in store for all of us.

I am grateful for the academic and financial support provided by the Department of

Economics. I received invaluable feedback from participants at Harvard's econometrics, macroeconomics, and international lunch seminars. Brenda Piquet, the glue that holds the department together, kindly guided me through administrative matters.

As always, family was the keystone of my graduate student life. My wonderful wife Sumin kept me, not just (mostly) sane, but fundamentally happy and excited through the academic peaks and valleys. I cherish the memory of the day the admissions email told us we would be able to spend our time in graduate school together. As if that were not fortunate enough, we even had the pleasure of living close to my sister-in-law Sujin and her partner Chris, fierce board game competitors and trusted confidants. My brothers Emil and Nikolaj were never more than a Skype call away and always ready for a Star Wars marathon during holidays. Finally, I owe my parents Annette and Jakob more than I could say. Their encouragement and unconditional support made everything possible.

In addition to the above-mentioned people, this dissertation has benefited from the input of several others. Chapter 1: I received helpful comments from Regis Barnichon, Varanya Chaubey, Gabriel Chodorow-Reich, Christian Matthes, Frank Schorfheide, Elie Tamer, Harald Uhlig, and seminar participants at Cambridge, Chicago Booth, Columbia, Harvard Kennedy School, MIT, Princeton, UCL, UPenn, and the 2015 DAEiNA meeting; I also thank Eric Sims for sharing his news shock code and Marco Lippi for providing a key part of the proof of Theorem A.1. Chapter 2: My coauthors and I thank Herman van Dijk, Allan Timmermann, and two anonymous referees. Chapter 3: Max Kasy, Adam McCloskey, and Emi Nakamura gave valuable feedback, and I thank Mark Gertler and Peter Karadi for making their data available online.

To

my wife,

my mother,

and the everlasting memory of my father.

Chapter 1

Bayesian Inference on Structural Impulse Response Functions

1.1 Introduction

Since Sims (1980), Structural Vector Autoregression (SVAR) analysis has been the most popular method for estimating the impulse response functions (IRFs) of observed macro variables to unobserved shocks without imposing a specific equilibrium model structure. However, the SVAR model has two well-known drawbacks. First, the under-identification of the parameters requires researchers to exploit prior information to estimate unknown features of the IRFs. Existing inference methods only exploit certain types of prior information, such as zero or sign restrictions, and these methods tend to implicitly impose unacknowledged restrictions. Second, the SVAR model does not allow for noninvertible IRFs. These can arise when the econometrician does not observe all variables in economic agents' information sets, as in models with news shocks or noisy signals.

I propose a new method for estimating structural IRFs: Bayesian inference on the Structural Vector Moving Average (SVMA) representation of the data. The parameters of this model are the IRFs, so prior information can be imposed by placing a flexible Bayesian prior distribution directly on the parameters of scientific interest. My SVMA approach thus

overcomes the two drawbacks of SVAR analysis. First, researchers can flexibly and transparently exploit all types of prior information about IRFs. Second, the SVMA model does not restrict the IRFs to be invertible *a priori*, so the model can be applied to a wider range of empirical questions than the SVAR model. To take the SVMA model to the data, I develop a posterior simulation algorithm that uses the Whittle likelihood approximation to speed up computations. As the IRFs are partially identified, I derive the frequentist asymptotic limit of the posterior distribution to show which features of the prior are dominated by the data.

The first key advantage of the SVMA model is that prior information about IRFs – the parameters of scientific interest – can be imposed in a direct, flexible, and transparent manner. In standard SVAR analysis the mapping between parameters and IRFs is indirect, and the IRFs are estimated by imposing zero or sign restrictions on short- or long-run impulse responses. In the SVMA model the parameters are the IRFs, so all types of prior information about IRFs may be exploited by placing a prior distribution on the parameters. While many prior choices are possible, I propose a multivariate Gaussian prior that facilitates graphical prior elicitation: Sketch the prior means for each impulse response in a plot, then place prior confidence bands around the means, and finally specify prior information about the smoothness (i.e., prior correlation) of the IRFs. In particular, researchers can easily and transparently exploit valuable prior information about the shapes and smoothness of IRFs.

The second key advantage of the SVMA model is that, unlike SVARs, it does not restrict IRFs to be invertible *a priori*, which broadens the applicability of the method. The IRFs are said to be invertible if the current shocks can be recovered as linear functions of current and past – but not future – data. As shown in the literature, *non*invertible IRFs can arise when the econometrician does not observe all variables in the economic agents' information sets, such as in macro models with news shocks or noisy signals. A long-standing problem for standard SVAR methods is that they cannot consistently estimate noninvertible IRFs because the SVAR model implicitly assumes invertibility. Proposed fixes in the SVAR literature either

exploit restrictive model assumptions or proxy variables for the shocks, which are not always available. In contrast, the SVMA model is generally applicable since its parametrization does not impose invertibility on the IRFs *a priori*.

I demonstrate the practical usefulness of the SVMA method through a simulation exercise and an empirical application. The simulations show that prior information about the smoothness of IRFs can sharpen posterior inference about unknown features of the IRFs, since smoother IRFs have fewer effective free parameters. Prior information about smoothness has not been explicitly exploited in the SVAR literature, because the shapes and smoothness of SVAR IRFs are complicated functions of the underlying SVAR parameters.

My empirical application estimates the effects of technological news shocks on the U.S. business cycle, using data on productivity, output, and the real interest rate. Technological news shocks – signals about future productivity increases – have received much attention in the recent macro literature. My analysis is the first to fully allow for noninvertible IRFs without dogmatically imposing a particular Dynamic Stochastic General Equilibrium (DSGE) model. I use the sticky-price DSGE model in E. Sims (2012) to guide prior elicitation. My results overwhelmingly indicate that the IRFs are noninvertible, implying that no SVAR can consistently estimate the IRFs in this dataset; nevertheless, most IRFs are precisely estimated by the SVMA procedure. The news shock is found to be unimportant for explaining movements in TFP and GDP, but it is an important driver of the real interest rate.

To conduct posterior inference about the IRFs, I develop a posterior simulation algorithm that exploits the Whittle (1953) likelihood approximation. Inference in the SVMA model is challenging due to the flexible parametrization, which explains the literature's preoccupation with the computationally convenient SVAR alternative. I overcome the computational challenges of the SVMA model by simulating from the posterior using Hamiltonian Monte Carlo (HMC), a Markov Chain Monte Carlo method that is well-suited to high-dimensional models (Neal, 2011). HMC evaluates the likelihood and score 100,000s of times in realistic

applications, so approximating the exact likelihood with the Whittle likelihood drastically reduces computation time. The resulting algorithm is fast, asymptotically efficient, and easy to apply, while allowing for both invertible and noninvertible IRFs.

Having established a method for computing the posterior, I derive its frequentist large-sample limit to show how the data updates the prior information. Because the IRFs are partially identified, some aspects of the prior information are not dominated by the data in large samples.[1] I establish new results on the frequentist limit of the posterior distribution for a large class of partially identified models under weaker conditions than assumed by Moon & Schorfheide (2012). I then specialize the results to the SVMA model with a non-dogmatic prior, allowing for noninvertible IRFs and non-Gaussian structural shocks. I show that, asymptotically, the role of the data is to pin down the true autocovariances of the data, which in turn pins down the reduced-form (Wold) impulse responses; all other information about structural impulse responses comes from the prior. Furthermore, I prove that the approximation error incurred by using the Whittle likelihood is negligible asymptotically.

As a key step in the asymptotic analysis, I show that the posterior distribution for the autocovariance function of essentially any covariance stationary time series is consistent for the true value. While the posterior is computed under the working assumption that the data is Gaussian and q-dependent, consistency obtains under general misspecification of the Whittle likelihood. Existing time series results on posterior consistency assume well-specified likelihood functions. The only assumptions I place on the data generating process are nonparametric covariance stationarity and weak dependence conditions, and the prior is unrestricted except its support must contain the true autocovariance function.

To aid readers who are familiar with SVAR analysis, I demonstrate how to transparently impose standard SVAR identifying restrictions in the SVMA framework, if desired. The

[1] Consistent with Phillips (1989), I use the term "partially identified" in the sense that a nontrivial function of the parameter vector is point identified, but the full parameter vector is not.

SVMA approach can easily accommodate exclusion and sign restrictions on short- and long-run (i.e., cumulative) impulse responses. The prior information can be imposed dogmatically (i.e., with 100% certainty) or non-dogmatically. External instruments can be exploited in the SVMA framework, as in SVARs, by expanding the vector of observed time series.

Literature. The SVMA estimation approach in this paper is more flexible than previous attempts in the literature, and it appears to be the first method for conducting valid inference about possibly noninvertible IRFs. Hansen & Sargent (1981) and Ito & Quah (1989) estimate SVMA models without assuming invertibility by maximizing the Whittle likelihood, but the only prior information they consider is a class of exact restrictions implied by rational expectations. Barnichon & Matthes (2015) propose a Bayesian approach to inference in SVMA models, but they restrict attention to recursively identified models and they center the prior at SVAR-implied IRFs. None of these three papers develop valid procedures for doing inference on IRFs that may be partially identified and noninvertible.[2] Moreover, each of the three papers impose parametric structures on the IRFs, while I show how to maintain computational tractability with potentially unrestricted IRFs.

A few SVAR papers have attempted to exploit general types of prior information about IRFs, but these methods are less flexible than the SVMA approach. Furthermore, by assuming an underlying SVAR model, they automatically rule out noninvertible IRFs. Dwyer (1998) works with an inflexible trinomial prior on IRFs. Gordon & Boccanfuso (2001) translate a prior on IRFs into a "best-fitting" prior on SVAR parameters, but Kocięcki (2010) shows that their method neglects the Jacobian of the transformation. Kocięcki's fix requires the transformation to be one-to-one, which limits the ability to exploit prior information

[2]Standard errors in Hansen & Sargent (1981) are only valid when the prior restrictions point identify the IRFs. Barnichon & Matthes (2015) approximate the SVMA likelihood using an autoregressive formula that is explosive when the IRFs are noninvertible, causing serious numerical instability. Barnichon & Matthes focus on invertible IRFs and extend the model to allow for asymmetric and state-dependent effects of shocks.

about long-run responses, shapes, and smoothness. Baumeister & Hamilton (2015b), who improve on the method of Sims & Zha (1998), persuasively argue for an explicit Bayesian approach to imposing prior information on IRFs. Their Bayesian SVAR method allows for a fully flexible prior on impact impulse responses, but they assume invertibility, and their prior on longer-horizon impulse responses is implicit and chosen for computational convenience.

Outline. Section 1.2 reviews SVARs and then discusses the SVMA model, invertibility, identification, and prior elicitation. Section 1.3 outlines the posterior simulation method. Section 1.4 illustrates the SVMA approach through a small simulation study. Section 1.5 empirically estimates the role of technological news shocks in the U.S. business cycle. Section 1.6 derives the large-sample limit of the posterior distribution for a large class of partially identified models that includes the SVMA model. Section 1.7 shows that popular SVAR restrictions can be imposed in the SVMA framework. Section 1.8 suggests topics for future research. Applied readers may want to focus on Sections 1.2 to 1.5. Technical details are relegated to Appendix A.1; in particular, notation is defined in Appendix A.1.1. Proofs can be found in Appendix B.1.

1.2 Model, invertibility, and prior elicitation

In this section I describe the SVMA model and my method for imposing priors on IRFs. After reviewing the SVAR model and its shortcomings, I discuss the SVMA model, whose parameters are the IRFs of observed variables to unobserved shocks. Because the SVMA model does not restrict the IRFs to be invertible *a priori*, it can be applied to more empirical settings than the SVAR approach. The IRFs are under-identified, as they are in SVAR analysis. The lack of identification necessitates the use of prior information, which I impose by placing a prior distribution on the IRFs that lets researchers flexibly and transparently

exploit all types of prior information about IRFs.

1.2.1 SVARs and their shortcomings

I begin with a brief review of Structural Vector Autoregression (SVAR) estimation of impulse response functions. The parametrization of the SVAR model makes it difficult to transparently exploit certain types of valuable prior information about impulse responses. Moreover, SVARs are ill-suited for empirical applications in which the econometrician has less information than economic agents.

Modern dynamic macroeconomics is based on Frisch's (1933) *impulse-propagation* paradigm, which attaches primary importance to *impulse response functions* (IRFs). The economy is assumed to be driven by unpredictable shocks (impulses) whose effect on observable macro aggregates is known as the propagation mechanism. It has long been recognized that – in a linear setting – this paradigm is well captured by the Structural Vector Moving Average (SVMA) model (Hansen & Sargent, 1981; Watson, 1994, Sec. 4)

$$y_t = \Theta(L)\varepsilon_t, \quad \Theta(L) = \sum_{\ell=0}^{\infty} \Theta_\ell L^\ell, \tag{1.1}$$

where L denotes the lag operator, $y_t = (y_{1,t}, \ldots, y_{n,t})'$ is an n-dimensional vector of observed macro variables, and the structural shocks $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})'$ form a martingale difference sequence with $E(\varepsilon_t \varepsilon_t') = \text{diag}(\sigma)^2$, $\sigma = (\sigma_1, \ldots, \sigma_n)'$. Most linearized discrete-time macro models can be written in SVMA form. $\Theta_{ij,\ell}$, the (i,j) element of Θ_ℓ, is the *impulse response* of variable i to shock j at horizon ℓ after the shock's initial impact. The IRF $(\Theta_{ij,\ell})_{\ell \geq 0}$ is thus a key object of scientific interest in macroeconomics (Ramey, 2016).

For computational reasons, most researchers follow Sims (1980) and estimate structural

IRFs and shocks using a SVAR model

$$A(L)y_t = H\varepsilon_t, \quad A(L) = I_n - \sum_{\ell=1}^{m} A_\ell L^\ell, \tag{1.2}$$

where m is a finite lag length, and the matrices $A_1, \ldots, A_m$ and H are each $n \times n$. The SVAR and SVMA models are closely related: If the SVAR is stable – i.e., the polynomial $A(L)$ has a one-sided inverse – the SVAR model (1.2) implies that the data has an SVMA representation (1.1) with IRFs given by $\Theta(L) = \sum_{\ell=0}^{\infty} \Theta_\ell L^\ell = A(L)^{-1}H$. The SVAR model is computationally attractive because the parameters A_ℓ are regression coefficients.

The IRFs implied by the SVAR model are not identified from the data if the shocks are unobserved, as is usually the case.[3] While the VAR polynomial $A(L)$ can be recovered from a regression of y_t on its lags, the impact matrix H and shock standard deviations σ are not identified. Denote the reduced-form (Wold) forecast error by $u_{t|t-1} = y_t - \text{proj}(y_t \mid y_{t-1}, y_{t-2}, \ldots) = H\varepsilon_t$, where "proj" denotes population linear projection. Then the only information available from second moments of the data to identify H and σ is that $E(u_{t|t-1}u'_{t|t-1}) = H\,\text{diag}(\sigma)^2 H'$.[4] As knowledge of H is required to pin down the SVAR IRFs, the latter are under-identified. Thus, the goal of the SVAR literature is to exploit weak prior information about the model parameters to estimate unknown features of the IRFs.

One drawback of the SVAR model is that its parametrization makes it difficult to transparently exploit certain types of prior information. The IRFs $\Theta(L) = A(L)^{-1}H$ implied by the SVAR model are nonlinear functions of the parameters $(A(L), H)$, and impulse responses $\Theta_{ij,\ell}$ at long horizons ℓ are extrapolated from the short-run correlations of the data. Hence,

---

[3]If the structural shocks ε_t were known, the IRFs in the SVMA model (1.1) could be estimated by direct regressions of y_t on lags of ε_t (Jordà, 2005).

[4]Equivalently, if $E(u_{t|t-1}u'_{t|t-1}) = JJ'$ is the (identified) Cholesky decomposition of the forecast error covariance matrix, then all that the second moments of the data reveal about H and σ is that $H\,\text{diag}(\sigma) = JQ$ for some unknown $n \times n$ orthogonal matrix Q (Uhlig, 2005, Prop. A.1).

the overall shapes and smoothness of the model-implied IRFs depend indirectly on the SVAR parameters, which impedes the use of prior information about such features of the IRFs.[5] Instead, SVAR papers impose zero or sign restrictions on short- or long-run impulse responses to sharpen identification.[6] Because of the indirect parametrization of the IRFs, such SVAR identification schemes are known to impose additional unintended and unacknowledged prior information about IRFs.[7]

A second drawback of the SVAR model is the invertibility problem. The defining property of the SVAR model (1.2) is that the structural shocks $\varepsilon_t = (\varepsilon_{1,t}, \dots, \varepsilon_{n,t})'$ can be recovered linearly from the history $(y_t, y_{t-1}, \dots)$ of observed data, given knowledge of H and σ. This *invertibility* assumption – that the time-t shocks can be recovered from current and past, but not future, values of the observed data – is arbitrary and may be violated if the econometrician does not observe all variables relevant to the decisions of forward-looking economic agents. Indeed, the literature has demonstrated that data generated by interesting macroeconomic models, including models with news or noise shocks, cannot be represented as a SVAR for this reason. Section 1.2.3 discusses invertibility in greater detail and provides references.

To overcome the drawbacks of the SVAR model, I return to the basics and infer IRFs directly from the SVMA representation (1.1) of the data. The SVMA parameters are underidentified, so prior information must be imposed to learn about unknown features of the IRFs. Conveniently, the parameters of the SVMA model are the IRFs themselves, so all

---

[5]The shapes of the IRFs are governed by the magnitudes and imaginary parts of the roots of the VAR lag polynomial $A(L)$, and the roots are in turn complicated functions of the lag matrices $A_1, \dots, A_m$. See Geweke (1988, Sec. 2) for an illustration in the univariate case.

[6]Ramey (2016), Stock & Watson (2016), and Uhlig (2015) review SVAR identification schemes.

[7]See Arias, Rubio-Ramírez & Waggoner (2014, Sec. 5) and Baumeister & Hamilton (2015b, Sec. 3). For a trivial example, consider the AR(1) model $y_t = A_1 y_{t-1} + \varepsilon_t$ with $n = 1$ and $|A_1| < 1$. The IRF is $\Theta_\ell = A_1^\ell$, so imposing the sign restriction $\Theta_1 \geq 0$ implicitly also restricts $\Theta_\ell \geq 0$ for all $\ell \geq 2$.

types of prior information about IRFs can be imposed easily and transparently. Moreover, because the IRFs $\Theta(L)$ are unrestricted, the structural shocks ε_t need not be recoverable from only current and past values of the data. Hence, the SVMA model can handle applications in which the data may not have a SVAR representation, as in the examples described above.

1.2.2 SVMA model

I now discuss the SVMA model assumptions in detail and show that its parameters can be interpreted as IRFs. Then I illustrate the natural parametrization by example.

The SVMA model assumes the observed time series $y_t = (y_{1,t}, \ldots, y_{n,t})'$ are driven by current and lagged values of unobserved, unpredictable shocks $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})'$ (Hansen & Sargent, 1981). Although the shocks are unobserved, the researcher must have some degree of prior knowledge about their nature in order to estimate the SVMA parameters, as described in Section 1.2.5. For simplicity, I follow the SVAR literature in assuming that the number n of shocks is known and equals the number of observed series. However, most methods in this paper generalize to the case with more shocks than variables, cf. Section 1.8.

Assumption 1.1 (SVMA model)**.**

$$y_t = \Theta(L)\varepsilon_t,\ t \in \mathbb{Z}, \quad \Theta(L) = \sum_{\ell=0}^{q} \Theta_\ell L^\ell, \tag{1.3}$$

where L is the lag operator, q is the finite MA lag length, and $\Theta_0, \Theta_1, \ldots, \Theta_q$ are each $n \times n$ coefficient matrices. The shocks are serially and mutually unpredictable: For each t and j, $E(\varepsilon_{j,t} \mid \{\varepsilon_{k,t}\}_{k \neq j}, \{\varepsilon_s\}_{-\infty < s < t}) = 0$ and $E(\varepsilon_{j,t}^2) = \sigma_j^2$, where $\sigma_j > 0$.

For simplicity, I assume that the moving average (MA) lag length q is finite and known, but it is of course possible to estimate q using information criteria or Box-Jenkins methods. To fit persistent data q must be relatively large, which my computational strategy in Section 1.3 is well-suited for. The assumption that y_t has mean zero is made for notational

convenience and can easily be relaxed. Unlike in reduced-form Vector Autoregressive Moving Average (VARMA) modeling, the SVMA model allows $\Theta_0 \neq I_n$.

The SVMA and SVAR models are related but not equivalent. If the matrix lag polynomial $\Theta(L)$ has a one-sided inverse $D(L) = \sum_{\ell=0}^{\infty} D_\ell L^\ell = \Theta(L)^{-1}$, the SVMA structure (1.3) is compatible with an underlying SVAR model $D(L)y_t = \varepsilon_t$ (with lag length $m = \infty$). However, the fact that I do not constrain $\Theta(L)$ to have a one-sided inverse is key to allowing for noninvertible IRFs, as explained in Section 1.2.3. Assumption 1.1 imposes stationary, linear dynamics with time-invariant parameters, which is restrictive but standard in the SVAR literature.[8] The condition that ε_t form a martingale difference sequence with mutually unpredictable components is also standard and operationalizes the interpretation of ε_t as a vector of conceptually independent structural shocks.[9]

Unlike in SVARs, the parameters of the SVMA model have direct economic interpretations as impulse responses. Denote the (i, j) element of matrix Θ_ℓ by $\Theta_{ij,\ell}$. The index ℓ will be referred to as the horizon. For each $j \in \{1, \ldots, n\}$, choose an $i_j \in \{1, \ldots, n\}$ and normalize the impact response of variable i_j to shock j: $\Theta_{i_j j,0} = 1$. Then the parameter $\Theta_{ij,\ell}$ is the expected response at horizon ℓ of variable i to shock j, where the size of the shock is of a magnitude that raises variable i_j by one unit on impact:[10]

$$\Theta_{ij,\ell} = E(y_{i,t+\ell} \mid \varepsilon_{j,t} = 1) - E(y_{i,t+\ell} \mid \varepsilon_{j,t} = 0). \tag{1.4}$$

The *impulse response function* (IRF) of variable i to shock j is the $(q+1)$-dimensional vector $(\Theta_{ij,0}, \Theta_{ij,1}, \ldots, \Theta_{ij,q})'$. In addition to the impulse response parameters $\Theta_{ij,\ell}$, the model

[8] I briefly discuss nonstationarity, nonlinearity, and time-varying parameters in Section 1.8.

[9] See Leeper, Sims & Zha (1996, pp. 6–15) and Sims & Zha (2006, p. 252). They emphasize that the assumption of mutually unpredictable shocks deliberately departs from standard practice in classical linear simultaneous equation models due to the different interpretation of the error terms.

[10] Henceforth, moments of the data and shocks are implicitly conditioned on the parameters (Θ, σ).

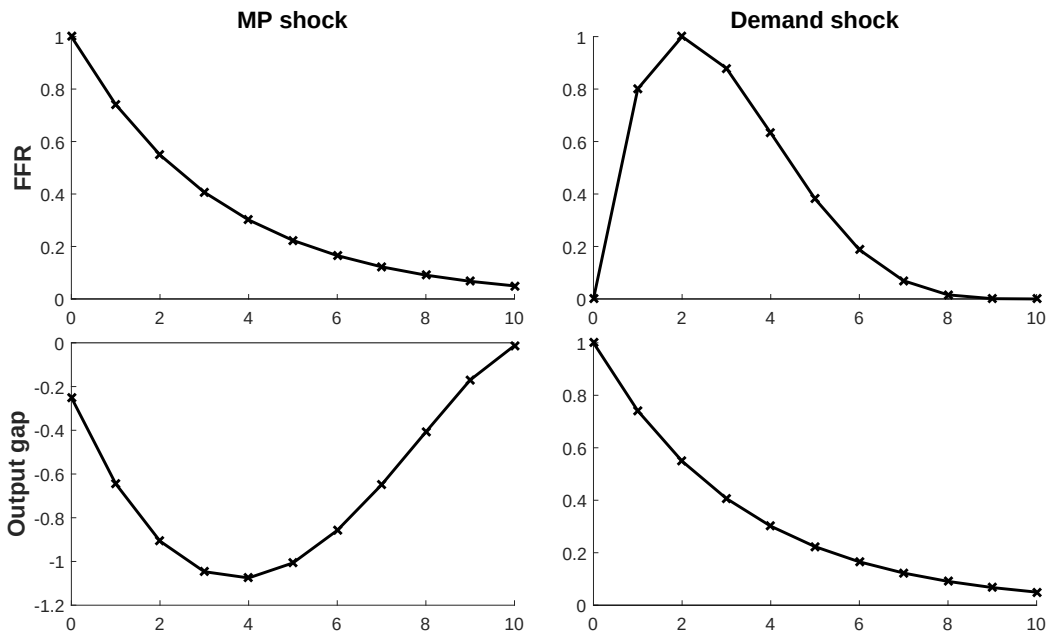

Figure 1.1: Hypothetical IRFs of two observed variables (along rows) to two unobserved shocks (along columns). The upper right display, say, shows the IRF of the FFR to the demand shock. The horizontal axes represent the impulse response horizon $\ell = 0, 1, \ldots, q$, where $q = 10$. IRFs in the left column are normalized so a positive monetary policy (MP) shock yields a 100 basis point increase in the FFR on impact; IRFs in the right column are normalized so a positive demand shock yields a 1 percentage point increase in the output gap on impact.

contains the shock standard deviation parameters σ_j, which govern the overall magnitudes of the responses to one-standard-deviation impulses to $\varepsilon_{j,t}$.

The parameters are best understood through an example. Figure 1.1 plots a hypothetical set of impulse responses for a bivariate application with two observed time series, the federal funds rate (FFR) $y_{1,t}$ and the output gap $y_{2,t}$, and two unobserved shocks, a monetary policy shock $\varepsilon_{1,t}$ and a demand shock $\varepsilon_{2,t}$. The figure imposes the normalizations $i_1 = 1$ and $i_2 = 2$, so that $\Theta_{21,3}$, say, is the horizon-3 impulse response of the output gap to a monetary policy shock that raises the FFR by 1 unit (100 basis points) on impact. As the figure shows, the impulse response parameters $\Theta_{ij,\ell}$ can be visualized jointly in a format that is familiar from theoretical macro modeling. Each impulse response (the crosses in the figure) corresponds to a distinct IRF parameter $\Theta_{ij,\ell}$. In contrast, the parameters in the SVAR model are only

indirectly related to IRFs and do not carry graphical intuition in and of themselves. The natural and flexible parametrization of the SVMA model facilitates the incorporation of prior information about IRFs, as described below.

Because I wish to estimate the IRFs using parametric Bayesian methods, it is necessary to strengthen Assumption 1.1 by assuming a specific distribution for the shocks ε_t. For concreteness I impose the working assumption that they are i.i.d. Gaussian.

Assumption 1.2 (Gaussian shocks)**.**

$$\varepsilon_t \overset{i.i.d.}{\sim} N(0, \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)),\ t \in \mathbb{Z}. \tag{1.5}$$

The Gaussianity assumption places the focus on the unconditional second-order properties of the data y_t, as is standard in the SVAR literature, but the assumption is not central to my analysis. Section 1.6 shows that if the Bayesian posterior distribution for the IRFs is computed under Assumption 1.2 and a non-dogmatic prior distribution, the large-sample limit of the posterior is robust to violations of the Gaussianity assumption. Moreover, the method for sampling from the posterior in Section 1.3 is readily adapted to non-Gaussian and/or heteroskedastic likelihoods, as discussed in Section 1.8.

1.2.3 Invertibility

One advantage of the SVMA model is that it allows for noninvertible IRFs. These can arise in applications in which the econometrician does not observe all variables in economic agents' information sets. Here I review the prevalence of noninvertible IRFs in macroeconomics and the SVAR model's inability to consistently estimate such IRFs. Because the SVMA model does not restrict IRFs to be invertible *a priori*, it is applicable to a broader set of empirical settings than the SVAR model.

The IRF parameters are *invertible* if the current shock ε_t can be recovered as a linear

function of current and past – but not future – values $(y_t, y_{t-1}, \dots)$ of the observed data, given knowledge of the parameters.[11] In this sense, *non*invertibility is caused by economically important variables being omitted from the econometrician's information set.[12] Invertibility is a property of the collection of n^2 IRFs, and an invertible collection of IRFs can be rendered noninvertible by removing or adding observed variables or shocks. See Hansen & Sargent (1981, 1991) and Lippi & Reichlin (1994) for extensive mathematical discussions of invertibility in SVMAs and SVARs.

Invertibility is not a compelling *a priori* restriction when estimating structural IRFs, for two reasons. First, the definition of invertibility is statistically motivated and has little economic content. For example, the reasonable-looking IRFs in Figure 1.1 happen to be noninvertible, but minor changes to the lower left IRF in the figure render the IRFs invertible. Second, interesting macro models generate noninvertible IRFs, such as models with news shocks or noisy signals.[13] Intuitively, upon receiving a signal about changes in policy or economic fundamentals that will occur sufficiently far into the future, economic agents change their current behavior much less than their future behavior. Thus, future – in addition to current and past – data is needed to distinguish the signal from other concurrent shocks.

By their very definition, SVARs implicitly restrict IRFs to be invertible, as discussed in Section 1.2.1. No SVAR identification strategy can therefore consistently estimate noninvertible IRFs. This fact has spawned an extensive literature trying to salvage the SVAR

---

[11] Precisely, the IRFs are invertible if ε_t lies in the closed linear span of $(y_t, y_{t-1}, \dots)$. Invertible MA representations are also referred to as "fundamental" in the literature.

[12] See Hansen & Sargent (1991), Sims & Zha (2006), Fernández-Villaverde, Rubio-Ramírez, Sargent & Watson (2007), Forni, Giannone, Lippi & Reichlin (2009), Leeper, Walker & Yang (2013), Forni, Gambetti & Sala (2014), and Lütkepohl (2014).

[13] See Alessi, Barigozzi & Capasso (2011, Sec. 4–6), Blanchard, L'Huillier & Lorenzoni (2013, Sec. II), Leeper et al. (2013, Sec. 2), and Beaudry & Portier (2014, Sec. 3.2).

approach. Some papers assume additional model structure,[14] while others rely on the availability of proxy variables for the shocks, thus ameliorating the invertibility issue.[15] These methods only produce reliable results under additional assumptions or if the requisite data is available, whereas the SVMA approach always yields correct inference about IRFs regardless of invertibility. If available, proxy variables can be incorporated in SVMA analysis to improve identification.

The SVMA model (1.3) is parametrized directly in terms of IRFs and does not impose invertibility *a priori* (Hansen & Sargent, 1981). In fact, the IRFs are invertible if and only if the polynomial $z \mapsto \det(\Theta(z))$ has no roots inside the unit circle.[16] In general, the structural shocks can be recovered from past, current, and *future* values of the observed data:[17]

$$\varepsilon_t = D(L)y_t, \quad D(L) = \sum_{\ell=-\infty}^{\infty} D_\ell L^\ell = \Theta(L)^{-1}.$$

Under Assumption 1.1, the structural shocks can thus be recovered from *multi-step* forecast errors: $\varepsilon_t = \sum_{\ell=0}^{\infty} D_\ell u_{t+\ell|t-1}$, where $u_{t+\ell|t-1} = y_{t+\ell} - \text{proj}(y_{t+\ell} \mid y_{t-1}, y_{t-2}, \dots)$ is the econometrician's $(\ell+1)$-step error. Only if the IRFs are invertible do we have $D_\ell = 0$ for $\ell \geq 1$, in which case ε_t is a linear function of the *one-step* (Wold) error $u_{t|t-1}$, as SVARs assume.

---

[14] Lippi & Reichlin (1994) and Klaeffing (2003) characterize the range of noninvertible IRFs consistent with a given estimated SVAR, while Mertens & Ravn (2010) and Forni, Gambetti, Lippi & Sala (2013) select a single such IRF using additional model restrictions. Lanne & Saikkonen (2013) develop asymptotic theory for a modified VAR model that allows for noninvertibility, but they do not consider structural estimation.

[15] Sims & Zha (2006), Fève & Jidoud (2012), Sims (2012), Beaudry & Portier (2014, Sec. 3.2), and Beaudry, Fève, Guay & Portier (2015) argue that noninvertibility need not cause large biases in SVAR estimation, especially if forward-looking variables are available. Forni et al. (2009) and Forni et al. (2014) use information from large panel data sets to ameliorate the omitted variables problem; based on the same idea, Giannone & Reichlin (2006) and Forni & Gambetti (2014) propose tests of invertibility.

[16] That is, if and only if $\Theta(L)^{-1}$ is a one-sided lag polynomial, so that the SVAR representation $\Theta(L)^{-1}y_t = \varepsilon_t$ obtains (Brockwell & Davis, 1991, Thm. 11.3.2 and Remark 1, p. 128).

[17] See for example Brockwell & Davis (1991, Thm. 3.1.3) and Lippi & Reichlin (1994, p. 312). The matrix lag polynomial $D(L) = \Theta(L)^{-1}$ is not well-defined in the knife-edge case $\det(\Theta(1)) = \det(\sum_{\ell=0}^{q} \Theta_\ell) = 0$.

For illustration, consider a univariate SVMA model with $n = q = 1$:

$$y_t = \varepsilon_t + \Theta_1 \varepsilon_{t-1}, \quad \Theta_1 \in \mathbb{R}, \quad E(\varepsilon_t^2) = \sigma^2. \tag{1.6}$$

If $|\Theta_1| \leq 1$, the IRF $\Theta = (1, \Theta_1)$ is invertible: The shock has the SVAR representation $\varepsilon_t = \sum_{\ell=0}^{\infty} (-\Theta_1)^\ell y_{t-\ell}$, so it can be recovered using current and past values of the observed data. On the other hand, if $|\Theta_1| > 1$, no SVAR representation for ε_t exists: $\varepsilon_t = -\sum_{\ell=1}^{\infty} (-\Theta_1)^{-\ell} y_{t+\ell}$, so *future* values of the data are required to recover the current structural shock. Clearly, the latter possibility is fully consistent with the SVMA model (1.6) but inconsistent with any SVAR model of the form (1.2).[18]

Bayesian analysis of the SVMA model can be carried out without reference to the invertibility of the IRFs. The formula for the Gaussian SVMA likelihood function is the same in either case, cf. Appendix A.1.3.1 and Hansen & Sargent (1981). Moreover, standard state-space methods can always be used to estimate the structural shocks, as demonstrated in Section 1.5. This contrasts sharply with SVAR analysis, where special tools are needed to handle noninvertible specifications. Since invertibility is a rather arcane issue without much economic content, it is helpful that the SVMA model allows the researcher to focus on matters that do have economic significance.

1.2.4 Identification

As in SVAR analysis, the IRFs in the SVMA model are only partially identified. The lack of identification arises because the model treats all shocks symmetrically and because noninvertible IRFs are not ruled out *a priori*.

---

[18] If $|\Theta_1| > 1$, an SVAR (with $m = \infty$) applied to the time series (1.6) estimates the incorrect invertible IRF $(1, 1/\Theta_1)$ and (Wold) "shock" $u_{t|t-1} = \varepsilon_t + (1 - \Theta_1^2) \sum_{\ell=1}^{\infty} (-\Theta_1)^{-\ell} \varepsilon_{t-\ell}$. This is because the SVMA parameters (Θ_1, σ) and $(1/\Theta_1, \sigma\Theta_1)$ are observationally equivalent, cf. Section 1.2.4.

Because of the linearity of the SVMA model and the assumption of Gaussian shocks, any two IRFs that give rise to the same autocovariance function (ACF) are observationally equivalent. Under Assumption 1.1, the matrix ACF of the time series $\{y_t\}$ is given by

$$\Gamma(k) = E(y_{t+k}y_t') = \begin{cases} \sum_{\ell=0}^{q-k} \Theta_{\ell+k} \operatorname{diag}(\sigma)^2 \Theta_\ell' & \text{if } 0 \leq k \leq q, \\ 0 & \text{if } k > q. \end{cases} \tag{1.7}$$

Under Assumptions 1.1 and 1.2, the observed vector time series y_t is a mean-zero strictly stationary Gaussian process, so the distribution of the data is completely characterized by the ACF $\Gamma(\cdot)$. The identified set $\mathcal{S}$ for the IRF parameters $\Theta = (\Theta_0, \Theta_1, \ldots, \Theta_q)$ and shock standard deviation parameters $\sigma = (\sigma_1, \ldots, \sigma_n)'$ is thus a function of the ACF:

$$\mathcal{S}(\Gamma) = \left\{ (\tilde{\Theta}_0, \ldots, \tilde{\Theta}_q) \in \Xi_\Theta, \tilde{\sigma} \in \Xi_\sigma \colon \sum_{\ell=0}^{q-k} \tilde{\Theta}_{\ell+k} \operatorname{diag}(\tilde{\sigma})^2 \tilde{\Theta}_\ell' = \Gamma(k), 0 \leq k \leq q \right\},$$

where $\Xi_\Theta = \{(\tilde{\Theta}_0, \ldots, \tilde{\Theta}_q) \in \mathbb{R}^{n \times n(q+1)} \colon \tilde{\Theta}_{ijj,0} = 1, 1 \leq j \leq n\}$ is the parameter space for Θ, and $\Xi_\sigma = \{(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_n)' \in \mathbb{R}^n \colon \tilde{\sigma}_j > 0, 1 \leq j \leq n\}$ is the parameter space for σ.[19] By definition, two parameter configurations contained in the same identified set give rise to the same value of the SVMA likelihood function under Gaussian shocks.

The identified set for the SVMA parameters is large in economic terms. Building on Hansen & Sargent (1981) and Lippi & Reichlin (1994), Appendix A.1.2 provides a constructive characterization of $\mathcal{S}(\Gamma)$. I summarize the main insights here.[20] The identified set contains uncountably many parameter configurations if the number n of shocks exceeds 1. The lack of identification is not just a technical curiosity but is of primary importance to

[19] If the shocks ε_t were known to have a non-Gaussian distribution, the identified set would change due to the additional information provided by higher-order moments of the data, cf. Section 1.6.3.

[20] The identification problem is not easily cast in the framework of interval identification, as $\mathcal{S}(\Gamma)$ is of strictly lower dimension than the parameter space $\Xi_\Theta \times \Xi_\sigma$. Still, expression (1.7) for $\operatorname{diag}(\Gamma(0))$ implies that the identified set for scaled impulse responses $\Psi_{ij,\ell} = \Theta_{ij,\ell}\sigma_j$ is bounded.

economic conclusions. For example, as in SVARs, for any observed ACF $\Gamma(\cdot)$, any horizon ℓ, any shock j, and any variable $i \neq i_j$, there exist IRFs in the identified set $\mathcal{S}(\Gamma)$ with $\Theta_{ij,\ell} = 0$.

One reason for under-identification, also present in SVARs (cf. Section 1.2.1), is that the assumptions so far treat the n shocks symmetrically: Without further restrictions, the model and data offer no way of distinguishing the first shock from the second shock, say, and consequently no way of separately identifying the IRFs to the first and second shocks. Mathematically, the two parameter configurations (Θ, σ) and $(\tilde{\Theta}, \tilde{\sigma})$ lie in the same identified set if there exists an orthogonal $n \times n$ matrix Q such that $\tilde{\Theta}\,\mathrm{diag}(\tilde{\sigma})Q = \Theta\,\mathrm{diag}(\sigma)$.

The second source of under-identification is that the SVMA model, unlike SVARs, does not arbitrarily restrict the IRFs to be invertible. For any noninvertible set of IRFs there always exists an observationally equivalent invertible set of IRFs (if $n > 1$, there exist several). If $nq > 1$, there are also several other observationally equivalent noninvertible IRFs. This identification issue arises even if, say, we impose exclusion restrictions on the elements of Θ_0 to exactly identify the correct orthogonal matrix Q in the previous paragraph.

Figure 1.2 illustrates the identification problem due to noninvertibility for a univariate model with $n = 1$ and $q = 4$: $y_t = \varepsilon_t + \sum_{\ell=1}^{4} \Theta_\ell \varepsilon_{t-\ell}$, $\Theta_\ell \in \mathbb{R}$, $E(\varepsilon_t^2) = \sigma^2$. The ACF in the left panel of the figure is consistent with the four IRFs shown in the right panel. The invertible IRF (drawn with a thick line) is the one that would be estimated by a SVAR (with lag length $m = \infty$). However, there exist three other IRFs that have very different economic implications but are equally consistent with the observed ACF.[21] If $n > 1$, the identification problem is even more severe, as described in Appendix A.1.2.

As the data alone does not suffice to distinguish between IRFs that have very different

[21] Similarly, in the special case $n = q = 1$, the parameters (Θ_1, σ) imply the same ACF as the parameters $(\tilde{\Theta}_1, \tilde{\sigma})$, where $\tilde{\Theta}_1 = 1/\Theta_1$ and $\tilde{\sigma} = \sigma\Theta_1$. If $|\Theta_1| \leq 1$, an SVAR would estimate the invertible IRF $(1, \Theta_1)$ for which most of the variation in y_t is due to the current shock ε_t. But the data would be equally consistent with the noninvertible IRF $(1, \tilde{\Theta}_1)$ for which y_t is mostly driven by the previous shock ε_{t-1}.

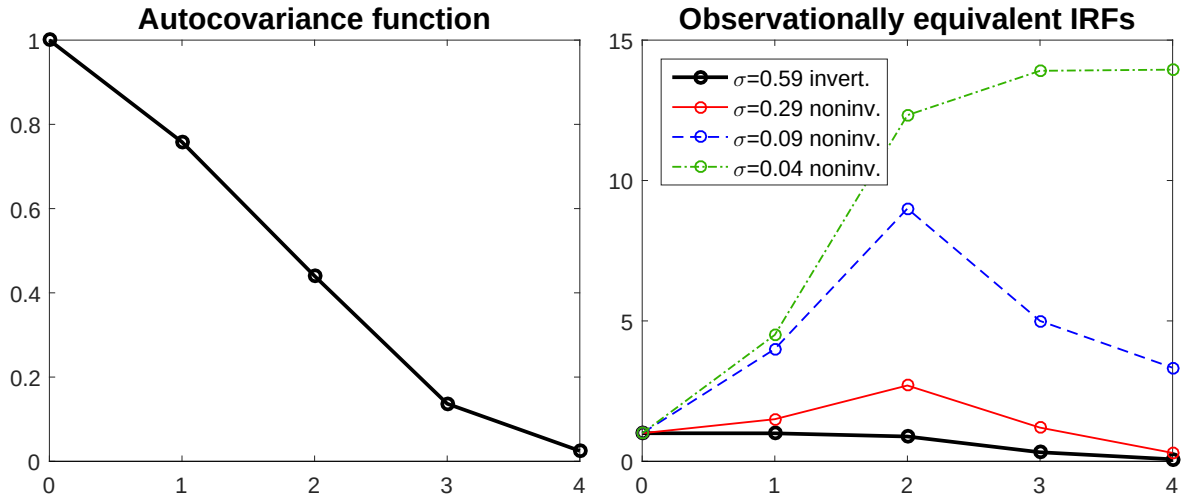

Figure 1.2: Example of IRFs that generate the same ACF, based on a univariate SVMA model with $n = 1$ and $q = 4$. The right panel shows the four IRFs that generate the particular ACF in the left panel; associated shock standard deviations are shown in the figure legend.

economic implications, it is necessary to leverage additional prior information.[22] In SVAR analysis the prior information is often referred to as the identification scheme, cf. Section 1.7. The next subsection describes the flexible and transparent approach to prior specification I adopt for the SVMA model.

1.2.5 Prior specification and elicitation

In addition to handling noninvertible IRFs, the other key advantage of the SVMA model is its natural parametrization, which allows prior information to be imposed directly on the IRFs through a transparent and flexible Bayesian prior distribution. Researchers often have access to more prior information about IRFs than what SVAR methods exploit. I explain how such information helps distinguish between observationally equivalent IRFs. Then I propose a prior elicitation procedure that imposes all types of prior information about IRFs

[22]Kline & Tamer (2015) develop methods for conducting Bayesian inference about the identified set in general models. Unfortunately, as argued above, hypotheses that only concern the identified set $\mathcal{S}(\Gamma)$ are rarely interesting in the context of estimating structural impulse responses $\Theta_{ij,\ell}$ because such hypotheses must treat all types of shocks symmetrically.

in a unified way. I highlight a Gaussian prior family that is convenient to visualize, but as Gaussianity is not essential for my approach, I discuss other choices of priors as well.

To impose prior information, the researcher must have some knowledge about the identity and effects of the unobserved shocks. As in the SVAR approach, the researcher postulates that, say, the first shock $\varepsilon_{1,t}$ is a monetary policy shock, the second shock $\varepsilon_{2,t}$ is a demand shock, etc.[23] Then prior information about the effects of the shocks, i.e., the IRFs, must be imposed. Prior information can be imposed dogmatically (with 100% certainty, as is common in SVAR analysis) or non-dogmatically (with less than 100% certainty).

Types and sources of prior information. Because the SVMA model is parametrized in terms of IRFs, it is possible to exploit many types of prior information. Researchers often have fairly weak – but not agnostic – prior information about *magnitudes* of certain impulse responses. For example, the impact response of the output gap to a monetary policy shock that lowers the FFR by 100 basis points is unlikely to exceed 2 percent. Researchers typically have more informative priors about the *signs* of certain impulse responses, e.g., the impact response of the output gap to a monetary policy shock that raises the federal funds rate. Researchers may also have quite informative beliefs about the *shapes* of IRFs, e.g., whether they are likely to be monotonic or hump-shaped (i.e., the effect gradually builds up and then peters out). Finally, researchers often have strong beliefs about the *smoothness* of IRFs, due to adjustment costs, time to build, and information frictions.

Prior information may arise from several sources, all of which can be integrated in the graphical prior elicitation procedure introduced below. First, researchers may be guided by structural macroeconomic models whose deep parameters have been calibrated to microeconometric data. Parameter and model uncertainty forbid treating model-implied IRFs as

[23]The order of the shocks is immaterial.

truth, but these may nevertheless be judged to be *a priori* likely, as in the empirical application in Section 1.5. Second, economic intuition and stylized models yield insight into the likely signs, shapes, and smoothness of the IRFs. Third, microeconometric evidence or macroeconometric studies on related datasets may provide relevant information.

Bayesian approach. Bayesian inference is a unified way to exploit all types of prior information about the IRFs Θ. In this approach an informative joint prior distribution is placed on the SVMA model parameters, i.e., the IRFs Θ and shock standard deviations σ.[24] Since there is no known flexible conjugate prior for MA models, I place a flexible multivariate prior distribution on the IRFs and shock standard deviations. The generality of this approach necessitates the use of simulation methods for conducting posterior inference about the structural parameters. The simulation method I propose in Section 1.3 works with any prior distribution for which the log density and its gradient can be computed, giving the researcher great flexibility.

The information in the prior and the data is synthesized in the posterior density, which is proportional to the product of the prior density and the likelihood function. As discussed in Section 1.2.4, the likelihood function does not have a unique maximum due to partial identification. The role of the prior is to attach weights to parameter values that are observationally equivalent based on the data but distinguishable based on prior information, as sketched in Figure 1.3.[25] The SVMA analysis thus depends crucially on the prior information imposed, just as SVAR analysis depends on the identification scheme. The frequentist asymptotics in

[24]Alternatively, one could specify a prior on the ACF Γ and a conditional prior for (Θ, σ) given Γ. This approach has the conceptual advantage that the data asymptotically dominates the prior for Γ but does not provide information about (Θ, σ) given Γ (cf. Section 1.6). However, in applications, prior information typically directly concerns the IRFs Θ, and it is unclear how to select a meaningful prior for Θ given Γ.

[25]From a subjectivist Bayesian perspective, as long as the prior is a proper probability distribution, the validity of posterior inference is unaffected by the under-identification of the parameters (Poirier, 1998).

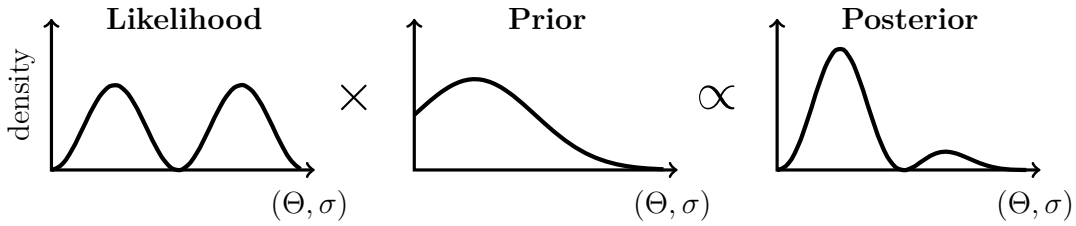

Figure 1.3: Conceptual illustration of how the likelihood function and the prior density combine to yield the posterior density. Even though the likelihood has multiple peaks of equal height, the posterior may be almost unimodal, depending on the strength of prior information.

Section 1.6 show formally that only some features of the prior information can be updated and falsified by the data. This is unavoidable due to the lack of identification, but it does underscore the need for a transparent and flexible prior elicitation procedure.

Gaussian prior. While many priors are possible, I first discuss an especially convenient multivariate Gaussian prior distribution. The assumption of Gaussianity means that the prior hyperparameters are easily visualized, as illustrated by example below. However, I stress that neither the overall SVMA approach nor the numerical methods in this paper rely on Gaussianity of the prior. I describe other possible prior choices below.

The multivariate Gaussian joint prior distribution on the impulse responses is given by

$$\Theta_{ij,\ell} \sim N(\mu_{ij,\ell}, \tau^2_{ij,\ell}), \quad 0 \leq \ell \leq q,$$

$$\text{Corr}(\Theta_{ij,\ell+k}, \Theta_{ij,\ell}) = \rho^k_{ij}, \quad 0 \leq \ell \leq \ell + k \leq q, \tag{1.8}$$

for each (i, j). This correlation structure means that the prior smoothness of IRF (i, j) is governed by ρ_{ij}, as illustrated below.[26] For simplicity, the IRFs $(\Theta_{ij,0}, \Theta_{ij,1}, \ldots, \Theta_{ij,q})$ are

[26]The prior has the equivalent autoregressive representation $(\Theta_{ij,\ell+1} - \mu_{ij,\ell+1})/\tau_{ij,\ell+1} = \rho_{ij}(\Theta_{ij,\ell} - \mu_{ij,\ell})/\tau_{ij,\ell} + (1 - \rho^2_{ij})\zeta_{ij,\ell+1}$, where $\zeta_{ij,\ell}$ is i.i.d. $N(0, 1)$. That is, if the true impulse response at horizon ℓ is above its prior mean, then we also find it likely that the true impulse response at horizon $\ell + 1$ is above its prior mean, and more likely the higher ρ_{ij} is.

a priori independent across (i, j) pairs. The normalized impulse responses have $\mu_{ijj,0} = 1$ and $\tau_{ijj,0} = 0$ for each j. The shock standard deviations $\sigma_1, \ldots, \sigma_n$ are *a priori* mutually independent and independent of the IRFs, with prior marginal distribution

$$\log \sigma_j \sim N(\mu_j^\sigma, (\tau_j^\sigma)^2)$$

for each j.[27] In practice, the prior variances $(\tau_j^\sigma)^2$ for the log shock standard deviations can be chosen to be a large number. Because the elements of σ scale the ACF, which is identified, the data will typically be quite informative about the standard deviations of the shocks, provided that the prior on the IRFs is sufficiently informative.

The key hyperparameters in this Gaussian prior are the prior means $\mu_{ij,\ell}$ and variances $\tau_{ij,\ell}^2$ of each impulse response, and the prior smoothness hyperparameter ρ_{ij} for each IRF. The prior means and variances can be elicited graphically by drawing a figure with a "best guess" for each IRF and then placing a 90% (say) prior confidence band around each IRF. Once these hyperparameters have been elicited, the prior smoothness ρ_{ij} of each IRF can be elicited by trial-and-error simulation from the multivariate Gaussian prior.[28]

The prior elicitation process is illustrated in Figures 1.4 and 1.5, which continue the bivariate example from Figure 1.1. The figures show a choice of prior means and 90% prior confidence bands for each of the impulse responses, directly implying suitable values for the $\mu_{ij,\ell}$ and $\tau_{ij,\ell}^2$ hyperparameters.[29] The prior distributions in the figures embed many different

---

[27] Alternatively, the prior on σ could be derived from a prior on the forecast error variance decomposition, cf. definition (1.9) in Section 1.5. I leave this possibility to future research.

[28] In principle, the ρ_{ij} values could be chosen by an empirical Bayes method. For each possible choice of ρ_{ij}, one could compute the marginal likelihood of the data (Chib, 2001, Sec. 10.2) and select the value of ρ_{ij} that maximizes the marginal likelihood. I leave this possibility to future research.

[29] The prior confidence bands in Figures 1.4 and 1.5 are pointwise bands that consider each horizon separately. This is the most common way to express uncertainty about impulse responses. Sims & Zha (1999, Sec. 6) recommend quantifying uncertainty about entire impulse response *functions*, i.e., uniform bands.

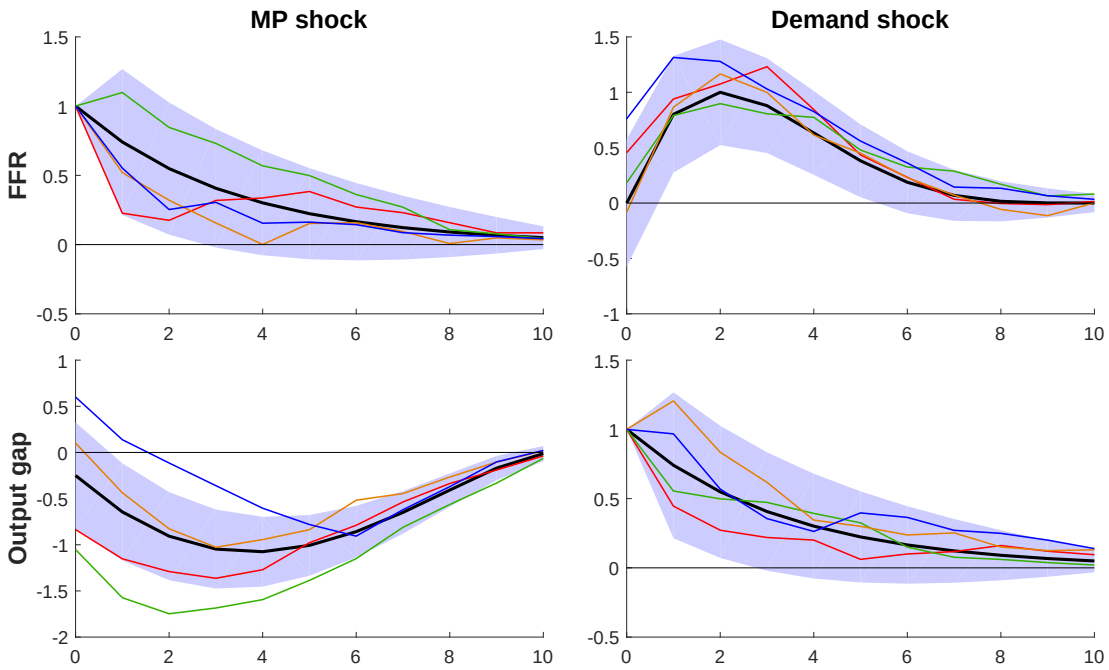

Figure 1.4: A choice of prior means (thick lines) and 90% prior confidence bands (shaded) for the four IRFs (Θ) in the bivariate example in Figure 1.1. Brightly colored lines are four draws from the multivariate Gaussian prior distribution with these mean and variance parameters and a smoothness hyperparameter of $\rho_{ij} = 0.9$ for all (i, j).

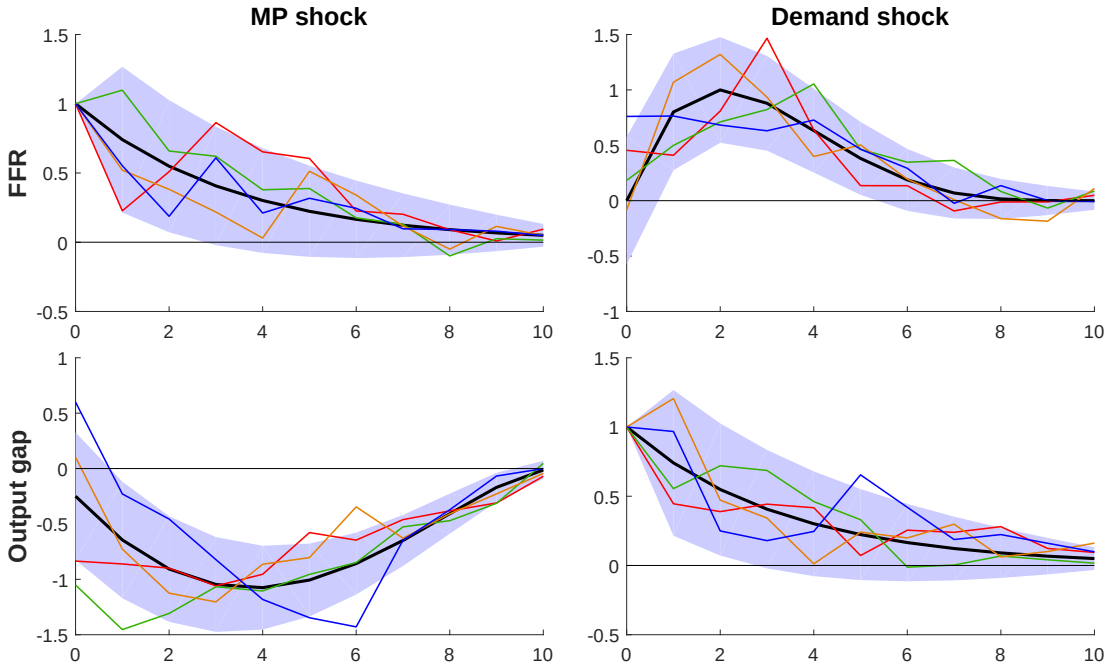

Figure 1.5: See caption for Figure 1.4. Here the smoothness parameter is $\rho_{ij} = 0.3$ for all (i, j).

kinds of prior information. For example, the IRF of the FFR to a positive demand shock is believed to be hump-shaped with high probability, and the IRF of the output gap to a contractionary monetary policy shock is believed to be negative at horizons 2–8 with high probability. Yet the prior expresses substantial uncertainty about features such as the sign and magnitude of the impact response of the output gap to a monetary policy shock.

After having elicited the prior means and variances, the smoothness hyperparameters can be chosen by trial-and-error simulations. Figure 1.4 also depicts four IRF draws from the multivariate Gaussian prior distribution with $\rho_{ij} = 0.9$ for all (i, j), while Figure 1.5 shows four draws with $\rho_{ij} = 0.3$. The latter draws are more jagged and erratic than the former draws, and many economists would agree that the jaggedness of the $\rho_{ij} = 0.9$ draws are more in line with their prior information about the smoothness of the true IRFs in this application.

The flexible and graphical SVMA prior elicitation procedure contrasts with prior specification in standard SVARs. As discussed in Sections 1.2.1 and 1.7, SVAR analyses exploit zero or sign restrictions on individual impulse responses or linear combinations thereof, while information about the shapes and smoothness of IRFs is neglected. Furthermore, prior restrictions on short- or long-run responses implicitly restrict other features of the IRFs, since the VAR model structure subtly constrains the possible shapes of the IRFs.

Bayesian analysis in the SVMA model is explicit about the prior restrictions on IRFs, and researchers can draw on standard Bayesian tools for conducting sensitivity analysis and model validation. The entire set of prior beliefs about IRFs is easily expressed graphically, unlike in SVAR analysis. The sensitivity of posterior inference with respect to features of the prior can be assessed using tools from the comprehensive Bayesian literature (Lopes & Tobias, 2011; Müller, 2012). Model validation and comparison can be carried out through the flexible framework of prior and posterior predictive checks and computation of Bayes

factors.[30] I give examples of prior sensitivity and model validation checks in the empirical application in Section 1.5. In contrast, robustness checks in SVAR analyses are typically limited to considering a small set of alternative identifying restrictions.

Other priors. The multivariate Gaussian prior distribution is flexible and easy to visualize but other prior choices are feasible as well. My inference procedure does not rely on Gaussianity of the prior, as the simulation method in Section 1.3 only requires that the log prior density and its gradient are computable. Hence, it is straight-forward to impose a different prior correlation structure than (1.8), or to impose heavy-tailed or asymmetric prior distributions on certain impulse responses. Section 1.7 gives examples of priors that transparently impose well-known identifying restrictions from the SVAR literature.

1.3 Bayesian computation

In this section I develop an algorithm to simulate from the posterior distribution of the IRFs. Because of the flexible and high-dimensional prior distribution placed on the IRFs, standard Markov Chain Monte Carlo (MCMC) methods are very cumbersome.[31] I employ a Hamiltonian Monte Carlo algorithm that uses the Whittle (1953) likelihood approximation to speed up computations. The algorithm is fast, asymptotically efficient, and easy to apply, and it allows for both invertible and noninvertible IRFs. If desired, a reweighting step can undo the Whittle approximation at the end.

I first define the posterior density of the structural parameters. Let T be the sample

[30] See Chib (2001, Ch. 10), Geweke (2010, Ch. 2), and Gelman et al. (2013, Ch. 6).

[31] Chib & Greenberg (1994) estimate univariate reduced-form Autoregressive Moving Average models by MCMC, but their algorithm is only effective in low-dimensional problems. Chan, Eisenstat & Koop (2015, see also references therein) perform Bayesian inference in possibly high-dimensional reduced-form VARMA models, but they impose statistical parameter normalizations that preclude structural estimation of IRFs.

size and $Y_T = (y_1', y_2', \ldots, y_T')'$ the data vector. Denote the prior density for the SVMA parameters by $\pi_{\Theta,\sigma}(\Theta, \sigma)$. The likelihood function of the SVMA model (1.3) depends on the parameters (Θ, σ) only through the scaled impulse responses $\Psi = (\Psi_0, \Psi_1, \ldots, \Psi_q)$, where $\Psi_\ell = \Theta_\ell \operatorname{diag}(\sigma)$ for $\ell = 0, 1, \ldots, q$. Let $p_{Y|\Psi}(Y_T \mid \Psi(\Theta, \sigma))$ denote the likelihood function, where the notation indicates that Ψ is a function of (Θ, σ). The posterior density is then

$$p_{\Theta,\sigma|Y}(\Theta, \sigma \mid Y_T) \propto p_{Y|\Psi}(Y_T \mid \Psi(\Theta, \sigma))\pi_{\Theta,\sigma}(\Theta, \sigma).$$

Hamiltonian Monte Carlo. To draw from the posterior distribution, I use a variant of MCMC known as Hamiltonian Monte Carlo (HMC). HMC is known to offer superior performance over other generic MCMC methods when the dimension of the parameter vector is high. In the SVMA model, the dimension of the full parameter vector is $n^2(q+1)$, which can easily be well into the 100s in realistic applications. Nevertheless, the HMC algorithm has no trouble producing draws from the posterior of the SVMA parameters.

HMC outperforms standard Random Walk Metropolis-Hastings algorithms because it exploits information contained in the gradient of the log posterior density to systematically explore the posterior distribution. See Neal (2011) for a very readable overview of HMC. I use the modified HMC algorithm by Hoffman & Gelman (2014), called the No-U-Turn Sampler (NUTS), which adaptively sets the HMC tuning parameters while still provably delivering draws from the posterior distribution.

As with other MCMC methods, the HMC algorithm delivers parameter draws from a Markov chain whose long-run distribution is the posterior distribution. After discarding a burn-in sample, the output of the HMC algorithm is a collection of parameter draws $(\Theta^{(1)}, \sigma^{(1)}), \ldots, (\Theta^{(N)}, \sigma^{(N)})$, each of which is (very nearly) distributed according to the pos-

terior distribution.[32] The number N of draws is chosen by the user. The draws are not independent, and plots of the autocorrelation functions of the draws are useful for gauging the reduction in effective sample size relative to the ideal of i.i.d. sampling (Chib, 2001, pp. 3579, 3596). In my experience, the proposed algorithm for the SVMA model yields autocorrelations that drop off to zero after only a few lags.

Likelihood, score and Whittle approximation. HMC requires that the log posterior density and its gradient can be computed quickly at any given parameter values. The gradient of the log posterior density equals the gradient of the log prior density plus the gradient of the log likelihood (the latter is henceforth referred to as the *score*). In most cases, such as with the Gaussian prior in Section 1.2.5, the log prior density and its gradient are easily computed. The log likelihood and the score are the bottlenecks. In the simulation study in the next section a typical full run of the HMC procedure requires 100,000s of evaluations of the likelihood and the score.

With Gaussian shocks (Assumption 1.2), the likelihood of the SVMA model (1.3) can be evaluated using the Kalman filter, cf. Appendix A.1.3.1, but a faster alternative is to use the Whittle (1953) approximation to the likelihood of a stationary Gaussian process. Appendix A.1.3.2 shows that both the Whittle log likelihood and the Whittle score for the SVMA model can be calculated efficiently using the Fast Fourier Transform.[33] When the MA lag length q is large, as in most applications, the Whittle likelihood is noticeably faster to compute than the exact likelihood, and massive computational savings arise from using

[32] Gelman et al. (2013, Ch. 11) discuss methods for checking that the chain has converged.

[33] Hansen & Sargent (1981), Ito & Quah (1989), and Christiano & Vigfusson (2003) also employ the Whittle likelihood for SVMA models. Qu & Tkachenko (2012a,b) and Sala (2015) use the Whittle likelihood to perform approximate Bayesian inference on DSGE models, but their Random-Walk Metropolis-Hastings simulation algorithm is less efficient than HMC. Moreover, the asymptotic theory in Qu & Tkachenko (2012b) assumes identification, unlike Section 1.6.

the Whittle approximation to the score.[34]

Numerical implementation. The HMC algorithm is easy to apply once the prior has been specified. I give further details on the Bayesian computations in Appendix A.1.4.1. As initial value for the HMC iterations I use a rough approximation to the posterior mode obtained using the constructive characterization of the identified set in Appendix A.1.2. The HMC algorithm I use adapts to the posterior standard deviations of individual parameters in a warm-up phase; this speeds up computations in some applications.

Reweighting. Appendix A.1.4.2 describes an optional reweighting step that translates the Whittle HMC draws into draws from the exact posterior $p_{\Theta,\sigma|Y}(\Theta, \sigma \mid Y_T)$. If the HMC algorithm is run with the Whittle likelihood and score replacing the exact likelihood and score, the algorithm yields draws from the "Whittle posterior" density $p^W_{\Theta,\sigma|Y}(\Theta, \sigma \mid Y_T) \propto p^W_{Y|\Psi}(Y_T \mid \Psi(\Theta, \sigma))\pi_{\Theta,\sigma}(\Theta, \sigma)$, where $p^W_{Y|\Psi}(Y_T \mid \Psi(\Theta, \sigma))$ is the Whittle likelihood. Reweighting can be used if the researcher seeks finite-sample optimal inference under the Gaussian SVMA model. The reweighting step is fast and does not require computation of score vectors.

The asymptotic analysis in Section 1.6.3 shows that the reweighting step has negligible effect in large samples, as the exact and Whittle posteriors converge to the same limit under weak nonparametric conditions. However, in applications where the MA lag length q is large relative to the sample size, the asymptotic distribution may not be a good approximation to the finite-sample posterior, and reweighting may have a non-negligible effect.

[34]The exact score can be approximated using finite differences, but this is highly time-consuming. The Koopman & Shephard (1992) analytical score formula is not applicable here due to the singular measurement density in the state-space representation of the SVMA model, cf. Appendix A.1.3.1.

1.4 Simulation study

To illustrate the workings of the SVMA approach, I conduct a small simulation study with two observed variables and two shocks. I show that prior information about the smoothness of the IRFs can substantially sharpen posterior inference. It is thus desirable to use an approach, like the SVMA approach, for which prior information about smoothness is directly controlled. I also illustrate the consequences of misspecifying the prior.

The illustration is based on the bivariate example from Section 1.2 with $n = 2$ and $q = 10$, cf. Figure 1.1. The number of parameters is $n^2(q+1) = 2^2(10+1) = 44$, smaller than the dimensionality of realistic empirical applications but sufficient to elucidate the flexibility, transparency, and effectiveness of the SVAR approach.

Parameters and prior. I consider a single parametrization, with a prior that is correctly centered but diffuse. The sample size is $T = 200$. The true IRF parameters Θ are the noninvertible ones plotted in Figure 1.1. The true shock standard deviations are $\sigma_1 = 1$ (monetary policy shock) and $\sigma_2 = 0.5$ (demand shock). I first show results for the prior specification in Figure 1.4 with $\rho_{ij} = 0.9$ for all (i, j). The prior is centered at the true values but it expresses significant prior uncertainty about the magnitudes of the individual impulse responses. The prior on $\sigma = (\sigma_1, \sigma_2)$ is median-unbiased for the true values but it is very diffuse, with prior standard deviation of $\log \sigma_j$ equal to $\tau_j^\sigma = 2$ for $j = 1, 2$.

Simulation settings. I simulate a single sample of artificial data from the Gaussian SVMA model and then run the HMC algorithm using the Whittle likelihood (I do not reweight the draws as in Appendix A.1.4.2). I take 10,000 MCMC steps, storing every 10th step and discarding the first 3,000 steps as burn-in.[35] The full computation takes less than 3

[35]The results are virtually identical in simulations with 100,000 MCMC steps.

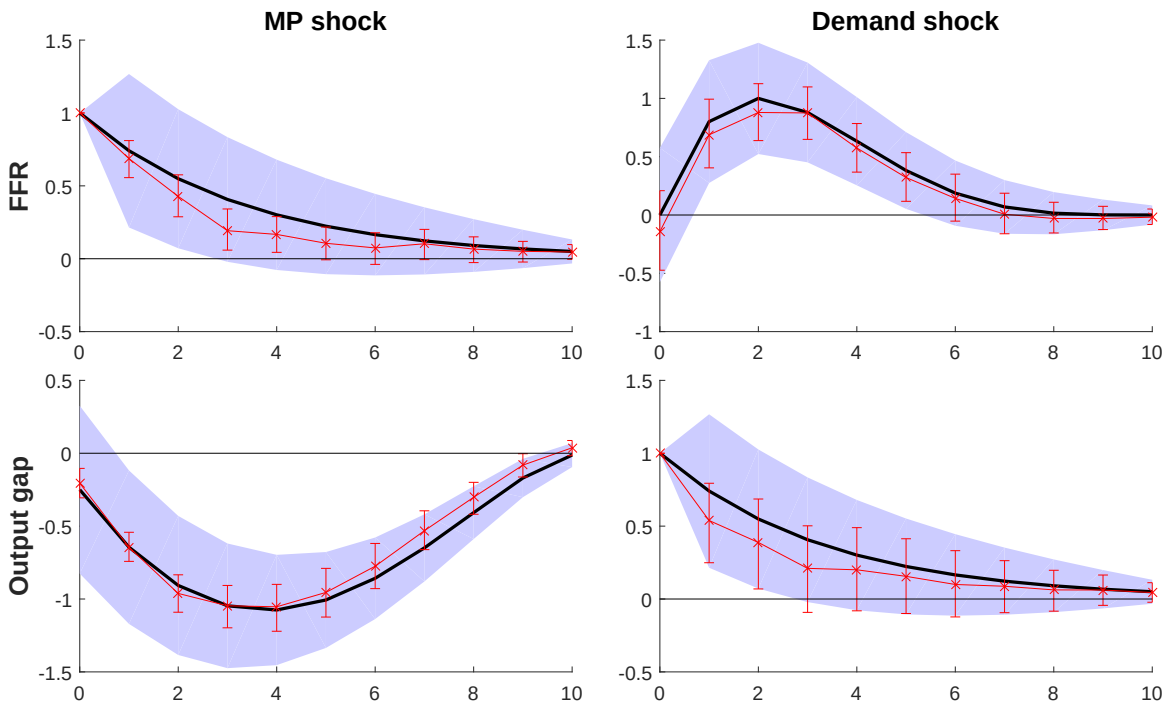

Figure 1.6: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with prior smoothness $\rho_{ij} = 0.9$. The plots show true values and prior means (thick lines), prior 90% confidence bands (shaded), posterior means (crosses), and posterior 5–95 percentile intervals (vertical bars).

hours in Matlab 8.6 on a personal laptop with 2.3 GHz Intel CPU. Appendix A.1.5.1 provides graphical diagnostics on the convergence and mixing of the MCMC chain.

Baseline results. Figure 1.6 shows that the posterior for the IRFs accurately estimates the true values and that the data serves to substantially reduce the prior uncertainty. The posterior means are generally close to the truth, although the means for two of the IRFs are slightly too low in this simulation. The 5–95 percentile posterior credible intervals are mostly much narrower than the prior 90% confidence bands, so this prior specification successfully allows the researcher to learn from the data about the magnitudes of the impulse responses. Figure 1.7 shows the posterior draws for the shock standard deviations and compares them with the prior distribution. The posterior draws are centered around the true values despite the very diffuse prior on σ. Overall, the inference method for this choice of prior works well,

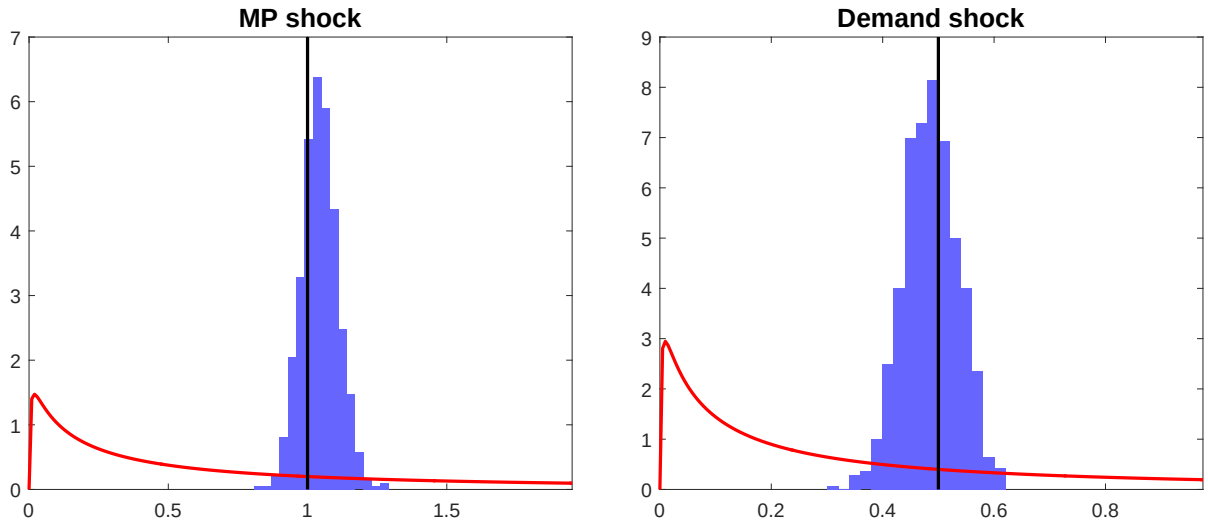

Figure 1.7: Summary of posterior shock standard deviation (σ) draws for the bivariate SVMA model with prior smoothness $\rho_{ij} = 0.9$. The plots show the true value (thick vertical line), prior density (curve), and histogram of posterior draws, for each σ_j, $j = 1, 2$.

despite the noninvertibility of the true IRFs.

Role of prior smoothness. To illustrate the importance of prior information about the smoothness of the IRFs, I run the HMC algorithm with the same specification as above, except that I set $\rho_{ij} = 0.3$ for all (i, j) in the prior, as in Figure 1.5. Figure 1.8 summarizes the posterior distribution of the IRFs corresponding to this alternative prior. Compared to Figure 1.4, the posterior credible intervals are much wider and the posterior means are less accurate estimates of the true IRFs.

The higher the degree of prior smoothness, the more do nearby impulse responses "learn from each other". Due to the prior correlation structure (1.8), any feature of the data that is informative about the impulse response $\Theta_{ij,\ell}$ is also informative about $\Theta_{ij,\ell+k}$; more so for smaller values of $|k|$, and more so for larger values of the smoothness hyperparameter ρ_{ij}. Hence, a higher degree of prior smoothness reduces the effective number of free parameters in the model. If the true IRFs are not smooth but the prior imposes a lot of smoothness, posterior inference can be very inaccurate. It is therefore important to use a framework,

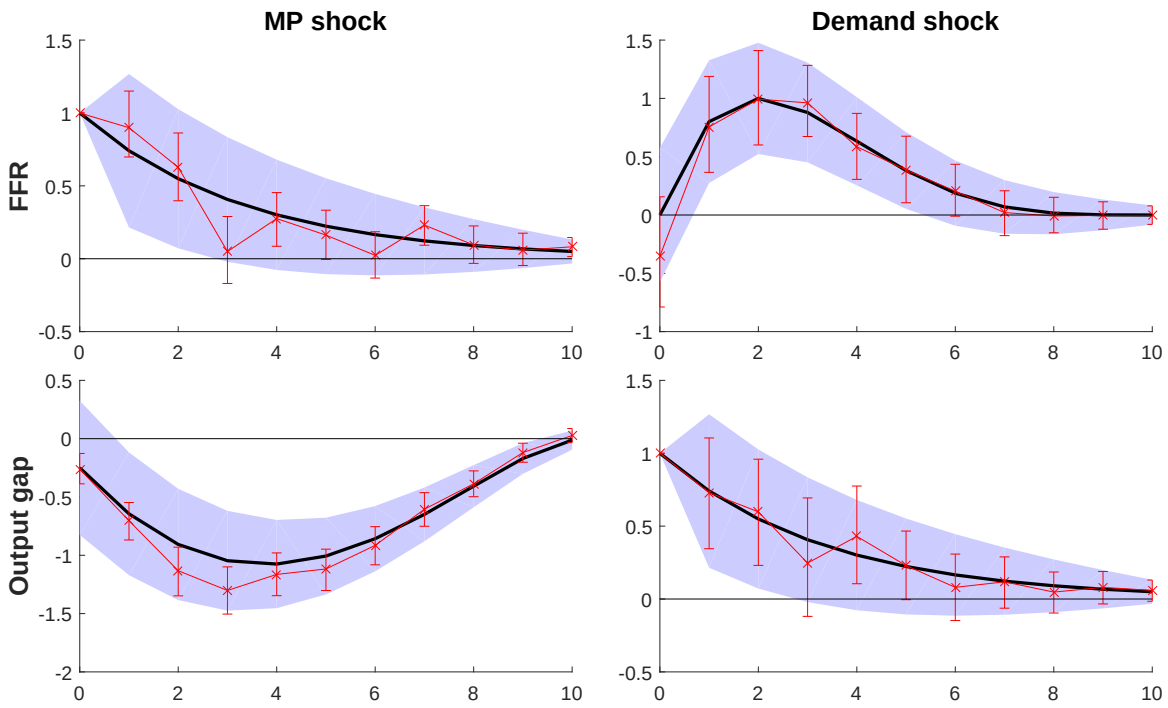

Figure 1.8: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with prior smoothness $\rho_{ij} = 0.3$. See caption for Figure 1.6.

like the SVMA approach, where prior smoothness is naturally parametrized and directly controlled. SVAR IRFs also impose smoothness *a priori*, but the degree of smoothness is implicitly controlled by the VAR lag length and restrictions on the VAR coefficients.

Misspecified priors. Appendix A.1.5.2 reports results for modifications of the baseline simulation above, maintaining the prior distribution but substantially modifying the true IRFs. I consider two such experiments: one in which the shocks have less persistent effects than the prior indicates, and one in which the true IRF of the output gap to a monetary policy shock is uniformly zero. In both cases, the inaccurate prior is overruled by the data, delivering reasonably accurate posterior inference. This happens because the implied prior distribution of the ACF is inconsistent with the true ACF. Since the data is informative about the latter, the posterior distribution puts more weight than the prior on parameters

that are consistent with the true ACF, as shown formally in Section 1.6.3.[36]

1.5 Application: News shocks and business cycles

I now use the SVMA method to infer the role of technological news shocks in the post-war U.S. business cycle. Following the literature, I define a technological news shock to be a signal about future productivity increases. My prior on IRFs is partially informed by a conventional sticky-price DSGE model, without imposing the model restrictions dogmatically. The analysis finds overwhelming evidence of noninvertible IRFs in my specification, yet most of the IRFs are estimated precisely. Furthermore, news shocks are relatively unimportant drivers of productivity and output growth, but more important for the real interest rate. Graphical diagnostics show that the posterior inference is insensitive to moderate changes in the prior; they also point to possible fruitful extensions of the model.

Technological news shocks have received great attention in the recent empirical and theoretical macro literature, but researchers have not yet reached a consensus on their importance, cf. the survey by Beaudry & Portier (2014). As explained in Section 1.2.3, theoretical macro models with news shocks often feature noninvertible IRFs, giving the SVMA method a distinct advantage over SVARs, as the latter assume away noninvertibility. My news shock analysis is the first to fully allow for noninvertible IRFs while refraining from dogmatically imposing a particular DSGE model structure (see the discussion at the end of this section).

Specification and data. I employ a SVMA model with three observed variables and three unobserved shocks: Total factor productivity (TFP) growth, real gross domestic prod-

[36]By the same token, if the true parameter values were chosen to be observationally equivalent to the prior medians in Figure 1.4 (i.e., they imply the same ACF), then the posterior would look the same as in Figures 1.6 and 1.7 up to simulation noise, even though the true parameters could be very different from the prior medians. Hence, not all misspecified priors can be corrected by the data, cf. Section 1.6.3.

uct (GDP) growth, and the real interest rate are assumed to be driven by a productivity shock, a technological news shock, and a monetary policy shock. I use quarterly data from 1954Q3–2007Q4, yielding sample size $T = 213$ (a quarter is lost when transforming to growth rates). I exclude data from 2008 to the present as my analysis ignores financial shocks.

TFP growth equals 100 times the log growth rate of TFP and is taken from the data appendix to Fernald (2014).[37] The remaining data is from the St. Louis Federal Reserve's FRED database.[38] Real GDP growth is given by 100 times the log growth rate of seasonally adjusted GDP per capita in chained dollars, as measured by the Bureau of Economic Analysis (NIPA Table 7.1, line 10). My real interest rate series equals the nominal policy interest rate minus the contemporaneous inflation rate.[39] The nominal policy rate is the average effective federal funds rate, expressed as a quarterly rate. The inflation rate equals 100 times the log growth rate in the seasonally adjusted implicit price deflator for the non-farm business sector, as reported by the Bureau of Labor Statistics.

Before running the analysis, I detrend the three data series to remove secular level changes that are arguably unrelated to the business cycle. Following Stock & Watson (2012a, Sec. I.C), I estimate the trend in each series using a biweight kernel smoother with a bandwidth of 100 quarters; the trends are then subtracted from the raw series. Figure 1.9 plots the raw data and the estimated time-varying trends.

I pick a MA lag length of $q = 16$ quarters based on two considerations. First, the Akaike Information Criterion (computed using the Whittle likelihood) selects $q = 13$. Second, the

---

[37]The TFP measure is based on a growth accounting method that adjusts for differing marginal products of capital across sectors as well as changes over time in labor quality and labor's share of income. Fernald (2014) also estimates utilization-adjusted TFP, but the adjustment is model-based and reliant on estimates from annual regressions on a separate dataset, so I prefer the simpler series. Data downloaded July 14, 2015.

[38]FRED series codes: A939RX0Q048SBEA (real GDP per capita), FEDFUNDS (effective federal funds rate), and IPDNBS (implicit price deflator, non-farm business sector). Data downloaded August 13, 2015.

[39]If agents form inflation expectations under the presumption that quarterly inflation follows a random walk, then my measure of the real interest rate equals the conventional *ex ante* real interest rate.

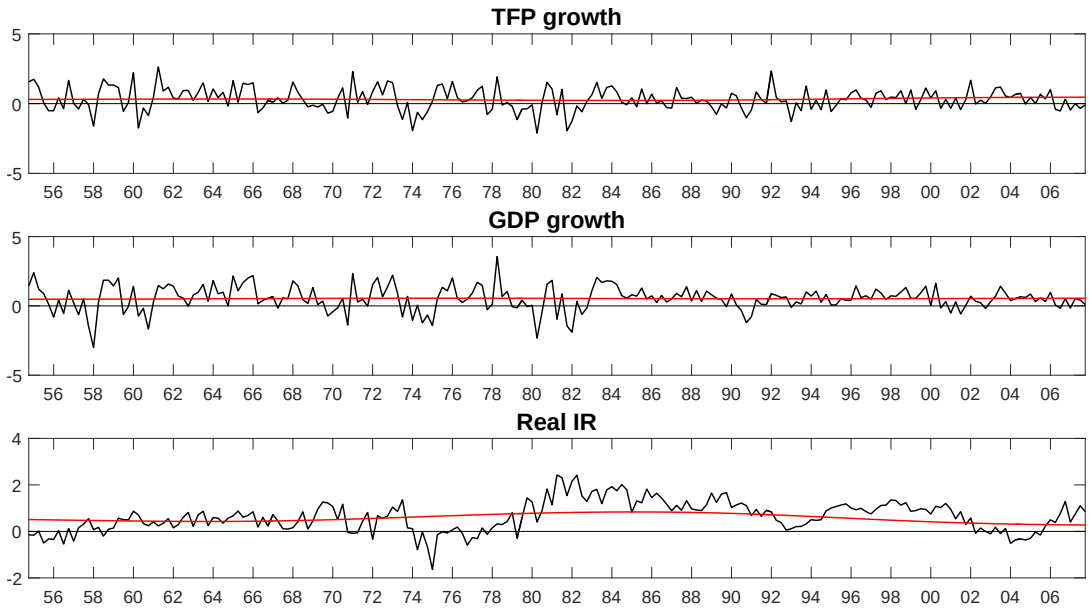

Figure 1.9: Raw data on TFP growth, GDP growth, and the real interest rate (IR), along with estimated time-varying trends (smooth curves). The final data used in the empirical analysis are differences between the raw series and the trends.

autocorrelation of the real interest rate equals 0.17 at lag 13 but is close to zero at lag 16.

Prior. The prior on the IRFs is of the multivariate Gaussian type introduced in Section 1.2.5, with hyperparameters informed by a conventional sticky-price DSGE model. The DSGE model is primarily used to guide the choice of prior means, and the model restrictions are not imposed dogmatically on the SVMA IRFs. Figure 1.10 plots the prior means and variances for the impulse responses, along with four draws from the joint prior distribution. The figure also shows the normalization that defines the scale of each shock.

The DSGE model used to inform the prior is the one developed by Sims (2012, Sec. 3). It is built around a standard New Keynesian structure with monopolistically competitive firms subject to a Calvo pricing friction, and the model adds capital accumulation, investment adjustment costs, internal habit formation in consumption, and interest rate smoothing in the Taylor rule. Within the DSGE model, the productivity and news shocks are, respectively,

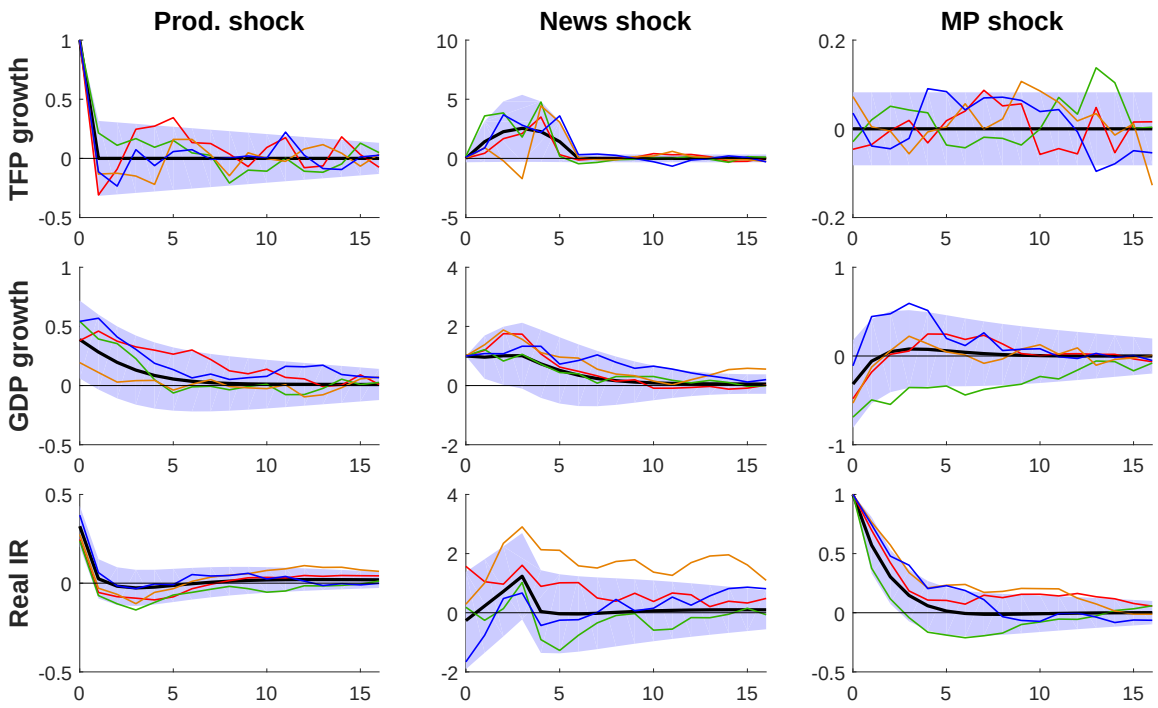

Figure 1.10: Prior means (thick lines), 90% prior confidence bands (shaded), and four random draws (brightly colored lines) from the prior for IRFs (Θ), news shock application. The impact impulse response is normalized to 1 in each IRF along the diagonal of the figure.

unanticipated and anticipated exogenous disturbances to the change in log TFP (cf. eq. 30–33 in Sims, 2012). The monetary policy shock is an unanticipated disturbance term in the Taylor rule (cf. eq. 35 in Sims, 2012). Detailed model assumptions and equilibrium conditions are described in Sims (2012, Sec. 3), but I repeat that I only use the DSGE model to guide the SVMA prior; the model restrictions are not imposed dogmatically.[40]

As prior means for the nine SVMA IRFs I use the corresponding IRFs implied by the log-

[40]My approach is distinct from IRF matching (Rotemberg & Woodford, 1997). In IRF matching, a SVAR is identified using exclusion restrictions, and then the structural parameters of a DSGE model are chosen so that the DSGE-implied IRFs match the estimated SVAR IRFs. In my procedure, the DSGE model informs the choice of prior on IRFs, but then the data is allowed to speak through a flexible SVMA model. I do not treat the DSGE model as truth, and I impose prior restrictions in a single stage. Ingram & Whiteman (1994) and Del Negro & Schorfheide (2004) apply related ideas to VAR models. Geweke (2010, Ch. 4.4) proposes a general method for letting DSGE models inform priors without imposing model restrictions dogmatically.

linearized DSGE model, with one exception mentioned below.[41] I use the baseline calibration of Sims (2012, Table 1), which assumes that news shocks are correctly anticipated TFP increases taking effect three quarters into the future. Because I am particularly uncertain that an anticipation horizon of three quarters is correct, I modify the prior means for the impulse responses of TFP growth to the news shock: The prior means smoothly increase and then decrease over the interval $\ell \in [0, 6]$, with a maximum value at $\ell = 3$ equal to half the DSGE-implied impulse response.[42]

The prior variances for the IRFs are chosen by combining information from economic intuition and DSGE calibration sensitivity experiments. For example, I adjust the prior variances for the IRFs so that the DSGE-implied IRFs mostly fall within the 90% prior bands when the anticipation horizon changes between nearby values. The 90% prior bands for the IRFs that correspond to the news shock are chosen quite large, and they mostly contain 0. In contrast, the prior bands corresponding to the monetary policy shock are narrower, expressing a strong belief that monetary policy shocks have a small (not necessarily zero) effect on TFP growth but a persistent positive effect on the real interest rate. The prior bands for the effects of productivity shocks on GDP growth and on the real interest rate are fairly wide, since these IRFs should theoretically be sensitive to the degree of nominal stickiness in the economy as well as to the Federal Reserve's information set and policy rule.

The prior expresses a belief that the IRFs for GDP growth and the real interest rate are quite smooth, while those for TFP growth are less smooth. Specifically, I set $\rho_{1j} = 0.5$ and $\rho_{2j} = \rho_{3j} = 0.9$ for $j = 1, 2, 3$. These choices are based on economic intuition and are consistent with standard calibrations of DSGE models. The ability to easily impose different

[41] The DSGE-implied IRFs for the real interest rate use the same definition of this variable as in the construction of the data series, i.e., nominal interest rate minus contemporaneous inflation. The IRFs are computed using Dynare 4.4.3 (Adjemian et al., 2011).

[42] A theoretically more satisfying way to deal with uncertainty about the anticipation horizon is to use a Gaussian mixture prior, where the categorical component label is the anticipation horizon.

degrees of prior smoothness across IRFs is unique to the SVMA approach; it would be much harder to achieve in a SVAR set-up.

The prior on the shock standard deviations is very diffuse. For each shock j, the prior mean μ_j^σ of $\log(\sigma_j)$ is set to $\log(0.5)$, while the prior standard deviation τ_j^σ is set to 2.[43] These values should of course depend on the units of the observed series.

As a consistency check, Appendix A.1.6.1 shows that the Bayesian computation procedure with the above prior accurately recovers the DSGE-implied IRFs from simulated data.

Results. Given my prior, the data is informative about most of the IRFs. Figure 1.11 summarizes the posterior distribution of each impulse response. Figure 1.12 plots the posterior distribution of long-run (i.e., cumulative) impulse responses $\sum_{\ell=0}^{q} \Theta_{ij,\ell}$ for each variable-shock combination (i, j). Figure 1.13 shows the posterior distribution of the forecast error variance decomposition (FEVD) of each variable i to each shock j at each horizon ℓ, defined as[44]

$$FEVD_{ij,\ell} = \frac{\mathrm{Var}(\sum_{k=0}^{q} \Theta_{ij,k}\varepsilon_{j,t+\ell-k} \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \dots)}{\mathrm{Var}(y_{i,t+\ell} \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \dots)} = \frac{\sum_{k=0}^{\ell} \Theta_{ij,k}^2 \sigma_j^2}{\sum_{b=1}^{n} \sum_{k=0}^{\ell} \Theta_{ib,k}^2 \sigma_b^2}. \tag{1.9}$$

$FEVD_{ij,\ell}$ is the fraction of the forecast error variance that would be eliminated if we knew all future realizations of shock j when forming ℓ-quarter-ahead forecasts of variable i at time t using knowledge of the shocks up to time $t-1$.

The posterior means for several IRFs differ substantially from the prior means, and the posterior 90% intervals are narrower than the prior 90% bands. The effects of productivity

[43] The prior is agnostic about the relative importance of the three shocks: Due to the diffuse prior on shock standard deviations, unreported simulations show that the prior 5th and 95th percentiles of the FEVD (cf. (1.9)) are very close to 0 and 1, respectively, for almost all (i, j, ℓ) combinations.

[44] The variances in the fraction are computed under the assumption that the shocks are serially and mutually independent. In the literature the FEVD is defined by conditioning on $(y_{t-1}, y_{t-2}, \dots)$ instead of $(\varepsilon_{t-1}, \varepsilon_{t-2}, \dots)$. This distinction matters when the IRFs are noninvertible. Baumeister & Hamilton (2015a) conduct inference on the FEVD in a Bayesian SVAR, assuming invertibility.

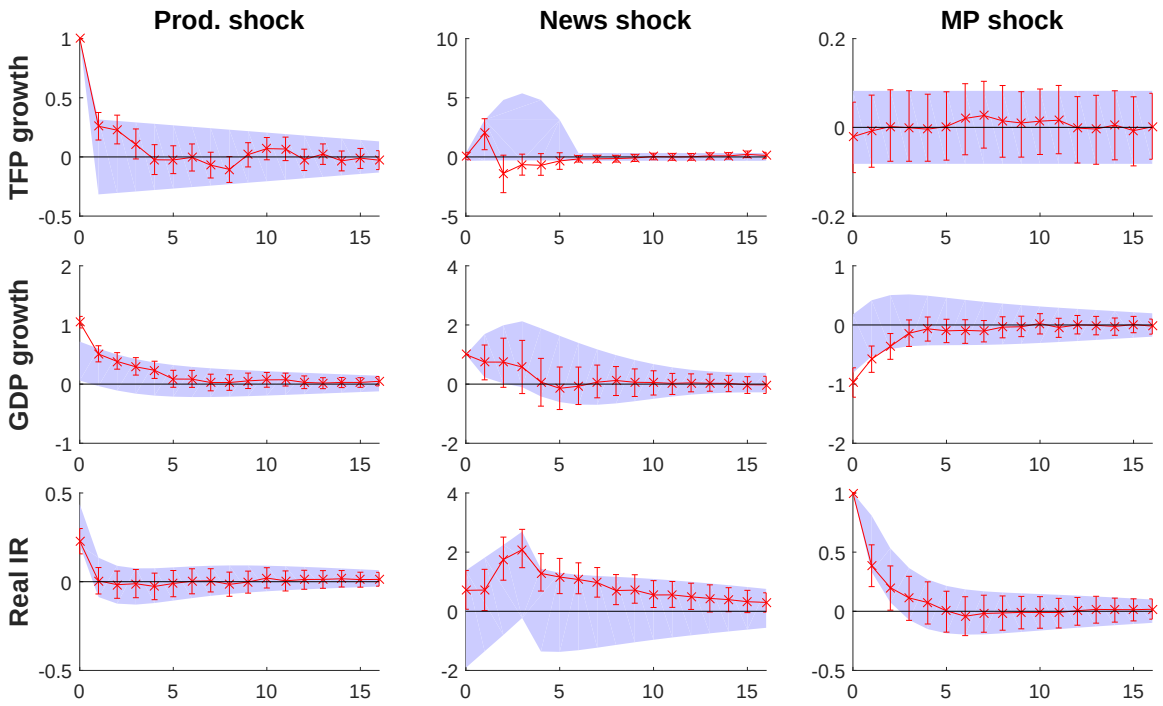

Figure 1.11: Summary of posterior IRF (Θ) draws, news shock application. The plots show prior 90% confidence bands (shaded), posterior means (crosses), and posterior 5–95 percentile intervals (vertical bars).

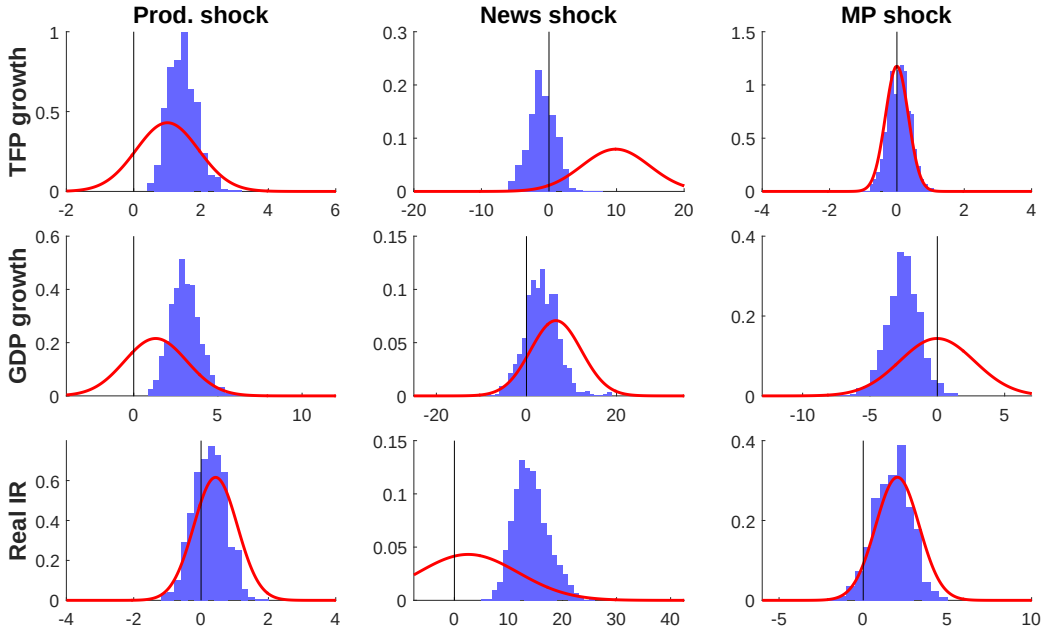

Figure 1.12: Histograms of posterior draws of long-run impulse responses $\sum_{\ell=0}^{q} \Theta_{ij,\ell}$ for each (i, j), news shock application. Curves are prior densities. Histograms and curves integrate to 1.

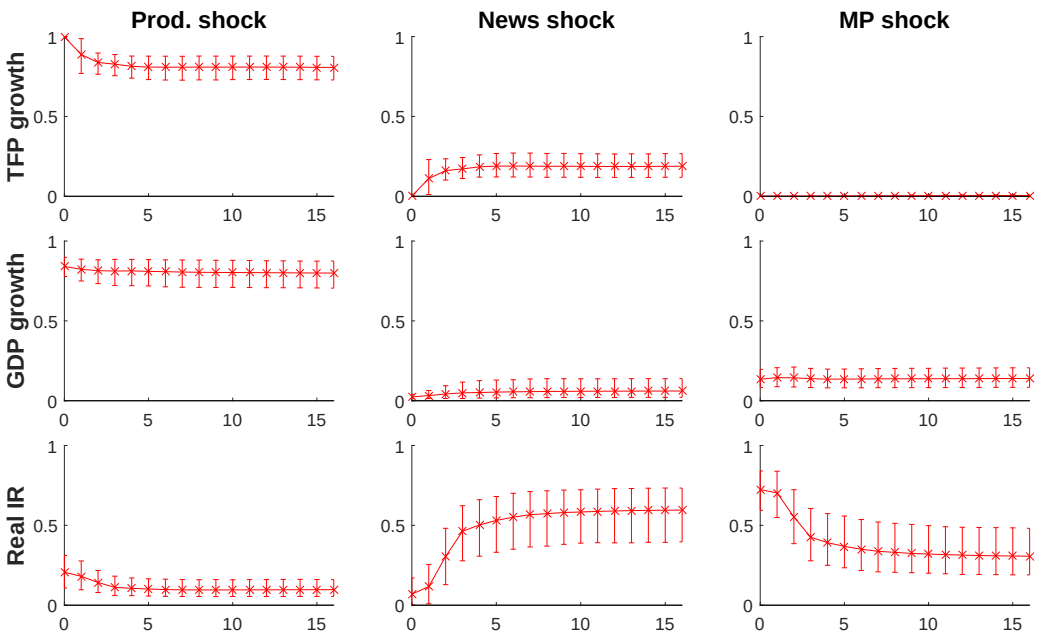

Figure 1.13: Summary of posterior draws of $FEVD_{ij,\ell}$ (1.9), news shock application. The figure shows posterior means (crosses) and posterior 5–95 percentile intervals (vertical bars). For each variable i and each horizon ℓ, the posterior means sum to 1 across the three shocks j.

and monetary policy shocks on TFP and GDP growth are especially precisely estimated. From the perspective of the prior beliefs, it is surprising to learn that the impact effect of productivity shocks on GDP growth is quite large, and the effect of monetary policy shocks on the real interest rate is not very persistent. Figure 1.12 shows that the monetary policy shock has negative and substantially non-neutral effects on the level of GDP in the long run, even though the prior distribution for this long-run response is centered around zero.

The IRFs corresponding to the news shock are not estimated as precisely as IRFs for the other shocks, but the data does noticeably update the prior. The IRF of TFP growth to the news shock indicates that future productivity increases are anticipated only one quarter ahead, and the increase is mostly reversed in the following quarters. According to the data, the long-run response of TFP to a news shock is unlikely to be substantially positive, implying that economic agents seldom correctly anticipate shifts in medium-run productivity levels. The news shock is found to have substantially less persistent effects on GDP growth than

predicted by the DSGE model. However, the effect of the news shock on the real interest rate is found to be large and persistent.

The news shock is not an important driver of TFP and GDP growth but is important for explaining real interest rate movements at longer horizons. According to Figure 1.13, the news shock contributes little to the forecast error variance for TFP and GDP growth at all horizons. The monetary policy shock is only slightly more important in explaining GDP growth, while the productivity shock is much more important by these measures. However, the monetary policy shock is important for explaining short-run movements in the real interest rate, while the news shock dominates longer-run movements in this series.

The data and prior provide overwhelming evidence that the IRFs are noninvertible. In Figure 1.14 I report a continuous measure of invertibility suggested by Watson (1994, p. 2901) and Sims & Zha (2006, p. 243). For each posterior parameter draw I compute the R^2 from a population regression of each shock $\varepsilon_{j,t}$ on current and 50 lags of past data $(y_t, y_{t-1}, \ldots, y_{t-50})$, assuming i.i.d. Gaussian shocks.[45] This R^2 value should be essentially 1 for all shocks if the IRFs are invertible, by definition. Instead, Figure 1.14 shows a high posterior probability that the news shock R^2 is below 0.3, despite the prior putting most weight on values near 1.[46]

The noninvertibility of the estimated IRFs is economically significant. Figure 1.15 summarizes the posterior distribution of those invertible IRFs that are closest to the actual (possibly noninvertible) IRFs. Specifically, for each posterior draw (Θ, σ) I compute the parameter vector $(\tilde{\Theta}, \tilde{\sigma})$ that minimizes the Frobenius distance $\|\Theta \operatorname{diag}(\sigma) - \tilde{\Theta} \operatorname{diag}(\tilde{\sigma})\|$ over

[45] Given the parameters, I run the Kalman filter in Appendix A.1.3.1 forward for 51 periods on data that is identically zero (due to Gaussianity, conditional variances do not depend on realized data values). This yields a final updated state prediction variance matrix $\operatorname{Var}(\operatorname{diag}(\sigma)^{-1}\varepsilon_{51} \mid y_{51}, \ldots, y_1)$ whose diagonal elements equal 1 minus the desired population R^2 values at the given parameters.

[46] Essentially no posterior IRF draws are exactly invertible; the prior probability is 0.06%.

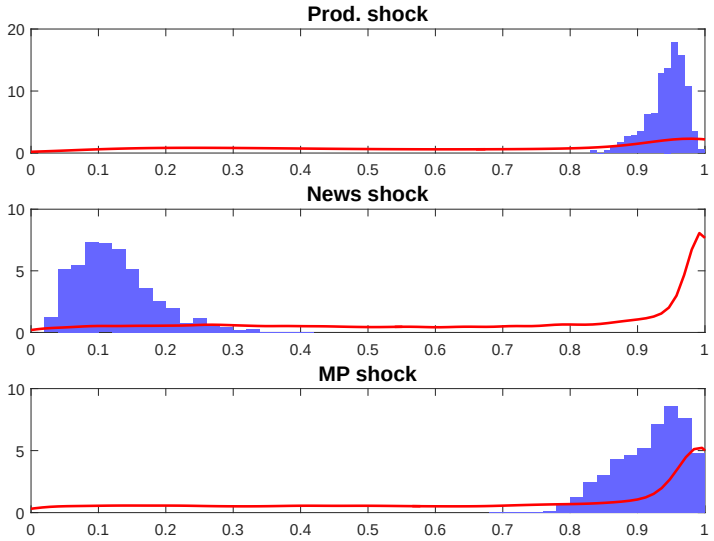

Figure 1.14: Histograms of posterior draws of the population R^2 values in regressions of each shock on current and 50 lagged values of the observed data, news shock application. Curves are kernel density estimates of the prior distribution of R^2s. Histograms and curves integrate to 1.

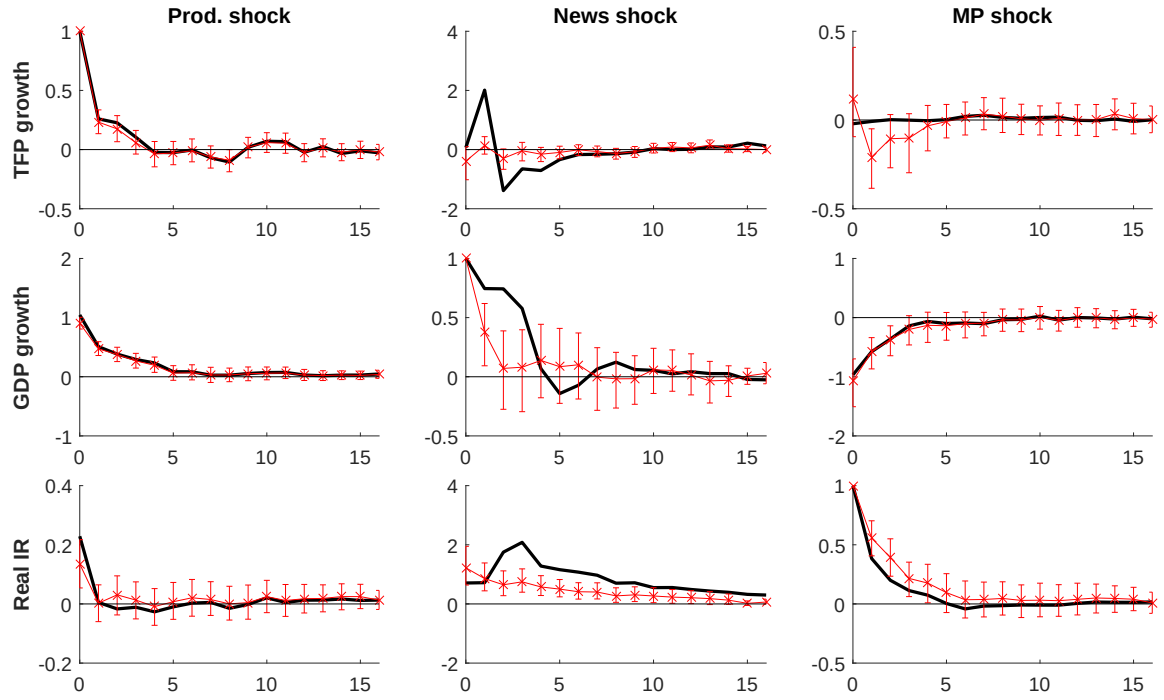

Figure 1.15: Posterior distribution of the invertible IRFs that are closest to the actual IRFs, news shock application. The figure shows posterior means of actual IRFs from Figure 1.11 (thick lines), posterior means of the closest invertible IRFs (crosses), and posterior 5–95 percentile intervals for these invertible IRFs (vertical bars).

parameters for which $\tilde{\Theta}$ is invertible and $(\tilde{\Theta}, \tilde{\sigma})$ generates the same ACF as (Θ, σ).[47] While the invertible IRFs for the productivity and monetary policy shocks are similar to the unrestricted IRFs, the invertible news shock IRFs look nothing like the actual estimated IRFs.[48] Thus, no SVAR identification scheme can deliver accurate inference about the effects of technological news shocks in this dataset.

Appendix A.1.6.2 uses standard state-space methods to estimate the structural shocks, which is straight-forward despite noninvertibility of the IRFs.

Prior sensitivity and model validation. In Appendix A.1.6.3 I show that the posterior inference is insensitive to moderate changes in the prior distribution. I use the Müller (2012) measure of local prior sensitivity, which allows me to graphically summarize the sensitivity of the posterior mean of each impulse response.

I conduct a battery of graphical posterior predictive checks to identify ways to improve the model's fit. As shown in Section 1.6, the posterior distribution of the parameters of the Gaussian SVMA model fits the unconditional second moments of the observed data well in large samples. In Appendix A.1.6.4 I investigate whether the model also matches higher moments and conditional time series properties. While the Gaussianity-based posterior analysis is robust to violations of Gaussianity and other model assumptions, cf. Section 1.6, the posterior predictive analysis points to ways the model could be improved to increase statistical efficiency. The analysis suggests it would be fruitful to extend the model to include stochastic volatility and nonlinearities. I briefly discuss such extensions in Section 1.8.

---

[47] According to Appendix A.1.2, $(\tilde{\Theta}, \tilde{\sigma})$ is obtained as follows. First apply transformation (ii) in Theorem A.1 several times to (Θ, σ) in order to flip all roots outside the unit circle. Denote the resulting invertible parameters by $(\check{\Theta}, \check{\sigma})$. Then $\tilde{\Theta}\,\mathrm{diag}(\tilde{\sigma}) = \check{\Theta}\,\mathrm{diag}(\check{\sigma})Q$, where Q is the orthogonal matrix that minimizes $\|\Theta\,\mathrm{diag}(\sigma) - \check{\Theta}\,\mathrm{diag}(\check{\sigma})Q\|$. This "orthogonal Procrustes problem" has a well-known solution.

[48] Figure 1.15 cannot be interpreted as the posterior distribution corresponding to a prior which truncates the prior from Figure 1.10 to the invertible region. It is difficult to sample from this truncated posterior, as essentially none of the unrestricted posterior draws are invertible, so an accept-reject scheme is inapplicable.

Comparison with the literature. My conclusion that technological news shocks are not important for explaining business cycles is consistent with the literature, but my method is the first to allow for noninvertibility without additional assumptions. Forni et al. (2014) estimate small effects of technological news shocks in a factor-augmented SVAR. Their empirical strategy may overcome the noninvertibility issue if technological news are well captured by the first few principal components of their large macroeconomic panel data set. They confirm that low-dimensional systems (without factors) are noninvertible. Papers that estimate fully-specified DSGE models with news shocks also tend to find a limited role for technological news, cf. the review by Beaudry & Portier (2014, Sec. 4.2.2). Unlike these papers, I do not dogmatically impose restrictions implied by a particular structural model.

Several SVAR papers on news shocks have used stock market data in an attempt to overcome the invertibility problem, cf. Beaudry & Portier (2014, Sec. 3). Such SVAR specifications may be valid if the stock market is a good proxy for the news shock, i.e., if the market responds immediately and forcefully upon arrival of technological news. On the other hand, if market movements are highly contaminated by other types of shocks, incorporating stock market data may lead to biased SVAR estimates. It would be interesting to incorporate stock market data into my analysis to fully reconcile my results with these SVAR analyses.

1.6 Asymptotic theory

To gain insight into how the data updates the prior information, I derive the asymptotic limit of the Bayesian posterior distribution from a frequentist point of view. I first derive general results on the frequentist asymptotics of Bayes procedures for a large class of partially identified models that includes the SVMA model. Then I specialize to the SVMA model and show that, asymptotically, the role of the data is to pin down the true autocovariances, whereas all other information about IRFs comes from the prior. The asymptotics imply

that the limiting form of the posterior is robust to violations of the assumption of Gaussian shocks and to the use of the Whittle likelihood in place of the exact likelihood.

1.6.1 General results for partially identified models

In this subsection I present a general result on the frequentist asymptotic limit of the Bayesian posterior distribution for a large class of partially identified models that includes the SVMA model. Due to the lack of identification, the asymptotic analysis is nonstandard, as the data does not dominate all aspects of the prior distribution in large samples.

Consider a general model for which the data vector Y_T is independent of the parameter of interest θ, conditional on a second parameter Γ.[49] In other words, the likelihood function of the data Y_T only depends on θ through Γ. This property holds for models with a partially identified parameter θ, as explained in Poirier (1998). Because I will restrict attention to models in which the parameter Γ is identified, I refer to Γ as the reduced-form parameter, while θ is called the structural parameter. The parameter spaces for Γ and θ are denoted Ξ_Γ and Ξ_θ, respectively, and these are assumed to be finite-dimensional Euclidean and equipped with the Frobenius norm $\|\cdot\|$.

As an illustration, consider the SVMA model with data vector $Y_T = (y_1', \ldots, y_T')'$. Let $\Gamma = (\Gamma(0), \ldots, \Gamma(q))$ be the ACF of the observed time series, and let θ denote a single IRF, for example the IRF of the first variable to the first shock, i.e., $\theta = (\Theta_{11,0}, \ldots, \Theta_{11,q})'$. I explain below why I focus on a single IRF. Since the distribution of the stationary Gaussian process y_t only depends on θ through the ACF Γ, we have $Y_T \perp\!\!\!\perp \theta \mid \Gamma$.

In any model satisfying $Y_T \perp\!\!\!\perp \theta \mid \Gamma$, the prior information about θ *conditional* on Γ is not updated by the data Y_T, but the data is informative about Γ. Let $P_{\theta|Y}(\cdot \mid Y_T)$ denote the posterior probability measure for θ given data Y_T, and let similarly $P_{\Gamma|Y}(\cdot \mid Y_T)$ denote the

[49] T denotes the sample size, but the model does not have to be a time series model.

posterior measure for Γ. For any $\tilde{\Gamma} \in \Xi_\Gamma$, let $\Pi_{\theta|\Gamma}(\cdot \mid \tilde{\Gamma})$ denote the conditional prior measure for θ given Γ, evaluated at $\Gamma = \tilde{\Gamma}$. As in Moon & Schorfheide (2012, Sec. 3), decompose

$$P_{\theta|Y}(\mathcal{A} \mid Y_T) = \int_{\Xi_\Gamma} \Pi_{\theta|\Gamma}(\mathcal{A} \mid \Gamma) P_{\Gamma|Y}(d\Gamma \mid Y_T) \tag{1.10}$$

for any measurable set $\mathcal{A} \subset \Xi_\theta$. Let Γ_0 denote the true value of Γ. If the reduced-form parameter Γ_0 is identified, the posterior $P_{\Gamma|Y}(\cdot \mid Y_T)$ for Γ will typically concentrate around Γ_0 in large samples, so that the posterior for θ is well approximated by $P_{\theta|Y}(\cdot \mid Y_T) \approx \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)$, the conditional prior for θ given Γ at the true Γ_0.

The following lemma formalizes the intuition about the asymptotic limit of the posterior distribution for θ. Define the L_1 norm $\|P\|_{L_1} = \sup_{|h| \leq 1} \int |h(x)| P(dx)$ on the space of signed measures, where P is any signed measure and the supremum is over all scalar real-valued measurable functions $h(\cdot)$ bounded in absolute value by 1.[50]

Lemma 1.1. *Let the posterior measure $P_{\theta|Y}(\cdot \mid Y_T)$ satisfy the decomposition* (1.10)*. All stochastic limits below are taken under the true probability measure of the data. Assume:*

(i) The map $\tilde{\Gamma} \mapsto \Pi_{\theta|\Gamma}(\theta \mid \tilde{\Gamma})$ is continuous at Γ_0 with respect to the L_1 norm $\| \cdot \|_{L_1}$.[51]

(ii) For any neighborhood $\mathcal{U}$ of Γ_0 in Ξ_Γ, $P_{\Gamma|Y}(\mathcal{U} \mid Y_T) \xrightarrow{p} 1$ as $T \to \infty$.

Then as $T \to \infty$,

$$\|P_{\theta|Y}(\cdot \mid Y_T) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1} \xrightarrow{p} 0.$$

---

[50]The L_1 distance $\|P_1 - P_2\|_{L_1}$ between two probability measures P_1 and P_2 equals twice the total variation distance (TVD) between P_1 and P_2. TVD is an important metric, as convergence in TVD implies convergence of Bayes point estimators under certain side conditions (van der Vaart, 1998, Ch. 10.3).

[51]Denote the underlying probability sample space by Ω, and let $\mathcal{B}_\theta$ be the Borel sigma-algebra on Ξ_θ. Formally, assumption (i) requires the existence of a function $\varsigma \colon \mathcal{B}_\theta \times \Xi_\Gamma \to \mathbb{R}_+$ such that $\{\varsigma(B, \Gamma(o))\}_{B \in \mathcal{B}_\theta,\, o \in \Omega}$ is a version of the conditional probability measure of θ given Γ, and such that $\|\varsigma(\cdot, \Gamma_k) - \varsigma(\cdot, \Gamma_0)\|_{L_1} \to 0$ as $k \to \infty$ for any sequence $\{\Gamma_k\}_{k \geq 1}$ satisfying $\Gamma_k \to \Gamma_0$ and $\Gamma_k \in \Xi_\Gamma$.

If furthermore $\hat{\Gamma}$ *is a consistent estimator of* Γ_0*, i.e.,* $\hat{\Gamma} \xrightarrow{p} \Gamma_0$*, then*

$$\|P_{\theta|Y}(\cdot \mid Y_T) - \Pi_{\theta|\Gamma}(\cdot \mid \hat{\Gamma})\|_{L_1} \xrightarrow{p} 0.$$

In addition to stating the asymptotic form of the posterior distribution, Lemma 1.1 yields three main insights. First, the posterior for θ given the data does not collapse to a point asymptotically, a consequence of the lack of identification.[52] Second, the sampling uncertainty about the true reduced-form parameter Γ_0, which is identified in the sense of assumption (ii), is asymptotically negligible relative to the uncertainty about θ given knowledge of Γ_0. Third, in large samples, the way the data disciplines the prior information on θ is through the consistent estimator $\hat{\Gamma}$ of Γ_0.

Lemma 1.1 gives weaker and simpler conditions for result (ii) in Theorem 1 of Moon & Schorfheide (2012). Lipschitz continuity in Γ of the conditional prior measure $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma)$ (their Assumption 2) is weakened to continuity, and the high-level assumption of asymptotic normality of the posterior for Γ (their Assumption 1) is weakened to posterior consistency.

Assumption (i) invokes continuity with respect to Γ of the conditional prior of θ given Γ. This assumption is satisfied in many models with partially identified parameters, if θ is chosen appropriately. The assumption is unlikely to be satisfied in other contexts. For example, if θ were identified because there existed a function mapping Γ to θ, and Γ were identified, then assumption (i) could not be satisfied. More generally, assumption (i) will typically not be satisfied if the identified set for θ lies in a lower-dimensional subspace of Ξ_θ.[53] If continuity of $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma)$ does not hold, assumption (ii) on the limiting posterior distribution

[52] As emphasized by Gustafson (2015, pp. 35, 59–61), the Bayesian approach to partial identification explicitly acknowledges the role of prior information even in infinite samples. This stands in contrast with traditional "identified" models, for which the potential bias due to misspecification of the identifying restrictions is often unacknowledged and difficult to characterize.

[53] For a discussion of this point, see Remarks 2 and 3, pp. 768–770, in Moon & Schorfheide (2012).

for Γ can be strengthened to derive an asymptotic approximation to the posterior for θ, cf. Moon & Schorfheide (2012, Sec. 3).

Assumption (ii) invokes posterior consistency for Γ_0, i.e., the posterior for the reduced-form parameter Γ must concentrate on small neighborhoods of the true value Γ_0 in large samples. While assumption (i) is a condition on the prior, assumption (ii) may be viewed as a condition on the likelihood of the model, although assumption (ii) does require that the true reduced-form parameter Γ_0 is in the support of the marginal prior distribution on Γ. As long as the reduced-form parameter Γ_0 is identified, posterior consistency holds under very weak regularity conditions, as discussed in Appendix A.1.7.1 and in the next subsection.[54]

As the proof of Lemma 1.1 shows, the likelihood function used to calculate the posterior measure does not have to be correctly specified. That is, if $\tilde{\Gamma} \mapsto p_{Y|\Gamma}(Y_T \mid \tilde{\Gamma})$ denotes the likelihood function for Γ used to compute the posterior $P_{\Gamma|Y}(\cdot \mid Y_T)$, then $p_{Y|\Gamma}(Y_T \mid \Gamma_0)$ need not be the true density of the data. As long as $P_{\Gamma|Y}(\cdot \mid Y_T)$ is a probability measure that satisfies assumption (ii), where the convergence in probability occurs under the true probability measure of the data, then the conclusion of the lemma follows. This observation is helpful when I derive the limit of the Whittle posterior for the SVMA model.

1.6.2 Posterior consistency for the autocovariance function

I now show that the posterior consistency assumption for the reduced-form parameter Γ in Lemma 1.1 is satisfied in almost all stationary time series models for which Γ can be chosen to be the ACF, as in the SVMA model. The result below supposes that the posterior measure for the ACF Γ is computed using the Whittle likelihood under the working assumption that the time series is stationary Gaussian and q-dependent, i.e., the autocovariances after lag q are zero. This is the case for the SVMA model. I show that the Whittle posterior is

[54] See also Ghosh & Ramamoorthi (2003, Ch. 1.3) and Choudhuri, Ghosal & Roy (2005, Ch. 3).

consistent for the true ACF (up to lag q) even if the true data generating process is in fact not Gaussian or q-dependent.[55]

The only restrictions imposed on the underlying true data generating process are the following nonparametric stationarity and weak dependence assumptions.

Assumption 1.3. *$\{y_t\}$ is an n-dimensional time series satisfying the following assumptions. All limits and expectations below are taken under the true probability measure of the data.*

(i) *$\{y_t\}$ is a covariance stationary time series with mean zero.*

(ii) *$\sum_{k=-\infty}^{\infty} \|\Gamma_0(k)\| < \infty$, where the true ACF is defined by $\Gamma_0(k) = E(y_{t+k}y_t')$, $k \in \mathbb{Z}$.*

(iii) *$\inf_{\omega\in[0,\pi)} \det\left(\sum_{k=-\infty}^{\infty} e^{-\imath k\omega}\Gamma_0(k)\right) > 0$.*

(iv) *For any fixed integer $k \geq 0$, $T^{-1}\sum_{t=k+1}^{T} y_t y_{t-k}' \xrightarrow{p} \Gamma_0(k)$ as $T \to \infty$.*

The assumption imposes four weak conditions on $\{y_t\}$. First, the time series must be covariance stationary to ensure that the true ACF $\Gamma_0(\cdot)$ is well-defined (as usual, the mean-zero assumption can be easily relaxed). Second, the process is assumed to be weakly dependent, in the sense that the matrix ACF is summable, implying that the spectral density is well-defined. Third, the true spectral density must be uniformly non-singular, meaning that the process has full rank, is strictly nondeterministic, and has a positive definite ACF. Fourth, I assume the weak law of large numbers applies to the sample autocovariances.[56]

---

[55] In the case of i.i.d. data, posterior consistency in misspecified models has been investigated in detail, see Ramamoorthi, Sriram & Martin (2015) and references therein. Shalizi (2009) places high-level assumptions on the prior and likelihood to derive posterior consistency under misspecification with dependent data. Müller (2013) discusses decision theoretic properties of Bayes estimators when the model is mispecified. Tamaki (2008) derives a large-sample Gaussian approximation to the Whittle-based posterior under a correctly specified parametric spectral density and further regularity conditions.

[56] Phillips & Solo (1992) and Davidson (1994, Ch. 19) give sufficient conditions for the law of large numbers for dependent data.

To state the posterior consistency result, I first define the posterior measure. Let

$$\mathbb{T}_{n,q} = \Bigg\{ \{\Gamma(k)\}_{0\leq k\leq q} \colon \Gamma(\cdot) \in \mathbb{R}^{n\times n},\ \Gamma(0) = \Gamma(0)', \\ \inf_{\omega\in[0,\pi)} \det\left(\Gamma(0) + \sum_{k=1}^{q} \{e^{-\imath\omega k}\Gamma(k) + e^{\imath\omega k}\Gamma(k)'\}\right) > 0 \Bigg\}$$

be the space of ACFs for n-dimensional full-rank nondeterministic q-dependent processes. Let $p^W_{Y|\Gamma}(Y_T \mid \Gamma)$ denote the Whittle approximation to the likelihood of a stationary Gaussian process with ACF Γ.[57] Let $\Pi_\Gamma(\cdot)$ be a prior measure on the space $\mathbb{T}_{n,q}$. The associated Whittle posterior measure for $\{\Gamma_0(k)\}_{0\leq k\leq q}$ given the data Y_T is given by

$$P^W_{\Gamma|Y}(\mathcal{A} \mid Y_T) = \frac{\int_{\mathcal{A}} p^W_{Y|\Gamma}(Y_T \mid \Gamma)\Pi_\Gamma(d\Gamma)}{\int_{\mathbb{T}_{n,q}} p^W_{Y|\Gamma}(Y_T \mid \Gamma)\Pi_\Gamma(d\Gamma)}, \tag{1.11}$$

where $\mathcal{A}$ is a measurable subset of $\mathbb{T}_{n,q}$.

Theorem 1.1. *Let Assumption 1.3 hold. Assume that $\{\Gamma_0(k)\}_{0\leq k\leq q}$ is in the support of $\Pi_\Gamma(\cdot)$. Then the Whittle posterior for $\{\Gamma_0(k)\}_{0\leq k\leq q}$ is consistent, i.e., for any neighborhood $\mathcal{U}$ of $\{\Gamma_0(k)\}_{0\leq k\leq q}$ in $\mathbb{T}_{n,q}$, we have*

$$P^W_{\Gamma|Y}(\mathcal{U} \mid Y_T) \xrightarrow{p} 1,$$

as $T \to \infty$ under the true probability measure of the data.

The SVMA model (1.3) and (1.5) is an example of a model with a stationary Gaussian and q-dependent likelihood. Hence, when applied to the SVMA model, Theorem 1.1 states that if the prior measure on the SVMA parameters induces a prior measure on Γ which has the true ACF $\{\Gamma_0(k)\}_{0\leq k\leq q}$ in its support, then the model-implied Whittle posterior for

[57] The precise functional form is stated in Appendix A.1.7.2.

Γ precisely pins down the true ACF in large samples. This result is exploited in the next subsection.

While the measure $P^W_{\Gamma|Y}(\mathcal{A} \mid Y_T)$ is computed using the Whittle likelihood and therefore exploits the working assumption that the data is Gaussian and q-dependent, Theorem 1.1 shows that posterior consistency for the true ACF (up to lag q) holds even for time series that are not Gaussian or q-dependent. The only restrictions placed on the true distribution of the data are the stationarity and weak dependence conditions in Assumption 1.3. Theorem 1.1 is silent about posterior inference on autocovariances at lags higher than q, although the true higher-order autocovariances are allowed to be nonzero.

Theorem 1.1 places no restrictions on the prior $\Pi_\Gamma(\cdot)$ on the ACF, except that the true ACF Γ_0 lies in its support. This level of generality is helpful below when I derive the properties of the SVMA posterior, since no closed-form expression is available for the prior on the ACF that is induced by any given prior on the IRFs and shock standard deviations.

The intermediate results I derive in Appendix A.1.7 to prove Theorem 1.1 may be useful in other contexts. My proof of Theorem 1.1 is based on the general Lemma A.2 in Appendix A.1.7.1, which gives sufficient conditions for posterior consistency in any model, time series or otherwise. Another component of the proof is Lemma A.3, which builds on Dunsmuir & Hannan (1976) to show posterior consistency for the reduced-form (Wold) IRFs in an invertible MA model with q lags, where the posterior is computed using the Whittle likelihood, but the data only has to satisfy Assumption 1.3. No assumptions are placed on the prior, except that the true reduced-form IRFs must be contained in its support.

1.6.3 Limiting posterior distribution in the SVMA model

I finally specialize the general asymptotic results from the previous subsections to the SVMA model with a non-dogmatic prior on IRFs. The asymptotics allow for noninvertibility and non-Gaussian structural shocks. The frequentist large-sample approximation to the Bayesian

posterior shows that the role of the data is to pin down the true autocovariances of the data, which in turn pins down the reduced-form (Wold) IRFs, while all other information about the structural IRFs comes from the prior. I also argue that the limiting form of the posterior is the same whether the Whittle likelihood or the exact likelihood is used.

SET-UP AND MAIN RESULT. To map the SVMA model into the general framework, let θ denote the IRFs and shock standard deviation corresponding to the first shock, and let Γ denote the ACF of the data. That is, $\theta = (\{\Theta_{i1,\ell}\}_{1\leq i\leq n,\, 0\leq \ell\leq q}, \sigma_1)$ and $\Gamma = (\Gamma(0), \ldots, \Gamma(k))$. I now apply Lemma 1.1 and Theorem 1.1 to the SVMA model, which will give a simple description of the limiting form of the Whittle posterior $P^W_{\theta|Y}(\cdot \mid Y_T)$ for all the structural parameters pertaining to the first shock. This analysis of course applies to each of the other shocks.

I choose θ to be the IRFs and shock standard deviation corresponding to a single shock in order to satisfy the prior continuity assumption in Lemma 1.1. In the SVMA model,

$$\Gamma(k) = \sigma_1^2 \sum_{\ell=0}^{q-k} \Theta_{:1,\ell+k}\Theta'_{:1,\ell} + \sum_{j=2}^{n} \sigma_j^2 \sum_{\ell=0}^{q-k} \Theta_{:j,\ell+k}\Theta'_{:j,\ell}, \quad k = 0, 1, \ldots, q, \tag{1.12}$$

where $\Theta_{:j,\ell} = (\Theta_{1j,\ell}, \ldots, \Theta_{nj,\ell})'$. If $\theta = (\{\Theta_{i1,\ell}\}_{1\leq i\leq n,\, 0\leq \ell\leq q}, \sigma_1)$ and there are two or more shocks ($n \geq 2$), then the above equations for $k = 0, 1, \ldots, q$ are of the form $\Gamma = G(\theta) + U$, where $G(\cdot)$ is a matrix-valued function and U is a function only of structural parameters pertaining to shocks $j \geq 2$. θ and U are *a priori* independent provided that the n^2 IRFs and n shock standard deviations are *a priori* mutually independent (for example, the multivariate Gaussian prior in Section 1.2.5 imposes such independence). In this case, the reduced-form parameter Γ equals a function of the structural parameter θ plus *a priori* independent "noise" U. If the prior on the IRFs is non-dogmatic so that U is continuously distributed, we can

expect the conditional prior distribution of θ given Γ to be continuous in Γ.[58]

On the other hand, the conditional prior distribution for θ given Γ would not be continuous in Γ if I had picked θ to be *all* IRFs and shock standard deviations. If $\theta = (\Theta, \sigma)$, then Γ would equal a deterministic function of θ, cf. (1.12), and so continuity of the conditional prior $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma)$ would not obtain. Hence, Lemma 1.1 is not useful for deriving the limit of the *joint* posterior of all structural parameters of the SVMA model.

The main theorem below states the limiting form of the Whittle posterior under general choices for the prior on IRFs and shock standard deviations. That is, I do not assume the multivariate Gaussian prior from Section 1.2.5. I also do not restrict the prior to the region of invertible IRFs, unlike the implicit priors used in SVAR analysis. Let $\Pi_{\Theta,\sigma}(\cdot)$ denote any prior measure for (Θ, σ) on the space $\Xi_\Theta \times \Xi_\sigma$. Through equation (1.7), this prior induces a joint prior measure $\Pi_{\Theta,\sigma,\Gamma}(\cdot)$ on (Θ, σ, Γ), which in turn implies marginal prior measures $\Pi_\theta(\cdot)$ and $\Pi_\Gamma(\cdot)$ for θ and Γ as well as the conditional prior measure $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma)$ for θ given Γ. Let $P^W_{\theta|Y}(\cdot \mid Y_T)$ denote the Whittle posterior measure for θ computed using the Whittle SVMA likelihood, cf. Section 1.3, and the prior $\Pi_{\Theta,\sigma}(\cdot)$.

Theorem 1.2. *Let the data $Y_T = (y_1', \ldots, y_T')'$ be generated from a time series $\{y_t\}$ satisfying Assumption 1.3 (but not necessarily Assumptions 1.1 and 1.2). Assume that the prior $\Pi_{\Theta,\sigma}(\cdot)$ for (Θ, σ) has full support on $\Xi_\Theta \times \Xi_\sigma$. If the induced conditional prior $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma)$ satisfies the continuity assumption (i) of Lemma 1.1, then the Whittle posterior satisfies*

$$\|P^W_{\theta|Y}(\cdot \mid Y_T) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1} \xrightarrow{p} 0,$$

where the stochastic limit is taken as $T \to \infty$ under the true probability measure of the data. If $\hat{\Gamma} = \{\hat{\Gamma}(k)\}_{0 \le k \le q}$ denotes the sample autocovariances up to order q, then the above two

[58] This paragraph is inspired by Remark 3, pp. 769–770, in Moon & Schorfheide (2012).

convergence statements also hold with Γ_0 *replaced by* $\hat{\Gamma}$.[59]

Continuity of the conditional prior $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma)$ is stated as a high-level assumption in Theorem 1.2. I conjecture that prior continuity holds for the multivariate Gaussian prior introduced in Section 1.2.5, for the reasons discussed below equation (1.12), but I have not yet been able to prove this result formally.

An important caveat on the results in this subsection is that the MA lag length q is considered fixed as the sample size T tends to infinity. In applications where q is large relative to T, i.e., when the data is very persistent, these asymptotics may not be a good guide to the finite-sample behavior of the posterior. Nevertheless, the fixed-q asymptotics do shed light on the interplay between the SVMA model, the prior, and the data.[60]

How the data updates the prior. According to Theorem 1.2, the posterior for the structural parameters θ does not collapse to a point asymptotically, but the data does pin down the true ACF Γ_0. Equivalently, the data reveals the true *reduced-form* IRFs and innovation variance matrix, or more precisely, reveals the Wold representation of the observed time series y_t (Hannan, 1970, Thm. $2''$, p. 158). Due to the under-identification of the SVMA model, many different *structural* IRFs are observationally equivalent with a given set of Wold IRFs, cf. Appendix A.1.2. In large samples, the prior is the only source of information able to discriminate between different structural IRFs that are consistent with the true ACF.

Unlike SVARs, the SVMA approach does not infer long-horizon IRFs from short-run dynamics of the data. In large samples the SVMA posterior depends on the data through

---

[59]The proof of Theorem 1.2 follows easily from Lemma 1.1 and Theorem 1.1. $P^W_{\theta|Y}(\cdot \mid Y_T)$ satisfies the general decomposition (1.10) for partially identified models, where the Whittle posterior for Γ has the general form (1.11) for q-dependent Gaussian time series. Theorem 1.1 gives posterior consistency for Γ_0, which is assumption (ii) in Lemma 1.1. Posterior consistency for Γ_0 requires the induced prior measure $\Pi_\Gamma(\cdot)$ to have Γ_0 in its support, which is guaranteed by the assumption of full support for the prior $\Pi_{\Theta,\sigma}(\cdot)$.

[60]I conjecture that my results can be extended to allow for the asymptotic thought experiment $q = q(T) = O(T^\nu)$, for appropriate $\nu > 0$ and under additional nonparametric conditions.

the empirical autocovariances $\hat{\Gamma}$ out to lag q. Inference about long-horizon impulse responses is informed by the empirical autocovariances at the same long horizons (as well as other horizons).[61] In contrast, most SVAR estimation procedures extrapolate long-horizon IRFs from the first few empirical autocorrelations of the data. In this sense, the SVMA approach lets the data influence IRF inference more flexibly than SVAR analysis, although the degree to which the data influences the posterior depends on the prior.

Theorem 1.2 shows to what extent the data can falsify the prior beliefs. The data indicates whether the induced prior $\Pi_\Gamma(\cdot)$ on the ACF is at odds with the true ACF Γ_0.[62] For example, if the prior distribution on IRFs imposes a strong (but non-dogmatic) belief that $\{y_t\}$ is very persistent, but the actual data generating process is not persistent, the posterior will in large samples put most mass on IRFs that imply low persistence, as illustrated in Appendix A.1.5.2. On the other hand, if the prior distribution on IRFs is tightly concentrated around parameters (Θ, σ) that lie in the true identified set $\mathcal{S}(\Gamma_0)$, then the posterior also concentrates around (Θ, σ), regardless of how close (Θ, σ) are to the true structural parameters.

Robustness to misspecified likelihood. Theorem 1.2 states that the posterior measure, which is computed using the Whittle likelihood and thus under the working assumption of a Gaussian SVMA model, converges to $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)$ regardless of whether the Gaussian SVMA model is correctly specified.[63] The only restrictions on the true data generating process are the stationarity and weak dependence conditions in Assumption 1.3. Of course, the IRF parameters only have a structural economic interpretation if the basic SVMA model

[61] The local projection method of Jordà (2005) shares this feature but assumes that shocks are observed.

[62] As a tool for prior elicitation, prior predictive checks can be used to gauge whether the induced prior distribution on the ACF is inconsistent with the observed sample ACF.

[63] Baumeister & Hamilton (2015b) derive an analogous result for Bayesian inference in the SVAR model with a particular family of prior distributions and assuming invertibility.

structure in Assumption 1.1 holds. In this case, the ACF has the form (1.7), so the conditional prior $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)$ imposes valid restrictions on the structural parameters. Thus, under Assumptions 1.1 and 1.3, the large-sample shape of the Whittle SVMA posterior provides valid information about θ even when the shocks are non-Gaussian or heteroskedastic (i.e., $E(\varepsilon_{j,t}^2 \mid \{\varepsilon_s\}_{s<t})$ is non-constant).[64]

The asymptotic robustness to non-Gaussianity of the shocks is a consequence of the negligible importance of the uncertainty surrounding estimation of the true ACF Γ_0. As in the general Lemma 1.1, the latter uncertainty gets dominated in large samples by the conditional prior uncertainty about the structural parameters θ given knowledge of Γ_0. Because the sampling distribution of any efficient estimator of Γ_0 in general depends on fourth moments of the data, the sampling distribution is sensitive to departures from Gaussianity, but this sensitivity does not matter for the first-order asymptotic limit of the posterior for θ.

My results do not and cannot imply that Bayesian inference based on the Gaussian SVMA model is asymptotically equivalent to *optimal* Bayesian inference under non-Gaussian shocks. If the SVMA likelihood were computed under the assumption that the structural shocks ε_t are i.i.d. Student-t distributed, say, then the asymptotic limit of the posterior would differ from $\Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)$. Indeed, if the shocks are known to be non-Gaussian, then higher-order cumulants of the data have identifying power, the empirical ACF does not constitute an asymptotically sufficient statistic for the IRFs, and it may no longer be the case that every invertible set of IRFs can be matched with an observationally equivalent set of noninvertible IRFs (Lanne & Saikkonen, 2013; Gospodinov & Ng, 2015).

However, Bayesian inference based on non-Gaussian shocks is less robust than Gaussian inference. Intuitively, while the expectation of the Gaussian or Whittle (quasi) log likelihood

[64]Standard arguments show that Assumption 1.1 implies Assumption 1.3 under two additional conditions: The true polynomial $\Theta(z)$ cannot have any roots exactly on the unit circle (but the true IRFs may be invertible or noninvertible), and the shocks ε_t must have enough moments to ensure consistency of $\hat{\Gamma}$.

function depends only on second moments of the data, the expectation of a non-Gaussian log likelihood function generally depends also on higher moments. Hence, Bayesian inference computed under non-Gaussian shocks is misleading asymptotically if a failure of the distributional assumptions causes misspecification of higher-order moments, even if second moments are correctly specified.[65] It is an interesting question how to exploit the identifying power of non-Gaussian shocks without unduly compromising computational tractability or robustness to misspecification.

Theorem 1.2 also implies that the error incurred in using the Whittle approximation to the SVMA likelihood is negligible in large samples, in the sense that the data pins down the true ACF in large samples even when the Whittle approximation is used. This is true whether or not the data distribution is the one implied by the Gaussian SVMA model, as long as Assumption 1.3 holds. As discussed in Section 1.3, the reweighting step in Appendix A.1.4.2 therefore makes no difference asymptotically.

1.7 Comparison with SVAR methods

To aid readers who are familiar with SVARs, this section shows that standard SVAR identifying restrictions can be transparently imposed through specific prior choices in the SVMA model, if desired. The SVMA approach easily accommodates exclusion and sign restrictions on short- and long-run impulse responses. External instruments can be exploited in the SVMA framework by expanding the vector of observed time series. Both dogmatic and non-dogmatic prior restrictions are feasible. For extensive discussion of SVAR identification schemes, see Ramey (2016), Stock & Watson (2016), and Uhlig (2015).

The most popular identifying restrictions in the literature are exclusion (i.e., zero) re-

[65] Consider the trivial SVMA model $y_t = \varepsilon_t$, $E(\varepsilon_t^2) = \sigma^2$ ($n = 1$, $q = 0$). It is well known that the Gaussian MLE $\hat{\sigma}^2 = T^{-1} \sum_t y_t^2$ of $\sigma^2 = \Gamma(0)$ enjoys unique robustness properties.

strictions on short-run (i.e., impact) impulse responses: $\Theta_{ij,0} = 0$ for certain pairs (i, j). These short-run exclusion restrictions include so-called "recursive" or "Cholesky" orderings, in which the Θ_0 matrix is assumed triangular. Exclusion restrictions on impulse responses (at horizon 0 or higher) can be incorporated in the SVMA framework by simply setting the corresponding $\Theta_{ij,\ell}$ parameters equal to zero and dropping them from the parameter vector. Prior elicitation and posterior computation for the remaining parameters are unchanged.

Another popular type of identifying restrictions are exclusion restrictions on long-run (i.e., cumulative) impulse responses: $\sum_{\ell=0}^{q} \Theta_{ij,\ell} = 0$ for certain pairs (i, j). Long-run exclusion restrictions can be accommodated in the SVMA model by restricting $\Theta_{ij,q} = -\sum_{\ell=0}^{q-1} \Theta_{ij,\ell}$ when evaluating the likelihood. The first q impulse responses $(\Theta_{ij,0}, \ldots, \Theta_{ij,q-1})$ are treated as free parameters whose prior must be specified by the researcher. When evaluating the score in the HMC procedure, cf. Section 1.3, the chain rule must be used to incorporate the effect that a change in $\Theta_{ij,\ell}$ $(\ell < q)$ has on the implied value for $\Theta_{ij,q}$.

Short- or long-run exclusion restrictions are special cases of linear restrictions on the IRF parameters. Suppose we have prior information that $C \operatorname{vec}(\Theta) = d$, where C is a known full-rank matrix and d is a known vector.[66] Let $C^{\perp}$ be a matrix such that $(C', C^{\perp})$ is a square invertible matrix and $CC^{\perp} = 0$. We can then reparametrize $\operatorname{vec}(\Theta) = C^{\perp}\psi + C'(CC')^{-1}d$, where ψ is an unrestricted vector. Given a prior for ψ,[67] posterior inference in the SVMA model can be carried out as in Section 1.3, except that Θ is treated as a known linear function of the free parameters ψ. Again, the chain rule provides the score with respect to ψ.

The preceding discussion dealt with *dogmatic* prior restrictions that impose exclusion restrictions with 100% prior certainty, but in many cases *non-dogmatic* restrictions are more

---

[66] These restrictions should include the normalizations $\Theta_{i_j j,0} = 1$ for $j = 1, \ldots, n$.

[67] For example, a prior for ψ can be elicited as follows. Elicit a tentative multivariate Gaussian prior for Θ that is approximately consistent with the linear restrictions. Then obtain the prior for ψ from the relationship $\psi = (C^{\perp\prime}C^{\perp})^{-1}C^{\perp\prime}\{\operatorname{vec}(\Theta) - C'(CC')^{-1}d\}$. In general, subtle issues (the Borel paradox) arise when a restricted prior is obtained from an unrestricted one (Drèze & Richard, 1983, Sec. 1.3).

credible.[68] Multivariate Gaussian priors can easily handle non-dogmatic prior restrictions. A prior belief that the impulse response $\Theta_{ij,\ell}$ is close to zero with high probability is imposed by choosing prior mean $\mu_{ij,\ell} = 0$ along with a small value for the prior variance $\tau^2_{ij,\ell}$ (see the notation in Section 1.2.5). To impose a prior belief that the long-run impulse response $\sum_{\ell=0}^{q} \Theta_{ij,\ell}$ is close to zero with high probability, imagine that $\Theta_{ij,q} = -\sum_{\ell=0}^{q-1} \Theta_{ij,\ell} + \nu_{ij}$, where ν_{ij} is mean-zero independent Gaussian noise with a small variance. Given a choice of Gaussian prior for the first q impulse responses $(\Theta_{ij,0}, \ldots, \Theta_{ij,q-1})$, this relationship fully specifies the prior mean vector and covariance matrix of the entire IRF $(\Theta_{ij,0}, \ldots, \Theta_{ij,q})$. These considerations only concern the functional form of the prior density for Θ; evaluation of the likelihood and score is carried out exactly as in Section 1.3.

Many recent SVAR papers exploit sign restrictions on impulse responses: $\Theta_{ij,\ell} \geq 0$ or $\Theta_{ij,\ell} \leq 0$ for certain triplets (i, j, ℓ). Dogmatic sign restrictions can be imposed in the SVMA framework by simply restricting the IRF parameter space Ξ_Θ to the subspace where the inequality constraints hold. This may require some care when running the HMC procedure, but the standard reparametrization $\Theta_{ij,\ell} = \pm \exp\{\log(\pm\Theta_{ij,\ell})\}$ should work (see also Neal, 2011, Sec. 5.1). If the researcher is uncomfortable imposing much more prior information than the sign restrictions, the prior distribution for the impulse responses in question can be chosen to be diffuse (e.g., truncated Gaussian with large variance).[69]

However, researchers often have more prior information about impulse responses than just their signs, and this can be exploited in the SVMA approach.[70] If an impulse response is viewed as being likely to be positive, then very small positive values ought to be less

---

[68]The distinction between dogmatic (exact) and non-dogmatic ("stochastic") identifying restrictions is familiar from the Bayesian literature on simultaneous equation models (Drèze & Richard, 1983).

[69]Poirier (1998, Sec. 4) warns against entirely flat priors in partially identified models.

[70]Similar points have been made in the context of SVARs by Kilian & Murphy (2012), and even more explicitly by Baumeister & Hamilton (2015c). While these papers focus on prior information about impact impulse responses, the SVMA approach facilitates imposing information about longer-horizon responses.

likely than somewhat larger values, up to a point. Additionally, extremely large values for the impulse responses can often be ruled out *a priori*. The multivariate Gaussian prior distribution in Section 1.2.5 is capable of expressing a strong but non-dogmatic prior belief that certain impulse responses have certain signs, while at the same time imposing weak information about magnitudes and ruling out extreme values.[71]

While computationally attractive, the well-known Uhlig (2005) inference procedure for sign-identified SVARs is less transparent than the SVMA approach. Uhlig (2005) uses a conjugate prior on the reduced-form VAR parameters and a uniform conditional prior for the structural parameters given the reduced-form parameters.[72] Because the prior is mainly chosen for computational convenience, it is not very flexible and does not easily allow for non-dogmatic sign restrictions. Furthermore, Baumeister & Hamilton (2015b) show that the Uhlig (2005) procedure imposes unintended and unacknowledged prior information in addition to the acknowledged sign restrictions.[73] In contrast, the SVMA prior is flexible and all restrictions that it imposes can be transparently visualized.

The SVMA approach can exploit the identifying power of external instruments. An external instrument is an observed variable z_t that is correlated with one of the structural shocks but uncorrelated with the other shocks (Stock & Watson, 2008, 2012a; Mertens & Ravn, 2013). If such an instrument is available, it can be incorporated in the analysis by

---

[71] In some applications, a non-symmetric (e.g., log-normal) prior distribution may better express a strong prior belief in the sign of an impulse response, while imposing weak prior restrictions on the magnitude.

[72] Using the notation of Footnote 4 and setting $\Sigma = H \operatorname{diag}(\sigma)^2 H'$, the prior for the reduced-form parameters $(A_1, \ldots, A_m, \Sigma)$ is a normal-inverse-Wishart distribution, while the prior for the orthogonal matrix Q given the reduced-form parameters is Haar measure restricted to the space where the sign restrictions hold. See Arias et al. (2014) for a full discussion. The Uhlig (2005) prior implies a particular informative prior distribution on IRFs, which could in principle be imposed in the SVMA model. Doing so is not desirable, as the main justification for the Uhlig (2005) prior is its convenience in SVAR analysis.

[73] Giacomini & Kitagawa (2015) develop a robust Bayes SVAR approach that imposes dogmatic exclusion and sign restrictions without imposing any other identifying restrictions. Their goals are diametrically opposed to mine, since one of the motivations of the SVMA approach is to allow for as many types of prior information as possible, including information about magnitudes, shapes, and smoothness.

adding z_t to the vector y_t of observed variables. Suppose we add it as the first element ($i = 1$), and that z_t is an instrument for the first structural shock ($j = 1$). The properties of the external instrument then imply that we have a strong prior belief that $\Theta_{1j,0}$ is (close to) zero for $j = 2, 3, \ldots, n$. Depending on the application, we may also have reason to believe that the non-impact impulse responses $\Theta_{1j,\ell}$ are (close to) zero for $\ell \geq 1$. Such prior beliefs can be imposed like any other exclusion restrictions.

The SVMA IRFs can be restricted to be invertible, if desired. As explained in Section 1.2.3, SVARs implicitly assume that the IRFs are invertible, although this is more of a bug than a feature. If, for some reason, the researcher wants to impose invertibility *a priori* in SVMA analysis, simply restrict the IRF parameter space Ξ_Θ to the invertible subspace $\{\Theta\colon \det(\sum_{\ell=0}^{q} \Theta_\ell z^\ell) \neq 0 \; \forall \, z \in \mathbb{C} \text{ s.t. } |z| < 1\}$.[74] Since this represents a nonlinear constraint on the parameter space, it is easiest to carry out posterior inference by first employing the HMC procedure on the unrestricted parameter space and afterwards discarding all posterior draws of Θ that are noninvertible. If the procedure ends up discarding a high fraction of posterior draws, the invertibility restriction is called into question.

1.8 Topics for future research

I conclude by listing some – primarily technical – avenues for future research.

Many SVAR papers seek to identify the IRFs to a single shock (while allowing for other shocks). It would be interesting to investigate whether prior elicitation and Bayesian computations in the SVMA approach can be simplified in the case of single-shock identification.

Whereas the SVMA model in Section 1.2 does not restrict the IRFs at all before prior

[74] If $\det(\Theta_0) = 0$, the IRFs are noninvertible. If $\det(\Theta_0) \neq 0$, the roots of the polynomial $\det(\sum_{\ell=0}^{q} \Theta_\ell z^\ell)$ equal the roots of $\det(I_n + \sum_{\ell=1}^{q} \Theta_0^{-1} \Theta_\ell z^\ell)$. The latter roots can be obtained as reciprocals of the eigenvalues of the polynomial's companion matrix (e.g., Hamilton, 1994, Prop. 10.1).

information is imposed, an alternative modeling strategy is to impose a flexible parametric structure on the SVMA IRFs.[75] Each IRF could be parametrized as a polynomial, trigonometric, or spline function of the horizon, say. The likelihood and score formulas would be essentially unchanged, and computational advantages due to reduced dimensionality may outweigh the loss of flexibility in specifying the IRF prior. Parametrizing the IRFs in this way would even permit an infinite MA lag length ($q = \infty$), while allowing for noninvertibility.

Following the SVAR literature, I assumed that the number of shocks equals the number of observed variables. Yet the SVMA likelihood and Whittle approximation may be evaluated in essentially the same way if the number of shocks exceeds the number of variables. Hence, given a prior on the IRFs, the same Bayesian computations can be applied. The asymptotic analysis also carries through essentially unchanged. However, the characterization of the identified set in Appendix A.1.2 is substantially affected by allowing for more shocks than variables. It would be interesting to extend the identification analysis to the general case.

Also in line with the SVAR literature, this paper focused on a linear stationary model with constant parameters and Gaussian shocks. However, the key advantages of the SVMA approach – the natural IRF parametrization and the ability to allow for noninvertibility – do not rely on these specific assumptions. As long as the likelihood can be evaluated, Bayesian computation is possible in principle. Future work could explore the computational feasibility and robustness properties of SVMA inference that explicitly incorporates nonstationarity, deterministic components, nonlinearities, time-varying parameters, non-Gaussian shocks, or stochastic volatility.[76] The Whittle likelihood approximation used in Section 1.3 may work poorly in non-Gaussian models, but other computational tricks may be available.

Prior knowledge that one of the shocks is more volatile than usual on a known set of

---

[75] This idea is explored by Hansen & Sargent (1981, p. 44) and Barnichon & Matthes (2015).

[76] In the context of cointegrated time series, an analog of the SVMA approach is to do Bayesian inference on the parameters of the Granger representation.

dates can improve identification of the IRFs, as in Rigobon (2003). The SVMA model can be extended along these lines by allowing the standard deviation of the shock in question to switch deterministically between two different values, while assuming constancy of all other structural parameters. It is straight-forward to modify the Kalman filter in Appendix A.1.3.1 to compute the corresponding likelihood function.

A very diffuse prior on IRFs leads to a multimodal posterior distribution due to under-identification. In extreme cases multimodality may cause problems for the basic HMC procedure. It may be possible to exploit the constructive characterization of the identified set in Appendix A.1.2 to extend the algorithm so that it occasionally jumps between approximate modes of the posterior. In case of multimodality, posterior uncertainty should be summarized by highest posterior density intervals instead of equal-tailed (e.g., 5–95 percentile) intervals.

Finally, it may be possible to simplify posterior simulation by leveraging the asymptotic theory in Section 1.6.3. For example, to do exact inference, an approximation to the asymptotic posterior limit could be used as a proposal distribution for importance sampling.

Chapter 2

Consistent Factor Estimation in Dynamic Factor Models with Structural Instability

This paper was coauthored with Brandon J. Bates, James H. Stock & Mark W. Watson. It was published in Journal of Econometrics 177(2), special issue on "Dynamic Econometric Modeling and Forecasting", Timmermann, A. & van Dijk, H. (Eds.), 289–304, December 2013. I thank my coauthors and Elsevier for their permission to reproduce the work here.

2.1 Introduction

Dynamic factor models (DFMs) provide a flexible framework for simultaneously modeling a large number of macroeconomic time series.[1] In a DFM, a potentially large number of observed time series variables are modeled as depending on a small number of unobserved factors, which account for the widespread co-movements of the observed series. Although there is now a large body of theory for the analysis of high-dimensional DFMs, nearly all

[1]The early work on DFMs considered a small number of time series. DFMs were introduced by Geweke (1977), and early low-dimensional applications include Sargent & Sims (1977), Engle & Watson (1981), Watson & Engle (1983), Sargent (1989) and Stock & Watson (1989). Work over the past fifteen years has focused on methods that facilitate the analysis of a large number of time series, see Forni, Hallin, Lippi & Reichlin (2000) and Stock & Watson (2002) for early contributions. For recent contributions and discussions of this large literature see Bai & Ng (2008), Eickmeier & Ziegler (2008), Chudik & Pesaran (2011) and Stock & Watson (2011).

of this theory has been developed for the case in which the DFM parameters are stable, in particular, in which there are no changes in the factor loadings (the coefficients on the factors); among the few exceptions are Stock & Watson (2002, 2009) and Breitung & Eickmeier (2011). This assumption of parameter stability is at odds with broad evidence of time variation in many macroeconomic forecasting relations. Recently, a number of empirical DFM papers have explicitly allowed for structural instability, e.g., Banerjee, Marcellino & Masten (2008), Stock & Watson (2009), Eickmeier, Lemke & Marcellino (2015) and Korobilis (2013). However, theoretical guidance remains scant.

The goal of this paper is to characterize the type and magnitude of parameter instability that can be tolerated by a standard estimator of the factors, the principal components estimator, in a DFM when the coefficients of the model are unstable. In so doing, this paper contributes to a larger debate about how best to handle the instability that is widespread in macroeconomic forecasting relations. On the one hand, the conventional wisdom is that time series forecasts deteriorate when there are undetected structural breaks or unmodeled time-varying parameters, see for example Clements & Hendry (1998). This view underlies the large literatures on the detection of breaks and on models that incorporate breaks and time variation, for example by modeling the breaks as following a Markov process (Hamilton, 1989; Pesaran, Pettenuzzo & Timmermann, 2006). In the context of DFMs, Breitung & Eickmeier (2011) show that a one-time structural break in the factor loadings has the effect of introducing new factors, so that estimation of the factors ignoring the break leads to estimating too many factors.

On the other hand, a few recent papers have provided evidence that sometimes it can be better to ignore parameter instability when forecasting. Pesaran & Timmermann (2005) point out that whether to use pre-break data for estimating an autoregression trades off an increase in bias against a reduction in estimator variance, and they supply empirical evidence supporting the use of pre-break data for forecasting. Pesaran & Timmermann (2007) develop

tools to help ascertain in practice whether pre-break data should be used for estimation of single-equation time series forecasting models. In DFMs, Stock & Watson (2009) provide an empirical example using U.S. macroeconomic data from 1960–2007 in which full-sample estimates of the factors are preferable to subsample estimates, despite clear evidence of a break in many factor loadings around the beginning of the Great Moderation in 1984.

We therefore seek a precise theoretical understanding of the effect of instability in the factor loadings on the performance of principal components estimators of the factors. Specifically, we consider a DFM with N variables observed for T time periods and $r \ll N$ factors, where the $N \times r$ matrix of dynamic factor loadings Λ can vary over time. We write this time variation so that Λ at date t equals its value at date 0, plus a deviation; that is, $\Lambda_t = \Lambda_0 + h_{NT}\xi_t$. The term ξ_t is a possibly random disturbance, and h_{NT} is a deterministic scalar sequence in N and T which governs the scale of the deviation. Using this framework and standard assumptions in the literature (Bai & Ng, 2002, 2006a), we obtain general conditions on h_{NT} under which the principal components estimates are mean square consistent for the space spanned by the true factors. We then specialize these general results to three leading cases: i.i.d. deviations of Λ_t from Λ_0, random walk deviations that are independent across series, and an arbitrary one-time break that affects some or all of the series.

For the case in which Λ_t is a vector of independent random walks, Stock & Watson (2002) showed that the factor estimates are consistent if $h_{NT} = O(T^{-1})$. By using a different method of proof (which builds on Bai & Ng, 2002), we are able to weaken this result considerably and show that the estimated factors are consistent if $h_{NT} = o(T^{-1/2})$. We further show that, if $h_{NT} = O(1/\min\{N^{1/4}T^{1/2}, T^{3/4}\})$, the estimated factors achieve the mean square consistency rate of $1/\min\{N, T\}$, a rate initially established by Bai & Ng (2002) in the case of no time variation. Because the elements of ξ_t in the random walk case are themselves $O_p(t^{1/2})$, this means that deviations in the factor loadings on the order of $o_p(1)$ do not break the consistency of the principal components estimator. These rates are remarkable:

as a comparison, if the factors were observed so an efficient test for time variation could be performed, the test would have nontrivial power against random walk deviations in a $h_{NT} \propto T^{-1}$ neighborhood of zero (e.g., Stock & Watson, 1998b) and would have power of one against parameter deviations of the magnitude tolerated by the principal components estimator. Intuitively, the reason that the principal components estimator can handle such large changes in the coefficients is that, if these shifts have limited dependence across series, their effect can be reduced, and eliminated asymptotically, by averaging across series.

We further provide the rate of mean square consistency as a function of h_{NT}, both in general and specialized to the random walk case. The resulting consistency rate function is nonlinear and reflects the tradeoff between the magnitude of the instability and, through the relative rate N/T as T increases, the amount of cross-sectional information that can be used to "average out" this instability. To elaborate on the practical implications of the theory, we conduct a simulation study calibrated to the Stock & Watson (2009) dataset. The results confirm that the principal components estimator and derived diffusion index forecasts are robust to empirically relevant degrees of temporal instability in the factor loadings, although the precise quantitative conclusions depend on the assumed type of structural instability and the persistence of the factors. Interestingly, the robustness obtains even though the Bai & Ng (2002) information criterion estimator of the rank of the factor space appears to be asymptotically biased for some of our parametrizations.

The rest of the paper proceeds as follows. Section 2.2 lays out the model, the assumptions, and the three special cases. Our main result on consistency of the principal components estimator is presented in Section 2.3. Rank selection and diffusion index forecasting are discussed in Section 2.4. Section 2.5 provides Monte Carlo results, and Section 2.6 concludes.

2.2 Model and assumptions

2.2.1 Basic model and intuition

The model and notation follow Bai & Ng (2002) closely. Denote the observed data by X_{it} for $i = 1, \ldots, N$, $t = 1, \ldots, T$. It is assumed that the observed series are driven by a small, fixed number r of unobserved common factors F_{pt}, $p = 1, \ldots, r$, such that

$$X_{it} = \lambda_{it}' F_t + e_{it}.$$

Here $\lambda_{it} \in \mathbb{R}^r$ is the possibly time-varying factor loading of series i at time t, $F_t = (F_{1t}, \ldots, F_{rt})'$, and e_{it} is an idiosyncratic error. Define vectors $X_t = (X_{1t}, \ldots, X_{Nt})'$, $e_t = (e_{1t}, \ldots, e_{Nt})'$, $\Lambda_t = (\lambda_{1t}, \ldots, \lambda_{Nt})'$ and data matrices $X = (X_1, \ldots, X_T)'$, $F = (F_1, \ldots, F_T)'$. The initial factor loadings Λ_0 are fixed. We write the cumulative drift in the parameter loadings as

$$\Lambda_t - \Lambda_0 = h_{NT} \xi_t,$$

where h_{NT} is a deterministic scalar that may depend on N and T, while $\{\xi_t\}$ is a possibly degenerate random process of dimension $N \times r$, $\xi_t = (\xi_{1t}, \ldots, \xi_{Nt})'$ (in fact, it will be allowed to be a triangular array). Observe that

$$X_t = \Lambda_t F_t + e_t = \Lambda_0 F_t + e_t + w_t, \tag{2.1}$$

where $w_t = h_{NT} \xi_t F_t$. Our proof technique will be to treat w_t as another error term in the factor model.[2]

[2]As pointed out by our referees, a straight-forward approach would be to treat $e_t^* = e_t + w_t$ as a catch-all error term and provide conditions on h_{NT} and ξ_t such that e_t^* satisfies Assumption C in Bai & Ng (2002). Some of the examples below could be handled this way. However, in the case of random walk factor loadings,

To establish some intuition for why estimation of the factors is possible despite structural instability, let the number of factors be $r = 1$ and consider an independent random walk model for the time variation in the factor loadings, so that $\xi_{it} = \xi_{i,t-1} + \zeta_{it}$, where ζ_{it} is i.i.d. across i and t with mean 0 and variance σ_ζ^2, and suppose that Λ_0 is known. In addition, we look ahead to Assumption 2.2 and assume that $\Lambda_0'\Lambda_0/N \to D > 0$. Because Λ_0 is known, we can consider the estimator $\hat{F}_t(\Lambda_0) = (\Lambda_0'\Lambda_0)^{-1}\Lambda_0'X_t$. From (2.1),

$$\hat{F}_t(\Lambda_0) = F_t + (\Lambda_0'\Lambda_0)^{-1}\Lambda_0'e_t + (\Lambda_0'\Lambda_0)^{-1}\Lambda_0'w_t,$$

so

$$\hat{F}_t(\Lambda_0) - F_t \approx D^{-1}N^{-1}\sum_{i=1}^{N}\lambda_{i0}e_{it} + D^{-1}N^{-1}\sum_{i=1}^{N}\lambda_{i0}w_{it}.$$

The first term does not involve time-varying factor loadings and under limited cross-sectional dependence it is $O_p(N^{-1/2})$. Using the definition of w_t, the second term can be written

$$D^{-1}N^{-1}\sum_{i=1}^{N}\lambda_{i0}w_{it} = D^{-1}\left(h_{NT}N^{-1}\sum_{i=1}^{N}\lambda_{i0}\xi_{it}\right)F_t.$$

Since F_t is $O_p(1)$, this second term is the same order as the first, $O_p(N^{-1/2})$, provided that $h_{NT}N^{-1}\sum_{i=1}^{N}\lambda_{i0}\xi_{it}$ is $O_p(N^{-1/2})$. Under the independent random walk model, $\xi_{it} = O_p(T^{1/2})$, so

$$h_{NT}N^{-1}\sum_{i=1}^{N}\lambda_{i0}\xi_{it} = O_p(h_{NT}(T/N)^{1/2}),$$

which in turn is $O_p(N^{-1/2})$ if $h_{NT} = O(T^{-1/2})$. This informal reasoning suggests that the estimator $\hat{F}_t(\Lambda_0)$ satisfies $\hat{F}_t(\Lambda_0) = F_t + O_p(N^{-1/2})$ if $h_{NT} = cT^{-1/2}$.

In practice Λ_0 is not known so $\hat{F}_t(\Lambda_0)$ is not feasible. The principal components esti-

---

applying the Bai & Ng assumption to e_t^* would restrict the temporal dependence of ξ_t more severely than required by our Theorem 2.1 (cf. Assumption 2.3.2 below).

mator of F_t is $\hat{F}_t(\hat{\Lambda}^r)$, where $\hat{\Lambda}^r$ is the matrix of eigenvectors corresponding to the first r eigenvalues of the sample second moment matrix of X_t. The calculations below suggest that the estimation of Λ_0 by $\hat{\Lambda}^r$ reduces the amount of time variation that can be tolerated in the independent random walk case; setting $h_{NT} = cT^{-1/2}$ results in an $O_p(1)$ mean square discrepancy between $\hat{F}_t(\hat{\Lambda}^r)$ and F_t.

2.2.2 Examples of structural instability

For concreteness, we highlight three special cases that will receive extra attention in the following analysis. In these examples, the scalar h_{NT} is left unspecified for now. We will continue to set the number of factors r to 1 for ease of exposition.

Example 1 (white noise). All entries ξ_{it} are i.i.d. across i and t with mean zero and $E(\xi_{it}^4) < \infty$. The factor loadings Λ_t are then equal to the initial loading matrix Λ_0 plus uncorrelated noise.[3]

Example 2 (random walk). Entries ξ_{it} are given by $\xi_{it} = \sum_{s=1}^{t} \zeta_{is}$, where $\{\zeta_{is}\}$ is a random process that is i.i.d. across i and s with mean zero and $E(\zeta_{is}^4) < \infty$. In this example, the factor loadings evolve as cross-sectionally uncorrelated random walks.[4] Models of this type are often referred to as time-varying parameter models in the literature. DFMs with time-varying parameters have recently received attention in the empirical macro literature, cf. Eickmeier et al. (2015), Korobilis (2013) and references therein.

---

[3]As is clear from the subsequent calculations, our conclusions remain true if the disturbances are weakly dependent in the temporal and cross-sectional dimensions. In the interest of clarity we focus on the i.i.d. case.

[4]While conceptually clear, cross-sectional independence of the random walk innovations ζ_{it} is a stricter assumption than necessary for the subsequent treatment. It is straight-forward to modify the example to allow m-dependence or exponentially decreasing correlation across i, and all the results below go through for these modifications.

EXAMPLE 3 (SINGLE LARGE BREAK). Let $\bar{\tau} \in (0,1)$ be fixed and set $\kappa = [\bar{\tau}T]$, where $[\,\cdot\,]$ denotes the integer part. Let $\Delta \in \mathbb{R}^N$ be a shift parameter. We then define

$$\xi_t = \begin{cases} 0 & \text{for } t = 1, \ldots, \kappa \\ \Delta & \text{for } t = \kappa + 1, \ldots, T \end{cases}.$$

Breitung & Eickmeier (2011) demonstrate that a structurally unstable model of this kind may equivalently be written as a stable DFM with $2r$ dynamic factors. Deterministic parameter shifts have also been extensively studied in the context of structural break tests in the linear regression model.

2.2.3 Principal components estimation

We are interested in the properties of the principal components estimator of the factors, where estimation is carried out as if the factor loadings were constant over time. Let k denote the number of factors that are estimated. The principal components estimators of the loadings and factors are obtained by solving the minimization problem

$$V(k) = \min_{\Lambda^k, F^k} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \lambda_i^{k\prime} F_t^k)^2, \tag{2.2}$$

where the superscripts on Λ^k and F^k signify that there are k estimated factors. It is necessary to impose a normalization on the estimators to uniquely define the minimizers (see Bai & Ng, 2008, for a thorough treatment). Such restrictions are innocuous since the unobserved true factors F are only identifiable up to multiplication by a non-singular matrix. One estimator of F is obtained by first concentrating out Λ^k and imposing the normalization $F^{k\prime} F^k / T = I_k$. The resulting estimator $\tilde{F}^k$ is given by $\sqrt{T}$ times the matrix of eigenvectors corresponding to the largest k eigenvalues of the matrix XX'. A second estimator is obtained by first concentrating out F^k and imposing the normalization $\Lambda^{k\prime} \Lambda^k / N = I_k$. This estimator

equals $\bar{F}^k = X\bar{\Lambda}^k/N$, where $\bar{\Lambda}^k$ is $\sqrt{N}$ times the eigenvectors corresponding to the k largest eigenvalues of $X'X$. Following Bai & Ng (2002), we use a rescaled estimator

$$\hat{F}^k = \bar{F}^k(\bar{F}^{k\prime}\bar{F}^k/T)^{1/2}$$

in the following.

2.2.4 Assumptions

Our assumptions on the factors, initial loadings and the idiosyncratic errors are the same as in Bai & Ng (2002). The matrix norm is chosen to be the Frobenius norm $\|A\| = [\mathrm{tr}(A'A)]^{1/2}$. The subscripts i, j will denote cross-sectional indices, s, t will denote time indices and p, q will denote factor indices. $M \in (0, \infty)$ is a constant that is common to all the assumptions below. Finally, define $C_{NT} = \min\{N^{1/2}, T^{1/2}\}$. The following are Assumptions A–C in Bai & Ng (2002).

Assumption 2.1 (Factors)**.** *$E\|F_t\|^4 \leq M$ and $T^{-1}\sum_{t=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F$ as $T \to \infty$ for some positive definite matrix Σ_F.*

Assumption 2.2 (Initial factor loadings)**.** *$\|\lambda_{i0}\| \leq \bar{\lambda} < \infty$, and $\|\Lambda_0'\Lambda_0/N - D\| \to 0$ as $N \to \infty$ for some positive definite matrix $D \in \mathbb{R}^{r\times r}$.*

Assumption 2.3 (Idiosyncratic errors)**.** *The following conditions hold for all N and T.*

1. *$E(e_{it}) = 0$, $E|e_{it}|^8 \leq M$.*
2. *$\gamma_N(s,t) = E(e_s'e_t/N)$ exists for all (s,t). $|\gamma_N(s,s)| \leq M$ for all s. In addition, $T^{-1}\sum_{s,t=1}^{T} |\gamma_N(s,t)| \leq M$.*
3. *$\tau_{ij,ts} = E(e_{it}e_{js})$ exists for all (i,j,s,t). $|\tau_{ij,tt}| \leq |\tau_{ij}|$ for some τ_{ij} and for all t, while $N^{-1}\sum_{i,j=1}^{N} |\tau_{ij}| \leq M$. In addition, $(NT)^{-1}\sum_{i,j=1}^{N}\sum_{s,t=1}^{T} |\tau_{ij,ts}| \leq M$.*

4. For every (s,t), $E|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]|^4 \leq M$.

As mentioned by Bai & Ng (2002), the above assumptions allow for weak cross-sectional and temporal dependence of the idiosyncratic errors. Note that the factors do not need to be stationary to satisfy Assumption 2.1.

The assumptions we need on the factor loading innovations $h_{NT}\xi_t$ are summarized below. For now we require the existence of three envelope functions that bound the rates, in terms of N and T, at which certain sums of higher moments diverge. Their interpretation will be made clear in examples below. As we later state in Theorem 2.1, these rates determine the convergence rate of the principal components estimator of the factors.

Assumption 2.4 (Factor loading innovations)**.** *There exist envelope functions* $Q_1(N,T)$*,* $Q_2(N,T)$ *and* $Q_3(N,T)$ *such that the following conditions hold for all* N*,* T *and factor indices* $p_1, q_1, p_2, q_2 = 1, \ldots, r$.

1. $\sup_{s,t\leq T}\sum_{i,j=1}^{N}|E(\xi_{isp_1}\xi_{jtq_1}F_{sp_1}F_{tq_1})| \leq Q_1(N,T)$.

2. $\sum_{s,t=1}^{T}\sum_{i,j=1}^{N}|E(\xi_{isp_1}\xi_{jsq_1}F_{sp_1}F_{sq_1}F_{tp_2}F_{tq_2})| \leq Q_2(N,T)$.

3. $\sum_{s,t=1}^{T}\sum_{i,j=1}^{N}|E(\xi_{isp_1}\xi_{jsq_1}\xi_{itp_2}\xi_{jtq_2}F_{sp_1}F_{sq_1}F_{tp_2}F_{tq_2})| \leq Q_3(N,T)$.

While consistency of the principal components estimator will require limited dependence between the factor loading innovations and the factors themselves, full independence is not necessary. This is empirically appealing, as it is reasonable to expect that breaks in the factor relationships may occur at times when the factors deviate substantially from their long-run means. That being said, we remark that if the processes $\{\xi_t\}$ and $\{F_t\}$ are assumed to be independent (and given Assumption 2.1), two sufficient conditions for Assumption 2.4 are that there exist envelope functions $\tilde{Q}_1(N,T)$ and $\tilde{Q}_3(N,T)$ such that for all factor indices,

$$\sup_{s,t\leq T}\sum_{i,j=1}^{N}|E(\xi_{isp_1}\xi_{jtq_1})| \leq \tilde{Q}_1(N,T) \tag{2.3}$$

and

$$\sum_{s,t=1}^{T}\sum_{i,j=1}^{N}|E(\xi_{isp_1}\xi_{jsq_1}\xi_{itp_2}\xi_{jtq_2})| \leq \tilde{Q}_3(N,T). \tag{2.4}$$

Under the above conditions, Assumption 2.4 holds with $Q_1(N,T) \propto \tilde{Q}_1(N,T)$, $Q_2(N,T) \propto T^2\tilde{Q}_1(N,T)$ and $Q_3(N,T) \propto \tilde{Q}_3(N,T)$.

Finally, rather than expanding the list of moment conditions in Assumption 2.4, we simply impose independence between the idiosyncratic errors and the other variables. It is possible to relax this assumption at the cost of added complexity.[5]

Assumption 2.5 (Independence)**.** *For all (i,j,s,t), e_{it} is independent of (F_s, ξ_{js}).*

Examples (continued). For Examples 1 and 2 (white noise and random walk), assume that $\{\xi_t\}$ and $\{F_t\}$ are independent.

In Example 1 (white noise), the supremum on the left-hand side of (2.3) reduces to $NE(\xi_{it}^2)$. By writing out terms, it may be verified that the quadruple sum in condition (2.4) is bounded by an $O(NT^2) + O(N^2T)$ expression. Consequently, Assumption 2.4 holds with $Q_1(N,T) = O(N)$, $Q_2(N,T) = O(NT^2)$ and $Q_3(N,T) = O(NT^2) + O(N^2T)$.

In Example 2 (random walk), due to cross-sectional i.i.d.-ness we obtain

$$\begin{aligned}\sup_{s,t\leq T}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(\xi_{is}\xi_{jt})| &= N\sup_{s,t\leq T}|E(\xi_{is}\xi_{it})| \\ &= N\sup_{s,t\leq T}\min\{s,t\}E(\zeta_{i1}^2) \\ &= O(NT),\end{aligned}$$

so Assumption 2.4.1 and 2.4.2 hold with $Q_1(N,T) = O(NT)$ and $Q_2(N,T) = O(NT^3)$. A somewhat lengthier calculation gives that the quadruple sum in condition (2.4) is $O(N^2T^4)$,

[5]Bai & Ng (2006a) impose independence of $\{e_t\}$ and $\{F_t\}$ when providing inferential theory for regressions involving estimated factors.

so Assumption 2.4.3 holds with $Q_3(N,T) = O(N^2T^4)$.

In Example 3 (single large break), the supremum in inequality (2.3) evaluates as

$$\sum_{i=1}^{N} |\Delta_i| \sum_{j=1}^{N} |\Delta_j|.$$

Assume that $|\Delta_i| \leq M$ for some $M \in (0, \infty)$ that does not depend on N. We note for later reference that if $|\Delta_i| > 0$ for at most $O(N^{1/2})$ values of i, the expression above is $O(N)$. The same condition ensures that the left-hand side of condition (2.4) is $O(NT^2)$. Consequently, we can choose $Q_1(N,T) = O(N)$ and $Q_2(N,T) = Q_3(N,T) = O(NT^2)$ if at most $O(N^{1/2})$ series undergo a break.

2.3 Consistent estimation of the factor space

2.3.1 Main result

Our main result provides the mean square convergence rate of the usual principal components estimator under Assumptions 2.1 to 2.5. After stating the general theorem, we give sufficient conditions that ensure the same convergence rate that Bai & Ng (2002) obtained in a setting with constant factor loadings.

Theorem 2.1. *Let Assumptions 2.1 to 2.5 hold. For any fixed k,*

$$T^{-1} \sum_{t=1}^{T} \|\hat{F}_t^k - H^{k'} F_t\|^2 = O_p(R_{NT})$$

as $N, T \to \infty$, where

$$R_{NT} = \max \left\{ \frac{1}{C_{NT}^2}, \frac{h_{NT}^2}{N^2} Q_1(N,T), \frac{h_{NT}^2}{N^2T^2} Q_2(N,T), \frac{h_{NT}^4}{N^2T^2} Q_3(N,T) \right\},$$

and the $r \times k$ matrix H^k is given by

$$H^k = (\Lambda_0'\Lambda_0/N)(F'\tilde{F}^k/T).$$

See Appendix B.2 for the proof. If $R_{NT} \to 0$ as $N, T \to \infty$, the theorem implies that the r-dimensional space spanned by the true factors is estimated consistently in mean square (averaging over time) as $N, T \to \infty$. While we do not discuss it here, a similar statement concerning pointwise consistency of the factors (Bai & Ng, 2002, p. 198) may be achieved by slightly modifying Assumptions 2.3 and 2.4.

We now give sufficient conditions on the envelope functions in Assumption 2.4 such that the principal components estimator achieves the same convergence rate as in Theorem 1 of Bai & Ng (2002). This rate, C_{NT}^2, turns out to be central for other results in the literature on DFMs (Bai & Ng, 2002, 2006a). The following corollary is a straight-forward consequence of Theorem 2.1.

Corollary 2.1. *Under the assumptions of Theorem 2.1, and if additionally*

- $h_{NT}^2 Q_1(N,T) = O(N)$,
- $h_{NT}^2 Q_2(N,T) = O(NT^2)$,
- $h_{NT}^4 C_{NT}^2 Q_3(N,T) = O(N^2T^2)$,

it follows that, as $N, T \to \infty$,

$$C_{NT}^2 \left(T^{-1} \sum_{t=1}^{T} \|\hat{F}_t^k - H^{k'} F_t\|^2 \right) = O_p(1).$$

Examples (continued). In Section 2.2.4 we computed the envelope functions $Q_1(N,T)$, $Q_2(N,T)$ and $Q_3(N,T)$ for our three examples. From these calculations we note that if $h_{NT} = 1$, the model in Example 1 (white noise) satisfies the conditions of Corollary 2.1.

Hence, uncorrelated order-$O_p(1)$ white noise disturbances in the factor loadings do not affect the consistency of the principal components estimator.

Likewise, it follows from our calculations that the structural break process in Example 2 (random walk) satisfies the conditions of Corollary 2.1 if $h_{NT} = O(1/\min\{N^{1/4}T^{1/2}, T^{3/4}\})$. Moreover, a rate of $h_{NT} = o(T^{-1/2})$ is sufficient to achieve $R_{NT} = o(1)$ in Theorem 2.1, i.e., that the factor space is estimated consistently. This is a weaker rate requirement than the $O(T^{-1})$ scale factor imposed by Stock & Watson (2002).[6] To elaborate on the convergence rate in Theorem 2.1, suppose we set $N = [T^{\mu}]$ and $h_{NT} = cT^{-\gamma}$, $\mu, \gamma \geq 0$. Using the formula for R_{NT} and the random walk calculations in Section 2.2.4, we obtain

$$R_{NT} = O(\max\{T^{-1}, T^{-\mu}, T^{1-2\gamma-\mu}, T^{2-4\gamma}\}) = O(T^{m(\mu,\gamma)}), \tag{2.5}$$

where

$$m(\mu, \gamma) = \max\{-1, -\mu, 1-2\gamma-\mu, 2-4\gamma\} = \max\{-1, -\mu, 2-4\gamma\}. \tag{2.6}$$

This convergence rate exponent reflects the influence of the magnitude of the random walk deviations, as measured by γ, and the relative sizes of the cross-sectional and temporal dimensions, as measured by μ. Evidently, increasing the number of available series relative to the sample size improves the worst-case convergence rate, but only up to a point. The dependence of the convergence rate on γ is monotonic, as expected, but nonlinear.

For Example 3 (single large break), Corollary 2.1 and our calculations in Section 2.2.4 yield that if we set $h_{NT} = 1$, the principal components estimator achieves the Bai & Ng (2002) convergence rate, provided at most $O(N^{1/2})$ series undergo a break. A fraction $O(N^{-1/2})$ of the series may therefore experience an order-$O(1)$, perfectly correlated shift in their factor

[6]Empirical implementations of principal components estimation of structurally unstable DFMs, such as Eickmeier et al. (2015) and Korobilis (2013), rely on robustness of the estimator to small degrees of instability. Our theorem shows that the asymptotically allowable amount of instability is larger than hitherto assumed.

loadings without affecting the consistency of the estimator.

2.3.2 Detailed calculations for special cases

Theorem 2.1 shows the convergence rate of the principal components estimator but does not offer any information on the constant of proportionality, which in general will depend on the size of the various moments in Assumptions 2.1 to 2.4. In this subsection we consider examples in which we can say more about the speed of convergence.

For analytical tractability, we assume in this subsection that the initial factor loadings Λ_0 are 0 and the true number of factors r is 1. When $\Lambda_0 = 0$, the matrix H^k in Theorem 2.1 is equal to zero, so that consistency of the principal components estimator hinges on how fast the norm of $\hat{F}_t^k$ tends to zero in mean square.[7] As shown in Appendix A.2.1, when $\Lambda_0 = 0$ and $r = 1$,

$$T^{-1}\sum_{t=1}^{T} \|\hat{F}_t^k - H^{k\prime}F_t\|^2 = (NT)^{-2}\sum_{l=1}^{k}\omega_l^2,$$

where ω_l is the l-th largest eigenvalue of the $T \times T$ matrix XX'.

Example 1 (white noise, continued). Suppose the single factor is identically 1 ($F_t \equiv 1$), and N and T tend to infinity at the relative rate $\theta = \lim_{N\to\infty} T/N$, $\theta \in (0, \infty)$. Let the idiosyncratic errors e_{it} be i.i.d. across i and t with $E(e_{it}^2) = \sigma_e^2$. Denote $\sigma_\xi^2 = E(\xi_{it}^2)$. Appendix A.2.1 shows that if the number of estimated factors is $k = 1$, then

$$T^{-1}\sum_{t=1}^{T} \|\hat{F}_t^k - H^{k\prime}F_t\|^2 = T^{-2}(\sigma_e^2 + h_{NT}^2\sigma_\xi^2)^2(1+\sqrt{\theta})^4(1+o_p(1)). \tag{2.7}$$

When $h_{NT} = 1$, the right-hand side quantity is $O_p(T^{-2})$, which is stronger than the $O_p(C_{NT}^{-2})$ rate bound in Theorem 2.1. Introducing cross-sectional and temporal dependence in the

[7] Note that while $\Lambda_0 = 0$ violates Assumption 2.2, the proof of Theorem 2.1 does not rely on the matrix $D = \text{plim } \Lambda_0'\Lambda_0/N$ being positive definite.

idiosyncratic errors causes the left-hand side above to achieve the worst-case rate asymptotically, as noted by Bai & Ng (2002, pp. 199–200). According to the expression on the right-hand side of equation (2.7), h_{NT}^2 measures the importance of the factor loading disturbance variance relative to the idiosyncratic error variance. Furthermore, for given T, the mean square error of the principal components estimator increases with the ratio $\theta \approx T/N$.

Example 2 (random walk, continued). Suppose that the idiosyncratic errors are cross-sectionally i.i.d. Denote $\sigma_\zeta^2 = E(\zeta_{it}^2)$. If the number of estimated factors is $k = 1$, we show in Appendix A.2.1 that

$$E\left(T^{-1}\sum_{t=1}^{T}\|\hat{F}_t^k - H^{k\prime}F_t\|^2\right) \geq \left\{T^{-2}\sum_{s,t=1}^{T}\left[\gamma_N(s,t) + h_{NT}^2\sigma_\zeta^2\min\{s,t\}E(F_sF_t)\right]\right\}^2, \quad (2.8)$$

where $\gamma_N(s,t)$ is defined in Assumption 2.3. This lower bound on the expectation of the mean square error of the principal components estimator complements the upper rate bound in Theorem 2.1. The expression reinforces the intuition that the factor space will be poorly estimated in models with persistent errors (here e_{it} and $h_{NT}\xi_{it}'F_t$).

Without any prior knowledge about the factor process, a conservative benchmark sets $E(F_sF_t) = O(1)$. Note that $\sum_{s,t=1}^{T}\min\{s,t\} = \frac{1}{3}T^3 + O(T^2)$, and $\sum_{s,t=1}^{T}\gamma_N(s,t) = O(T)$ by Assumption 2.3. If $h_{NT} \geq T^{-1}$ asymptotically, the right-hand side of inequality (2.8) is then of order $h_{NT}^4T^2$. Together with Theorem 2.1, this establishes that there exist constants $\underline{C}, \overline{C} > 0$ such that

$$\underline{C} \leq (h_{NT}^2T)^{-2}E\left(T^{-1}\sum_{t=1}^{T}\|\hat{F}_t^k - H^{k\prime}F_t\|^2\right) \leq \overline{C}\max\{(h_{NT}^2TC_{NT})^{-2}, 1\}$$

for sufficiently large N and T.[8] The maximum on the right-hand side above tends to 1 as long as $h_{NT} \geq (TC_{NT})^{-1/2} = 1/\min\{N^{1/4}T^{1/2}, T^{3/4}\}$ asymptotically.[9] Thus, unless we have special knowledge about the factor process, we generically need $h_{NT} = o(T^{-1/2})$ for mean square consistency of the factors, while $h_{NT} = O(1/\min\{N^{1/4}T^{1/2}, T^{3/4}\})$ is generically necessary to achieve the Bai & Ng (2002) convergence rate.

Example 3 (single large break, continued). Here we consider a limiting case with $e_{it} \equiv 0$, so that all the variance in the observed data is due to structural instability. Suppose the single factor F_t satisfies $(T-\kappa)^{-1}\sum_{t=\kappa+1}^{T} F_t^2 \xrightarrow{p} \tilde{\Sigma}_F$ as $T \to \infty$. Then, regardless of the number of estimated factors k,

$$T^{-1}\sum_{t=1}^{T} \|\hat{F}_t^k - H^{k\prime}F_t\|^2 = \frac{h_{NT}^4}{N^2}\|\Delta\|^4(1-\bar{\tau})^2\tilde{\Sigma}_F^2(1+o_p(1)), \tag{2.9}$$

as shown in Appendix A.2.1. The result indicates that the mean square error of the principal components estimator is larger the smaller is $\bar{\tau}$ (the break fraction), the larger is $\tilde{\Sigma}_F$ (the post-break factor second moment), and the larger is $\|\Delta\|$ (the size of the break vector). Note that if the elements of Δ are uniformly bounded, $|\Delta_i| \leq M$, then $\|\Delta\|^2$ is on the order of the number of series undergoing a break. Denote this number by B_{NT}. The right-hand side above is then $O_p\left((h_{NT}^2 B_{NT}/N)^2\right)$, which is also the rate stated in the bound in Theorem 2.1, provided that $h_{NT} = 1$.

---

[8] The rate bound in Theorem 2.1 is in probability, but the proof given in Appendix B.2 shows that the bound holds in expectation as well.

[9] Recall that such rates for h_{NT} are exactly the ones we are most interested in, since any faster rate of decay for h_{NT} will lead to $R_{NT} = C_{NT}^{-2}$ in Theorem 2.1.

2.4 Rank selection and diffusion index forecasting

2.4.1 Estimating the number of factors

Bai & Ng (2002, 2006b) introduce a class of information criteria that consistently estimate the true number r of factors when the factor loadings are constant through time. Specifically, define the two classes of criteria

$$PC(k) = V(k) + kg(N,T), \quad IC(k) = \log V(k) + kg(N,T), \tag{2.10}$$

where $V(k)$ is the sum of squared residuals defined by (2.2), and $g(N,T)$ is a deterministic function satisfying $g(N,T) \to 0$, $C_{NT}^2 g(N,T) \to \infty$ as $N,T \to \infty$. Let $k_{\max} \geq r$ be an upper bound on the estimated rank. With constant factor loadings, a consistent estimate of r is then given by either $\hat{k} = \arg\min_{0\leq k\leq k_{\max}} PC(k)$ or $\hat{k} = \arg\min_{0\leq k\leq k_{\max}} IC(k)$.

Lemma 2 of Amengual & Watson (2007) establishes that these information criteria remain consistent for r when the data X are measured with an additive error, i.e., if the researcher instead observes $\tilde{X} = X + b$ for a $T \times N$ error matrix b that satisfies $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} b_{it}^2 = O_p(C_{NT}^{-2})$. By our decomposition (2.1) of X_t, time variation in the factor loadings may be seen as contributing an extra error term w_t to the usual terms $\Lambda_0 F_t + e_t$. The following result is therefore a direct consequence of Lemma 2 of Amengual & Watson (2007) and Markov's inequality.

Observation 2.1. *Let assumptions (A1)–(A9) in Amengual & Watson (2007) hold. If in addition*

$$h_{NT}^2 \sum_{i=1}^{N}\sum_{t=1}^{T} E[(\xi_{it}' F_t)^2] = O(\max\{N,T\}), \tag{2.11}$$

then $\arg\min_{0\leq k\leq k_{\max}} PC(k) \xrightarrow{p} r$ *and* $\arg\min_{0\leq k\leq k_{\max}} IC(k) \xrightarrow{p} r$ *as* $N,T \to \infty$.

In the interest of brevity we do not state the precise Amengual & Watson conditions here

but remark that they are very similar to our Assumptions 2.1 to 2.3 and 2.5. We now comment on how the sufficient condition in Observation 2.1 bears on our three examples of structural breaks. The finite-sample performance of the information criteria will be explored in Section 2.5.

EXAMPLES (CONTINUED). If $r = 1$ and ξ_{it} is independent of F_t, the left-hand side of condition (2.11) is of order $h_{NT}^2 \sum_{i=1}^N \sum_{t=1}^T E(\xi_{it}^2)$. In Example 1 (white noise), $\sum_{i=1}^N \sum_{t=1}^T E(\xi_{it}^2) = O(NT)$, so condition (2.11) holds if $h_{NT} = O(C_{NT}^{-1})$. The white noise disturbances must therefore vanish asymptotically, albeit slowly, for the Amengual & Watson (2007) result to ensure consistent estimation of the factor rank.

For Example 2 (random walk), $\sum_{i=1}^N \sum_{t=1}^T E(\xi_{it}^2) = O(NT^2)$, implying that we need $h_{NT} = O(1/\min\{T, (NT)^{1/2}\})$ to fulfill condition (2.11). In particular, the Stock & Watson (2002) assumption $h_{NT} = O(1/T)$ admits consistent estimation of the true number of factors using the Bai & Ng (2002) information criteria.

In Example 3 (single large break), we set $h_{NT} = 1$ as before. If $(T-\kappa)^{-1} \sum_{t=\kappa+1}^T E(F_t^2) = O(1)$, we get $\sum_{i=1}^N \sum_{t=1}^T E(\xi_{it}^2) = O(T\|\Delta\|^2)$, so $\|\Delta\|^2 = O(\max\{N/T, 1\})$ is needed to satisfy condition (2.11). As previously explained, if the elements of Δ are uniformly bounded, $\|\Delta\|^2$ is on the order of the number B_{NT} of series undergoing a break at time $t = \kappa + 1$. The fraction B_{NT}/N of series undergoing a break must therefore be of order at most C_{NT}^{-2} for the Amengual & Watson (2007) result to apply. The conclusion that large breaks are more problematic for rank estimation than for mean square consistency is not surprising given Breitung & Eickmeier's (2011) insight that the large break model (with non-vanishing break parameter) is equivalent to a DFM with $2r$ factors.

In summary, in all three of our examples we need more stringent assumptions on h_{NT} in order to ensure consistent estimation of r than we did for consistency of the principal components estimator. It is a topic for future research to determine whether these tentative

results can be improved upon.

2.4.2 Diffusion index forecasting

As an application of Corollary 2.1, consider the diffusion index model of Stock & Watson (1998a, 2002) and Bai & Ng (2006a). For ease of exposition we assume that the factors are the only explanatory variables, so the model is

$$y_{t+h} = \alpha' F_t + \varepsilon_{t+h}.$$

Here y_{t+h} is the scalar random variable that we seek to forecast, while ε_{t+h} is an idiosyncratic forecast error term that is independent of all other variables. We shall assume that the true number of factors r is known. Because the true factors F_t are not observable, one must forecast y_{t+h} using the estimated factors $\hat{F}_t$. Does the sampling variability in $\hat{F}$ influence the precision and asymptotic normality of the feasible estimates of α?

Let $\hat{F}$ be the principal components estimator with $k = r$ factors estimated and denote the $r \times r$ matrix H^r from Theorem 2.1 by H. Define $\delta = H^{-1}\alpha$ (note that due to the factors being unobservable, α is only identified up to multiplication by a nonsingular matrix) and let $\hat{\delta}$ be the least squares estimator in the feasible diffusion index regression of y_{t+h} on $\hat{F}_t$. Bai & Ng (2006a) show that

$$\sqrt{T}(\hat{\delta} - \delta) = (T^{-1}\hat{F}'\hat{F})^{-1}T^{-1/2}\hat{F}'\varepsilon - (T^{-1}\hat{F}'\hat{F})^{-1}[T^{-1/2}\hat{F}'(\hat{F} - FH)]H^{-1}\alpha, \qquad (2.12)$$

where $\varepsilon = (\varepsilon_{1+h}, \ldots, \varepsilon_{T+h})'$. Under the assumptions of Corollary 2.1, the Cauchy-Schwarz inequality yields

$$\|T^{-1/2}\hat{F}'(\hat{F} - FH)\|^2 \leq T\left(T^{-1}\sum_{t=1}^{T}\|\hat{F}_t\|^2\right)\left(T^{-1}\sum_{t=1}^{T}\|\hat{F}_t - H'F_t\|^2\right)$$

$$= TO_p(1)O_p(C_{NT}^{-2})$$
$$= O_p(\max\{1, T/N\}).$$

Similarly,

$$T^{-1/2}\hat{F}'\varepsilon = T^{-1/2}H'F'\varepsilon + T^{-1/2}(\hat{F} - FH)'\varepsilon = T^{-1/2}H'F'\varepsilon + O_p(\max\{1, T/N\}).$$

Suppose $T^{-1/2}F'\varepsilon = O_p(1)$, as implied by Assumption E in Bai & Ng (2006a). It is easy to show that $H = O_p(1)$. Provided $T = O(N)$, we thus obtain $\hat{\delta} - \delta = O_p(T^{-1/2})$, i.e., under the conditions of Corollary 2.1, the feasible diffusion regression estimator is consistent at the usual rate. The restrictions on h_{NT} for the three examples are discussed immediately following Corollary 2.1.[10]

2.5 Simulations

2.5.1 Design

To illustrate our results and assess their finite sample validity we conduct a Monte Carlo simulation study. Stock & Watson (2002) and Eickmeier et al. (2015) numerically evaluate the performance of the principal components estimator when the factor loadings evolve as random walks, and Banerjee et al. (2008) focus in particular on the effect of time variation in short samples. We provide additional evidence on the necessary scale factor h_{NT} for

[10] If $\alpha = 0$, which is often an interesting null hypothesis in applied work, the second term on the right-hand side of the decomposition (2.12) vanishes. Assume that $\{\varepsilon_{t+h}\}$ is independent of all other variables. Then, conditional on $\hat{F}$, the first term on the right-hand side of (2.12) will (under weak conditions) obey a central limit theorem, and so $\hat{\delta}$ should be unconditionally asymptotically normally distributed under the null $H_0\colon \alpha = 0$. Bai & Ng (2006a) prove that if the factor loadings are not subject to time variation, $\hat{\delta}$ will indeed be asymptotically normal, regardless of the true value of α, as long as $\sqrt{T}/N \to 0$. We expect that a similar result can be proved formally in our framework but leave this for future research.

the random walk case (our Example 2). Moreover, we consider data generating processes (DGPs) in which the factor loadings are subject to white noise disturbances (as in Example 1), as well as DGPs for which a subset of the series undergo one large break in their factor loadings (an analog of Example 3).

The design broadly follows Stock & Watson (2002):

$$X_{it} = \lambda_{it}' F_t + e_{it}, \quad F_{tp} = \rho F_{t-1,p} + u_{tp}, \quad (1 - aL)e_{it} = v_{it}, \quad y_{t+1} = \sum_{q=1}^{r} F_{tq} + \varepsilon_{t+1},$$

where $i = 1, \ldots, N$, $t = 1, \ldots, T$, $p = 1, \ldots, r$. The processes $\{u_{tp}\}$, $\{v_{it}\}$ and $\{\varepsilon_{t+1}\}$ are mutually independent, with u_{tp} and ε_{t+1} being i.i.d. standard normally distributed. To capture cross-sectional dependence of the idiosyncratic errors, we let $v_t = (v_{1t}, \ldots, v_{Nt})'$ be i.i.d. normally distributed with covariance matrix $\Omega = (\beta^{|i-j|})_{ij}$, as in Amengual & Watson (2007). The scalar ρ is the common AR(1) coefficient for the r factors, while a is the AR(1) coefficient for the idiosyncratic errors.

The initial values F_0 and e_0 for the factors and idiosyncratic errors are drawn from their respective stationary distributions. The initial factor loading matrix Λ_0 was chosen based on the population R^2 for the regression of $X_{i0} = \lambda_{i0}' F_0 + e_{i0}$ on F_0. Specifically, for each i we draw a value R_i^2 uniformly at random from the interval $[0, 0.8]$. We then set $\lambda_{i0p} = \lambda^*(R_i^2)\bar{\lambda}_{i0j}$, where $\bar{\lambda}_{i0j}$ is i.i.d. standard normal and independent of all other disturbances.[11] The scalar $\lambda^*(R_i^2)$ is given by the value for which $E[(\lambda_{0i}' F_0)^2 | R_i^2] / E[X_{i0}^2 | R_i^2] = R_i^2$, given the draw of R_i^2.[12]

We consider three different specifications for the evolution of factor loadings over time.

---

[11] We assumed above that Λ_0 is fixed for simplicity. It is not difficult to verify that Λ_0 could instead be random, provided that it is independent of all other random variables, $N^{-1}\Lambda_0'\Lambda_0 \xrightarrow{p} D$ for an $r \times r$ non-singular matrix D, and $E\|\lambda_i\|^4 < M$, as in Bai & Ng (2006a).

[12] Specifically, $[\lambda^*(R_i^2)]^2 = \frac{1-\rho^2}{r(1-a^2)} \frac{R_i^2}{1-R_i^2}$.

In the *white noise model* the loadings are given by

$$\lambda_{itp} = \lambda_{i0p} + d\xi_{itp},$$

$i = 1, \ldots, N$, $t = 1, \ldots, T$, $p = 1, \ldots, r$, where d is a constant and the disturbances ξ_{itp} are i.i.d. standard normal and independent of all other disturbances. Note that the standard deviation of $\lambda_{itp} - \lambda_{i0p}$ is d for all t.

In the *random walk model* we set

$$\lambda_{itp} = \lambda_{i,t-1,p} + cT^{-3/4}\zeta_{itp},$$

$i = 1, \ldots, N$, $t = 1, \ldots, T$, $p = 1, \ldots, r$, where c is a constant and the innovations ζ_{itp} are i.i.d. standard normal and independent of everything. Note that the $T^{3/4}$ rate is different from the rate of T used by Stock & Watson (2002) and Banerjee et al. (2008). In our design, the standard deviation of $\lambda_{iTp} - \lambda_{i0p}$ is $cT^{-1/4}$.

In the *large break model* we select a subset J of size $[bN^{1/2}]$ uniformly at random from the integers $\{1, \ldots, N\}$, where b is a constant. For $i \notin J$, we simply let $\lambda_{itp} = \lambda_{i0p}$ for all t. For $i \in J$, we set

$$\lambda_{itp} = \begin{cases} \lambda_{i0p} & \text{for } t \leq [0.5T] \\ \lambda_{i0p} + \Delta_p & \text{for } t > [0.5T] \end{cases}.$$

The shift Δ_p (which is the same for all $i \in J$) is distributed $\mathcal{N}(0, [\lambda^*(0.4)]^2)$, i.i.d. across $p = 1, \ldots, r$, so that the shift is of the same magnitude as the initial loading λ_{i0p}.[13] The fraction of series that undergo a shift in the large break model is $[bN^{1/2}]/N \approx bN^{-1/2}$.

The principal components estimator $\hat{F}^k$ described earlier is used to estimate the factors.

[13]This shift process satisfies Assumption 2.4 with envelope functions of the same order as was used for the deterministic break in Example 3.

Estimation of the factor rank r is done using the "IC_{p2}" information criterion of Bai & Ng (2002) with a maximum rank of $r_{\max} = 10$, and, for simplicity, a minimum estimated rank of 1. The criterion is of the IC type in definition (2.10) with $g(N,T) = (\log C_{NT}^2)(N+T)/(NT)$. We also consider principal components estimates that impose the true rank $k = r$. To evaluate the principal components estimator's performance, we compute a trace R^2 statistic for the multivariate regression of $\hat{F}$ onto F,

$$R^2_{\hat{F},F} = \frac{\hat{E}\|P_F\hat{F}\|^2}{\hat{E}\|\hat{F}\|^2},$$

where $\hat{E}$ denotes averaging over Monte Carlo repetitions and $P_F = F(F'F)^{-1}F'$. Corollary 2.1 states that this measure tends to 1 as $T \to \infty$. In each repetition we compute the feasible out-of-sample forecast $\hat{y}_{T+1|T} = \hat{\delta}'\hat{F}_T$, where $\hat{\delta}$ are the OLS coefficients in the regression of y_{t+1} onto $\hat{F}_t$ for $t \leq T-1$, as well as the infeasible forecast $\tilde{y}_{T+1|T} = \tilde{\delta}'F_T$, where $\tilde{\delta}$ is obtained by regressing y_{t+1} on the true factors F_t, $t \leq T-1$. The closeness of the feasible and infeasible forecasts is measured by the statistic

$$S^2_{\hat{y},\tilde{y}} = 1 - \frac{\hat{E}(\hat{y}_{T+1|T} - \tilde{y}_{T+1|T})^2}{\hat{E}(\hat{y}^2_{T+1|T})}.$$

The measures $R^2_{\hat{F},F}$ and $S^2_{\hat{y},\tilde{y}}$ were also used by Stock & Watson (2002).

2.5.2 Calibration

The free parameters are T, N, r, ρ, a, β, b, c and d. We set $r = 5$ throughout. In line with Stock & Watson (2002) and Amengual & Watson (2007), we consider $\rho = 0, 0.9$, $a = 0, 0.5$ and $\beta = 0, 0.5$.

To guide our choice of the crucial parameters b, c and d, we turn to the empirical analysis of Stock & Watson (2009). They fit a DFM to 144 quarterly U.S. macroeconomic time series

from 1959 to 2006, splitting the sample at the first quarter of 1984. Using their results, we compare the pre- and post-break estimated factor loadings. The ratio of the mean square changes in the factor loadings to the mean square pre-break factor loadings is 0.21. Assuming that the break date and factor loadings are known, the corresponding ratio in our large break DGP is

$$\frac{(Nr)^{-1}\sum_{i=1}^{N}\sum_{p=1}^{r}\Delta_p^2}{(Nr)^{-1}\sum_{i=1}^{N}\sum_{p=1}^{r}\lambda_{i0p}^2} = \frac{bN^{-1/2}[\lambda^*(0.4)]^2}{\int_0^{0.8}[\lambda^*(x)]^2dx/0.8} + o_p(1) = 0.66bN^{-1/2} + o_p(1),$$

regardless of the values of r, ρ, a and β. For $N = 144$ series, the value of the parameter b that brings the theoretical ratio in line with the observed one in the Stock & Watson (2009) dataset is $b = \sqrt{144} \cdot 0.21/0.66 = 3.7$. While we have ignored estimation error, it therefore seems empirically relevant to consider large break DGPs with a b of this magnitude. We pick $b = 3.5$ to be our benchmark value, which for $N = 100$ implies that $bN^{-1/2} = 35\%$ of the loadings undergo a break (for $N = 200$ and $N = 400$ the fraction is 25% and 18%, respectively). To stress test our conclusions, we also examine the extreme choice $b = 7$.

When calibrating the values of c and d, we take the following steps. Focusing on the parametrization $N = T = 200$ and $a = \beta = \rho = 0$, we first record the trace R^2 statistics for the large break DGPs with $b = 3.5$ and $b = 7$, respectively. We then determine round values of c and d such that the corresponding trace R^2 statistics for the random walk and white noise DGPs approximately match the above-mentioned two figures for the large break model. This yields $c = 2, 3.5$ and $d = 0.4, 0.7$. To compare the time variation with the scale of the initial factor loadings, note that with $a = \beta = \rho = 0$ and $r = 5$, the unconditional standard deviation of each initial factor loading is $\sqrt{\int_0^{0.8}[\lambda^*(x)]^2dx/0.8} = 0.45$. Because d is the standard deviation of $\lambda_{itp} - \lambda_{i0p}$ in the white noise model, the choice $d = 0.4$ creates fluctuations of about the same magnitude as the initial factor loadings. Similarly, the standard deviation of $\lambda_{iTp} - \lambda_{i0p}$ in the random walk model equals $cT^{-1/4} = 0.53$ for

$T = 200$ and $c = 2$. In Appendix A.2.2 we show that this amount of random walk parameter variation is of the same magnitude as the estimates for U.S. data presented in Eickmeier et al. (2015), while our $c = 3.5$ parametrization exhibits substantially more instability.[14]

2.5.3 Results

We perform 5,000 Monte Carlo repetitions for each DGP. To graphically illustrate the convergence properties of the principal components estimator, we first focus on the baseline set-up with $a = \beta = \rho = 0$, $N = T$ and $k = r$ (the true number of factors is known). We run simulations for a fine grid of T values, $T = 50, 100, 150, \ldots, 400$. The results are plotted in Figures 2.1 to 2.3, corresponding to the white noise, random walk and large break models, respectively. Each figure has two panels. The top panel shows the $R^2_{\hat{F},F}$ statistic as a function of the sample size T, for the three different choices of b, c or d. Similarly, the bottom panel shows the $S^2_{\hat{y},\tilde{y}}$ statistic. All figures confirm that, while time variation in the factor loadings, vanishing at the appropriate rate, does impact the precision of the principal components estimator, the performance improves as T increases, both in absolute terms and relative to the no-instability benchmark.

[14]For $T \geq 67$ our worst-case random walk DGP, $c = 3.5$, exhibits more time variation in factor loadings than any of the parametrizations considered by Stock & Watson (2002) and Banerjee et al. (2008).

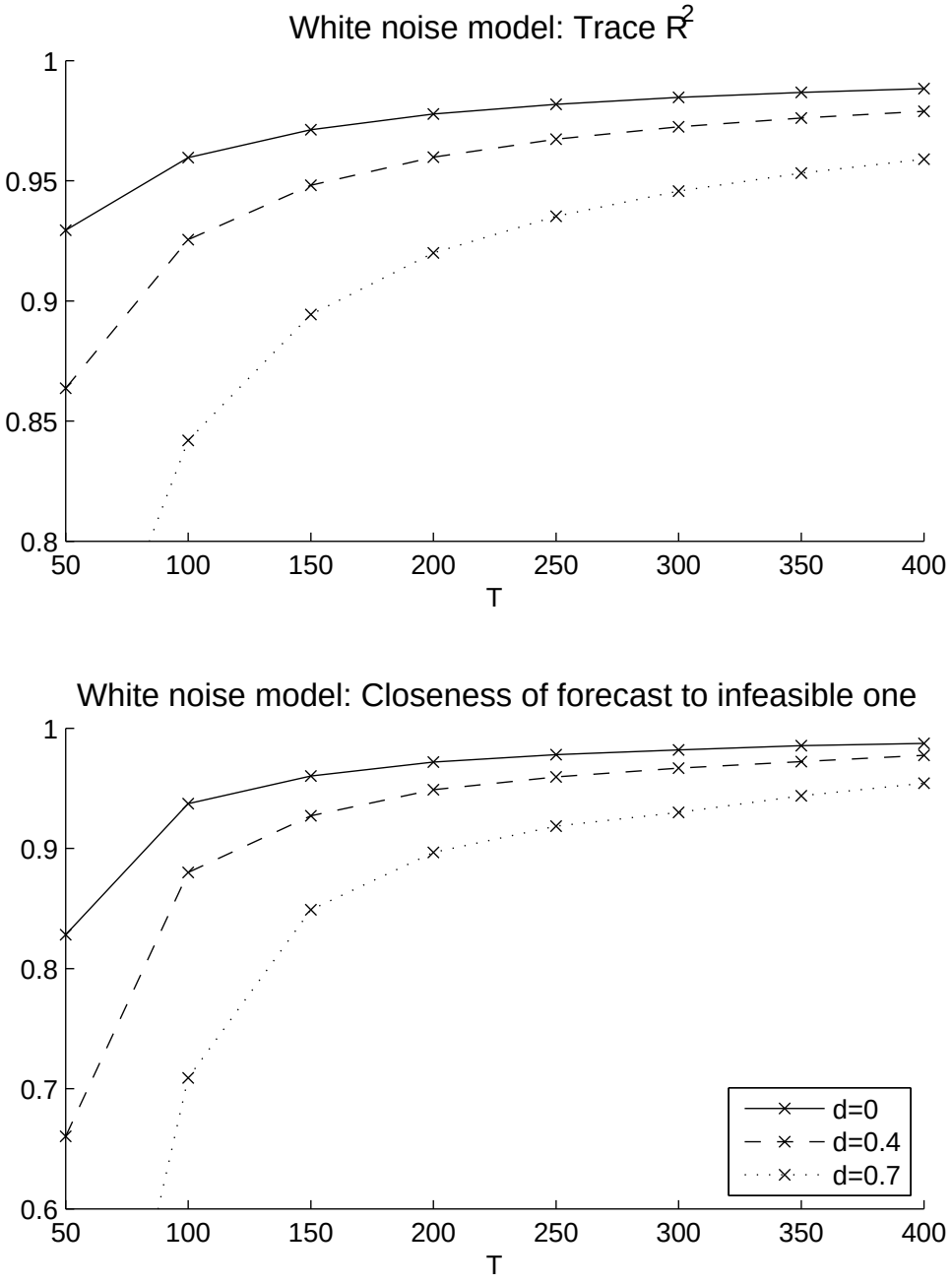

Figure 2.1: Simulation results for the white noise model, benchmark parameter and rate choices. Actual observations are marked with "x." Each is based on 5,000 Monte Carlo repetitions. The lines are piecewise linear interpolations.

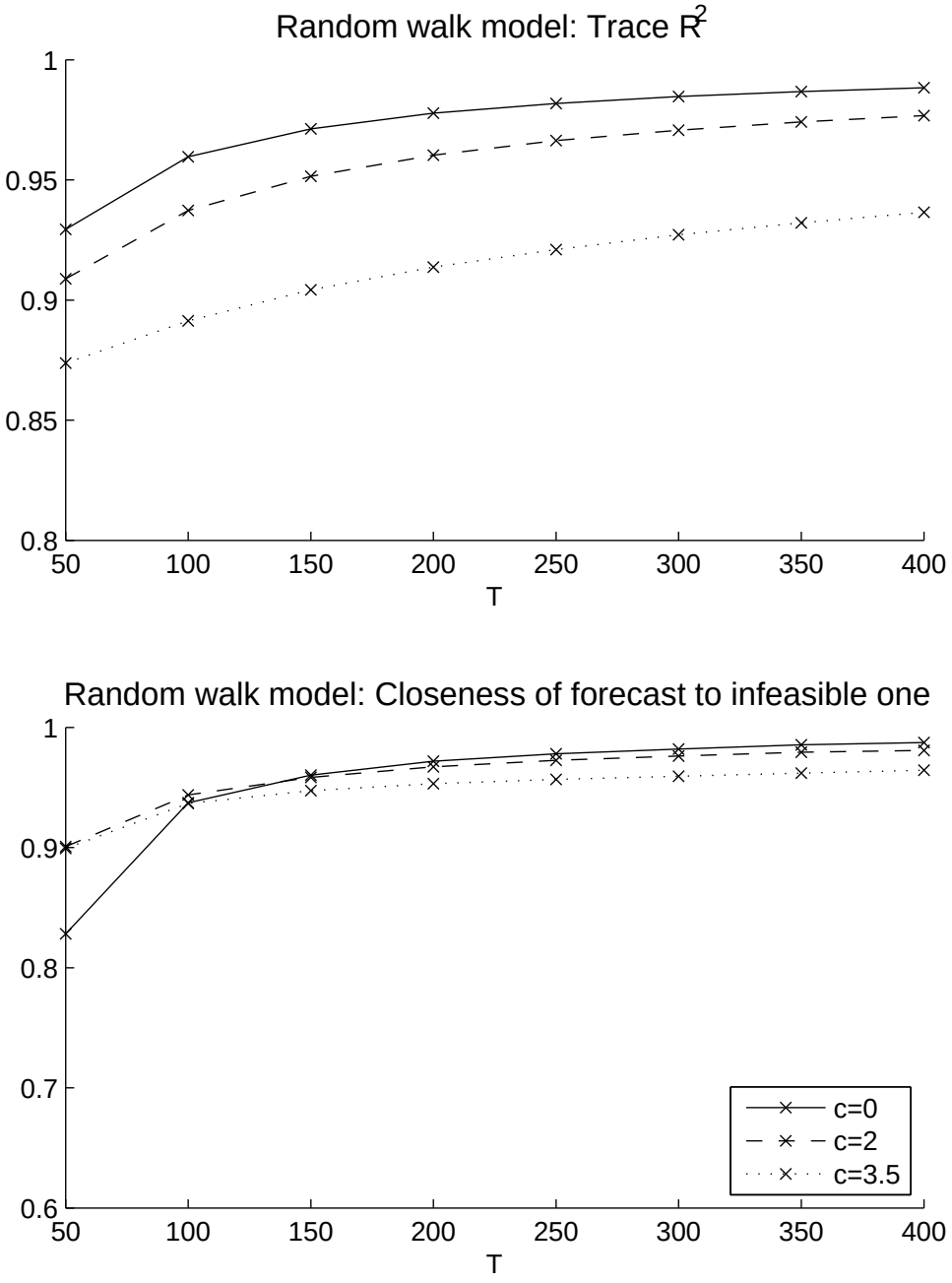

Figure 2.2: Simulation results for the random walk model, benchmark parameter and rate choices.

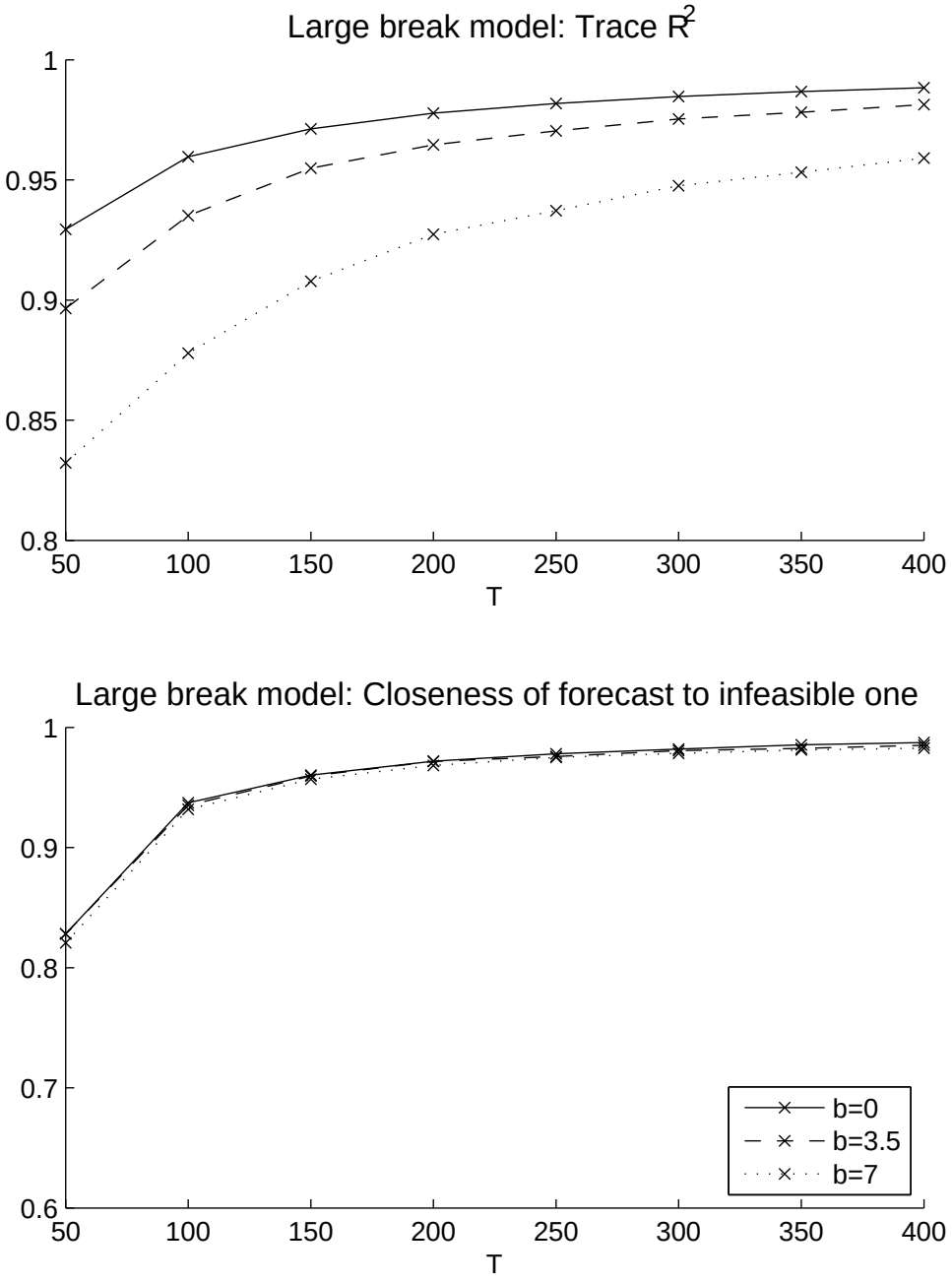

Figure 2.3: Simulation results for the large break model, benchmark parameter and rate choices.

Monte Carlo simulations: White noise model

		$d=0$					$d=0.4$					$d=0.7$				
		$k=r$		IC			$k=r$		IC			$k=r$		IC		
T	N	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$
$a=0,\ \beta=0,\ \rho=0$																
50	50	0.93	0.83	0.94	0.50	3.4	0.86	0.64	0.91	−0.83	1.6	0.71	0.25	0.84	−2.77	1.0
50	100	0.96	0.93	0.96	0.88	4.7	0.92	0.84	0.94	0.35	2.9	0.81	0.59	0.89	−1.72	1.1
100	100	0.96	0.94	0.96	0.93	5.0	0.93	0.88	0.93	0.66	3.9	0.84	0.70	0.90	−1.88	1.3
100	200	0.98	0.97	0.98	0.97	5.0	0.96	0.94	0.96	0.93	5.0	0.91	0.86	0.93	0.13	2.7
200	100	0.96	0.94	0.96	0.94	5.0	0.93	0.89	0.93	0.86	4.8	0.86	0.77	0.88	−0.20	2.4
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.95	0.96	0.95	5.0	0.92	0.90	0.93	0.70	4.1
200	400	0.99	0.99	0.99	0.99	5.0	0.98	0.97	0.98	0.97	5.0	0.96	0.95	0.96	0.94	5.0
$a=0.5,\ \beta=0,\ \rho=0$																
50	50	0.91	0.77	0.93	0.53	3.7	0.86	0.64	0.91	−0.42	2.0	0.75	0.38	0.86	−2.39	1.1
50	100	0.95	0.90	0.95	0.88	4.8	0.92	0.84	0.93	0.57	3.6	0.84	0.68	0.91	−1.01	1.4
100	100	0.96	0.93	0.96	0.92	5.0	0.93	0.89	0.93	0.78	4.4	0.87	0.77	0.91	−0.70	1.9
100	200	0.98	0.97	0.98	0.97	5.0	0.96	0.95	0.96	0.94	5.0	0.93	0.89	0.93	0.65	3.8
200	100	0.96	0.94	0.96	0.94	5.0	0.93	0.90	0.93	0.90	4.9	0.88	0.82	0.89	0.41	3.4
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.95	0.96	0.95	5.0	0.93	0.92	0.94	0.87	4.8
200	400	0.99	0.99	0.99	0.99	5.0	0.98	0.98	0.98	0.98	5.0	0.96	0.96	0.96	0.96	5.0
$a=0,\ \beta=0.5,\ \rho=0$																
50	50	0.91	0.76	0.93	0.53	3.7	0.85	0.58	0.90	−0.87	1.7	0.70	0.22	0.83	−3.09	1.0
50	100	0.95	0.91	0.96	0.87	4.7	0.92	0.83	0.93	0.39	3.0	0.80	0.57	0.89	−1.58	1.1
100	100	0.96	0.92	0.96	0.92	5.0	0.92	0.87	0.93	0.66	4.0	0.84	0.70	0.89	−1.90	1.3
100	200	0.98	0.97	0.98	0.97	5.0	0.96	0.94	0.96	0.93	5.0	0.91	0.86	0.93	0.15	2.7
200	100	0.96	0.94	0.95	0.94	5.0	0.92	0.88	0.92	0.86	4.8	0.85	0.76	0.88	−0.19	2.4
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.95	0.96	0.95	5.0	0.92	0.90	0.92	0.70	4.1
200	400	0.99	0.99	0.99	0.99	5.0	0.98	0.97	0.98	0.97	5.0	0.96	0.95	0.96	0.94	5.0
$a=0,\ \beta=0,\ \rho=0.9$																
50	50	0.95	0.81	0.97	0.43	2.3	0.61	0.03	0.84	−1.18	1.0	0.32	−0.96	0.52	−2.94	1.0
50	100	0.97	0.91	0.98	0.69	2.9	0.70	0.37	0.91	−0.75	1.0	0.38	−0.38	0.67	−1.74	1.0
100	100	0.97	0.94	0.97	0.81	3.9	0.75	0.43	0.89	−1.46	1.0	0.39	−0.43	0.67	−2.48	1.0
100	200	0.98	0.97	0.98	0.94	4.6	0.84	0.65	0.94	−0.75	1.3	0.48	−0.00	0.80	−1.60	1.0
200	100	0.97	0.94	0.97	0.93	4.9	0.79	0.55	0.88	−1.68	1.2	0.43	−0.50	0.68	−3.19	1.0
200	200	0.98	0.97	0.98	0.97	5.0	0.88	0.80	0.93	−0.52	1.7	0.57	0.11	0.81	−2.02	1.0
200	400	0.99	0.99	0.99	0.99	5.0	0.94	0.90	0.95	0.40	2.7	0.69	0.45	0.88	−1.99	1.0
$a=0.5,\ \beta=0.5,\ \rho=0.9$																
50	50	0.94	0.74	0.95	0.65	3.7	0.71	0.24	0.88	−1.05	1.0	0.41	−0.55	0.66	−1.91	1.0
50	100	0.97	0.86	0.97	0.83	4.5	0.80	0.51	0.93	−0.52	1.2	0.49	−0.09	0.79	−1.40	1.0
100	100	0.96	0.90	0.97	0.88	4.6	0.82	0.56	0.92	−1.05	1.3	0.50	−0.07	0.77	−2.04	1.0
100	200	0.98	0.96	0.98	0.95	4.9	0.90	0.76	0.95	−0.07	1.9	0.61	0.25	0.87	−1.51	1.0
200	100	0.96	0.93	0.96	0.92	5.0	0.84	0.66	0.90	−0.66	1.8	0.55	−0.07	0.76	−2.91	1.0
200	200	0.98	0.97	0.98	0.96	5.0	0.91	0.85	0.94	0.19	2.5	0.70	0.39	0.86	−1.93	1.0
200	400	0.99	0.98	0.99	0.98	5.0	0.95	0.93	0.96	0.71	3.6	0.80	0.64	0.92	−1.79	1.0

Table 2.1: Simulation results for DGPs with white noise disturbances in the factor loadings.

Monte Carlo simulations: Random walk model

		$c=0$					$c=2$					$c=3.5$				
		$k=r$		IC			$k=r$		IC			$k=r$		IC		
T	N	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$
$a=0,\ \beta=0,\ \rho=0$																
50	50	0.93	0.83	0.94	0.50	3.4	0.91	0.90	0.91	0.89	4.8	0.87	0.89	0.87	0.90	5.2
50	100	0.96	0.93	0.96	0.88	4.7	0.93	0.93	0.93	0.93	5.0	0.89	0.92	0.88	0.91	5.7
100	100	0.96	0.94	0.96	0.93	5.0	0.94	0.94	0.94	0.95	5.0	0.89	0.94	0.89	0.94	5.4
100	200	0.98	0.97	0.98	0.97	5.0	0.95	0.96	0.95	0.96	5.0	0.90	0.94	0.89	0.94	6.8
200	100	0.96	0.94	0.96	0.94	5.0	0.94	0.95	0.94	0.95	5.0	0.90	0.94	0.90	0.94	5.6
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.97	0.96	0.97	5.0	0.91	0.95	0.91	0.95	7.0
200	400	0.99	0.99	0.99	0.99	5.0	0.97	0.98	0.97	0.97	5.0	0.92	0.96	0.91	0.96	9.5
$a=0.5,\ \beta=0,\ \rho=0$																
50	50	0.91	0.77	0.93	0.53	3.7	0.91	0.90	0.91	0.87	4.7	0.88	0.89	0.88	0.90	5.1
50	100	0.95	0.90	0.95	0.88	4.8	0.94	0.94	0.94	0.94	5.0	0.90	0.92	0.89	0.92	5.5
100	100	0.96	0.93	0.96	0.92	5.0	0.94	0.95	0.94	0.95	5.0	0.90	0.94	0.90	0.94	5.2
100	200	0.98	0.97	0.98	0.97	5.0	0.96	0.96	0.96	0.96	5.0	0.91	0.95	0.91	0.94	6.1
200	100	0.96	0.94	0.96	0.94	5.0	0.95	0.95	0.95	0.95	5.0	0.92	0.95	0.91	0.95	5.2
200	200	0.98	0.97	0.98	0.97	5.0	0.97	0.97	0.97	0.97	5.0	0.93	0.96	0.92	0.96	5.8
200	400	0.99	0.99	0.99	0.99	5.0	0.98	0.98	0.98	0.98	5.0	0.93	0.96	0.93	0.96	8.4
$a=0,\ \beta=0.5,\ \rho=0$																
50	50	0.91	0.76	0.93	0.53	3.7	0.91	0.89	0.91	0.89	4.8	0.87	0.89	0.87	0.90	5.3
50	100	0.95	0.91	0.96	0.87	4.7	0.93	0.93	0.93	0.93	5.0	0.89	0.92	0.88	0.91	5.8
100	100	0.96	0.92	0.96	0.92	5.0	0.94	0.94	0.94	0.95	5.0	0.89	0.94	0.89	0.94	5.5
100	200	0.98	0.97	0.98	0.97	5.0	0.95	0.96	0.95	0.96	5.0	0.90	0.94	0.89	0.94	6.9
200	100	0.96	0.94	0.95	0.94	5.0	0.94	0.95	0.94	0.95	5.0	0.90	0.95	0.90	0.94	5.8
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.97	0.96	0.97	5.0	0.91	0.95	0.91	0.95	7.1
200	400	0.99	0.99	0.99	0.99	5.0	0.97	0.98	0.97	0.98	5.0	0.92	0.96	0.91	0.96	9.5
$a=0,\ \beta=0,\ \rho=0.9$																
50	50	0.95	0.81	0.97	0.43	2.3	0.94	0.90	0.94	0.91	5.5	0.94	0.90	0.94	0.92	7.3
50	100	0.97	0.91	0.98	0.69	2.9	0.95	0.92	0.95	0.94	6.4	0.94	0.92	0.94	0.94	8.6
100	100	0.97	0.94	0.97	0.81	3.9	0.93	0.93	0.92	0.95	6.7	0.91	0.93	0.91	0.95	9.2
100	200	0.98	0.97	0.98	0.94	4.6	0.93	0.94	0.93	0.97	7.8	0.92	0.94	0.91	0.96	9.9
200	100	0.97	0.94	0.97	0.93	4.9	0.91	0.94	0.90	0.95	7.4	0.88	0.94	0.88	0.96	9.9
200	200	0.98	0.97	0.98	0.97	5.0	0.92	0.95	0.91	0.97	8.3	0.89	0.95	0.88	0.97	10.0
200	400	0.99	0.99	0.99	0.99	5.0	0.92	0.95	0.91	0.98	9.6	0.89	0.95	0.88	0.97	10.0
$a=0.5,\ \beta=0.5,\ \rho=0.9$																
50	50	0.94	0.74	0.95	0.65	3.7	0.94	0.90	0.94	0.91	6.8	0.94	0.90	0.94	0.92	8.3
50	100	0.97	0.86	0.97	0.83	4.5	0.95	0.92	0.95	0.95	8.4	0.94	0.92	0.94	0.94	9.7
100	100	0.96	0.90	0.97	0.88	4.6	0.93	0.93	0.92	0.95	7.7	0.92	0.93	0.91	0.95	9.7
100	200	0.98	0.96	0.98	0.95	4.9	0.94	0.95	0.93	0.97	9.2	0.92	0.94	0.91	0.97	10.0
200	100	0.96	0.93	0.96	0.92	5.0	0.92	0.94	0.91	0.95	8.0	0.89	0.94	0.88	0.95	10.0
200	200	0.98	0.97	0.98	0.96	5.0	0.93	0.95	0.92	0.97	8.9	0.89	0.95	0.89	0.97	10.0
200	400	0.99	0.98	0.99	0.98	5.0	0.93	0.96	0.92	0.98	9.9	0.90	0.95	0.89	0.98	10.0

Table 2.2: Simulation results for DGPs with random walk factor loadings.

Monte Carlo simulations: Large break model

		$b=0$					$b=3.5$					$b=7$				
		$k=r$		IC			$k=r$		IC			$k=r$		IC		
T	N	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$R^2_{\hat{F},F}$	$S^2_{\hat{y},\tilde{y}}$	$\hat{E}(\hat{k})$
$a=0,\ \beta=0,\ \rho=0$																
50	50	0.93	0.83	0.94	0.50	3.4	0.90	0.83	0.90	0.59	3.6	0.83	0.82	0.84	0.58	3.7
50	100	0.96	0.93	0.96	0.88	4.7	0.94	0.92	0.94	0.89	4.7	0.89	0.91	0.89	0.89	4.8
100	100	0.96	0.94	0.96	0.93	5.0	0.94	0.94	0.94	0.93	5.0	0.88	0.93	0.88	0.93	5.1
100	200	0.98	0.97	0.98	0.97	5.0	0.97	0.97	0.96	0.97	5.1	0.93	0.96	0.93	0.97	5.3
200	100	0.96	0.94	0.96	0.94	5.0	0.93	0.94	0.93	0.94	5.2	0.87	0.94	0.87	0.94	5.5
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.97	0.96	0.97	5.2	0.93	0.97	0.92	0.97	5.6
200	400	0.99	0.99	0.99	0.99	5.0	0.98	0.98	0.98	0.99	5.2	0.96	0.98	0.95	0.99	5.6
$a=0.5,\ \beta=0,\ \rho=0$																
50	50	0.91	0.77	0.93	0.53	3.7	0.88	0.79	0.89	0.61	3.8	0.82	0.79	0.83	0.62	3.9
50	100	0.95	0.90	0.95	0.88	4.8	0.93	0.90	0.93	0.89	4.9	0.88	0.89	0.88	0.88	5.0
100	100	0.96	0.93	0.96	0.92	5.0	0.93	0.93	0.93	0.93	5.0	0.88	0.93	0.88	0.93	5.2
100	200	0.98	0.97	0.98	0.97	5.0	0.96	0.97	0.96	0.97	5.1	0.93	0.96	0.92	0.97	5.4
200	100	0.96	0.94	0.96	0.94	5.0	0.93	0.94	0.93	0.94	5.2	0.87	0.94	0.87	0.94	5.5
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.97	0.96	0.97	5.2	0.93	0.97	0.92	0.97	5.6
200	400	0.99	0.99	0.99	0.99	5.0	0.98	0.98	0.98	0.98	5.2	0.96	0.98	0.95	0.98	5.6
$a=0,\ \beta=0.5,\ \rho=0$																
50	50	0.91	0.76	0.93	0.53	3.7	0.87	0.77	0.88	0.61	3.9	0.80	0.77	0.81	0.61	4.1
50	100	0.95	0.91	0.96	0.87	4.7	0.93	0.90	0.93	0.89	4.8	0.88	0.90	0.88	0.88	5.0
100	100	0.96	0.92	0.96	0.92	5.0	0.93	0.93	0.93	0.93	5.1	0.87	0.92	0.87	0.93	5.3
100	200	0.98	0.97	0.98	0.97	5.0	0.96	0.96	0.96	0.96	5.1	0.93	0.96	0.92	0.97	5.4
200	100	0.96	0.94	0.95	0.94	5.0	0.93	0.94	0.92	0.94	5.3	0.86	0.93	0.86	0.94	5.7
200	200	0.98	0.97	0.98	0.97	5.0	0.96	0.97	0.96	0.97	5.3	0.92	0.97	0.91	0.97	5.7
200	400	0.99	0.99	0.99	0.99	5.0	0.98	0.98	0.98	0.98	5.3	0.96	0.98	0.95	0.99	5.7
$a=0,\ \beta=0,\ \rho=0.9$																
50	50	0.95	0.81	0.97	0.43	2.3	0.94	0.81	0.95	0.52	2.4	0.92	0.81	0.93	0.56	2.6
50	100	0.97	0.91	0.98	0.69	2.9	0.97	0.90	0.97	0.72	3.0	0.95	0.90	0.95	0.73	3.1
100	100	0.97	0.94	0.97	0.81	3.9	0.96	0.93	0.96	0.83	4.0	0.93	0.91	0.93	0.83	4.2
100	200	0.98	0.97	0.98	0.94	4.6	0.98	0.97	0.98	0.94	4.7	0.96	0.96	0.96	0.94	4.8
200	100	0.97	0.94	0.97	0.93	4.9	0.95	0.94	0.95	0.93	5.0	0.91	0.93	0.91	0.93	5.1
200	200	0.98	0.97	0.98	0.97	5.0	0.97	0.97	0.97	0.97	5.1	0.95	0.96	0.94	0.97	5.3
200	400	0.99	0.99	0.99	0.99	5.0	0.99	0.98	0.98	0.98	5.1	0.97	0.98	0.97	0.98	5.4
$a=0.5,\ \beta=0.5,\ \rho=0.9$																
50	50	0.94	0.74	0.95	0.65	3.7	0.93	0.75	0.94	0.70	4.1	0.90	0.75	0.91	0.71	4.3
50	100	0.97	0.86	0.97	0.83	4.5	0.96	0.85	0.96	0.84	4.8	0.94	0.86	0.94	0.85	5.0
100	100	0.96	0.90	0.97	0.88	4.6	0.95	0.89	0.95	0.88	4.8	0.92	0.88	0.92	0.88	5.0
100	200	0.98	0.96	0.98	0.95	4.9	0.97	0.95	0.97	0.95	5.2	0.95	0.94	0.95	0.95	5.4
200	100	0.96	0.93	0.96	0.92	5.0	0.94	0.92	0.94	0.93	5.3	0.90	0.91	0.90	0.93	5.6
200	200	0.98	0.97	0.98	0.96	5.0	0.97	0.96	0.97	0.97	5.3	0.94	0.95	0.94	0.97	5.6
200	400	0.99	0.98	0.99	0.98	5.0	0.98	0.98	0.98	0.98	5.3	0.97	0.98	0.96	0.98	5.7

Table 2.3: Simulation results for DGPs with a single large break in the factor loadings.

Tables 2.1 to 2.3 display a more comprehensive range of simulation results for the white noise, random walk and large break models, respectively. As explained above, we consider two values each for the instability parameters b, c and d, and each table compares those results to the no-instability benchmark ($b = c = d = 0$). The columns marked "$k = r$" impose knowledge of the true number of factors, while the columns marked "IC" correspond to simulations in which the factor rank is estimated using an information criterion. $\hat{E}(\hat{k})$ denotes the average estimated rank. We focus on dataset dimensions that are especially relevant for macroeconomic analyses with quarterly data, namely $T = 50, 100, 200$ and N either equal to, half or double the value of T.

Our first set of simulations has $a = \beta = \rho = 0$, i.e., no serial or cross-sectional dependence in the factors or idiosyncratic errors. For the empirically calibrated amount of instability (the middle five columns in each table), the $R^2_{\hat{F},F}$ and $S^2_{\hat{y},\tilde{y}}$ statistics are close to the no-instability benchmark as long as $N \geq T \geq 100$. The average estimated rank is also close to the truth $r = 5$ in these cases. Throughout Table 2.3 and Figure 2.2, the large break model does remarkably well in terms of the closeness $S^2_{\hat{y},\tilde{y}}$ of the feasible and infeasible forecasts, even when a majority of factor loadings undergo a break. As Figure 2.1 already demonstrated, the white noise model gives comparatively poor results for small T and when $N < T$, as predicted by our $\Lambda_0 = 0$ calculation, cf. expression (2.7). Increasing the amount of structural instability to extreme values (the right-most five columns of each table) substantially affects the results, more so than the introduction of moderate serial and cross-sectional correlation. The white noise model fares particularly poorly for $d = 0.7$, except when $N > T \geq 200$, and the estimated factor rank tends to severely undershoot the target for small sample sizes, as the common component in the data is diluted by the loading disturbances. For the random walk DGP, while the average estimated rank is hardly affected by moving from $c = 0$ to $c = 2$, extreme structural instability $c = 3.5$ does lead to significant deterioration of the performance of the information criterion; the continual evolution of the factor loadings over

time causes overestimation of the number of common factors. For the large break model the information criterion does much better, although it overshoots somewhat, as established by Breitung & Eickmeier (2011).

We consider separately the effects of introducing serial ($a = 0.5$) or cross-sectional ($\beta = 0.5$) dependence in the idiosyncratic errors. Moderate serial correlation in the errors is clearly a second-order issue.[15] Exponentially decreasing cross-sectional correlation of the above-mentioned magnitude has only a slightly larger impact. Furthermore, there appears to be no interesting interaction between dependence in the idiosyncratic errors and instability in the factor loadings.

Introducing persistence in the factors ($\rho = 0.9$) dramatically worsens the results for the white noise DGP. For the empirically calibrated amount of instability, $d = 0.4$, the $R^2_{\hat{F},F}$ and $S^2_{\hat{y},\tilde{y}}$ statistics are unacceptably poor, except perhaps for large sample sizes, and the estimated rank is much too low. For the random walk model, factor persistence has a more moderate, but still noticeable, effect. It causes overestimation of the number of factors, which only becomes worse as the sample size increases, and the convergence to 1 of the $R^2_{\hat{F},F}$ and $S^2_{\hat{y},\tilde{y}}$ statistics is not evident for $T \leq 200$.[16] However, the absolute impact of the factor loading instability is not alarming, even for $c = 3.5$, unless consistent estimation of r is viewed as a goal in and of itself. In contrast to the first two models, the large break model does not exhibit noticeable sensitivity to the persistence of the factors. Since serial correlation in the factors tends to bias downward the estimate of the factor rank, it actually partially corrects for the upward bias induced by the one-time loading break.[17]

---

[15] In fact, relative to the i.i.d. benchmark, the $a = 0.5$ results are somewhat better in cases in which the estimated rank is much too low.

[16] In unreported simulations, we have confirmed that these statistics do begin to improve for larger values of T.

[17] In Table 2.3, the large break model often performs better for $\rho = 0.9$ than for $\rho = 0$. The reason is that the denominators in the $R^2_{\hat{F},F}$ and $S^2_{\tilde{y},\tilde{y}}$ statistics tend to increase with the persistence of the factors. For

The last seven rows in the tables display results for the most empirically relevant case in which the factors are persistent and the idiosyncratic errors are both serially and cross-sectionally correlated ($a = \beta = 0.5, \rho = 0.9$). As expected based on the discussion above, these figures are similar to those for $a = \beta = 0, \rho = 0.9$, and we find no interesting compounding effects of the various departures from the baseline parametrization.

We summarize the findings of the simulation study as follows.

- Empirically calibrated structural instability of the random walk or large break variety does not, on average, markedly impact the estimation of the factor space or diffusion index forecasts. Increasing the temporal instability by an order of magnitude does not overturn this conclusion.

- The impact of white noise disturbances is a lot more sensitive to the sample size, to the ratio of N to T (higher is better), and to the persistence of the factors (lower is better). The numbers in Table 2.1 arguably overstate this sensitivity, since d was calibrated based on a setting with $\rho = 0$ and $N = T = 200$, which is relatively favorable for the white noise model. In a sense, Table 2.1 documents how well the principal components estimator deals with substantial white noise disturbances when the sample size and relative dimension N/T are both large.

- The correlation structure of the idiosyncratic errors is not an important concern in the exponential design we consider here. We have also tried the linearly decreasing correlation structure of Bai & Ng (2002, section 6). As expected, such a set-up yields worse convergence rates than those exhibited in Tables 2.1 to 2.3, although the results are sensitive to the choice of correlation parameters.

- Estimation of the factor rank r is governed by somewhat different forces than estimation

---

the two other models, the detrimental impact on the numerators outweigh this effect.

of the factor space or diffusion index forecasting, as we anticipated in Section 2.4.1. Relative to the no-instability benchmark, the Bai & Ng (2002) information criterion estimator is generally biased downward in the white noise model, whereas it is biased upward in the large break model and (especially) the random walk model. In the latter two models, there is no indication that this bias vanishes as $N, T \to \infty$ for the choices of h_{NT} and $\|\Delta\|^2$ that we have considered here. However, overestimation of r is not a problem, on average, for diffusion index forecasting.

2.5.4 Rate of convergence

We now turn to the more detailed asymptotic rates stated in Theorem 2.1. Our method of proof and the calculations in Section 2.3.2 suggest that it may not in general be possible to improve upon the R_{NT} rate for our three examples of break processes. To investigate this claim, we carry out two exercises. First, we set $N = T$ and execute a separate set of simulations in which $\lambda_{itp} - \lambda_{i0p} = dT^{1/4}\xi_{itp}$ for the white noise model, $\lambda_{itp} - \lambda_{i,t-1,p} = cT^{-1/2}\zeta_{itp}$ for the random walk model, and the number of shifting series in the large break model is set to $[bN]$. These three rates all (just) violate the conditions for mean square consistency in Theorem 2.1. To make the results comparable to Figures 2.1 to 2.3, we scale down our choices of b, c and d so that the amount of time variation in the two experiments coincide for $T = 200$. All other parameters are unchanged. See Figures 2.4 to 2.6 for the results. As hypothesized, for the random walk and large break models the trace R^2 curve flattens out for large T, instead of converging with the no-instability curve as in Figures 2.2 and 2.3. For the white noise model, convergence seems to still obtain with $h_{NT} = dT^{1/4}$.[18] It would be interesting to explore whether temporal or cross-sectional dependence in the

[18]This is consistent with the calculations in Section 2.3.2, which showed that $h_{NT} = o(T^{1/2})$ is necessary and sufficient for mean square consistency when $\Lambda_0 = 0$, $F_t \equiv 1$, $k = r = 1$ and $T/N \to \theta \in (0, \infty)$, cf. equation (2.7).

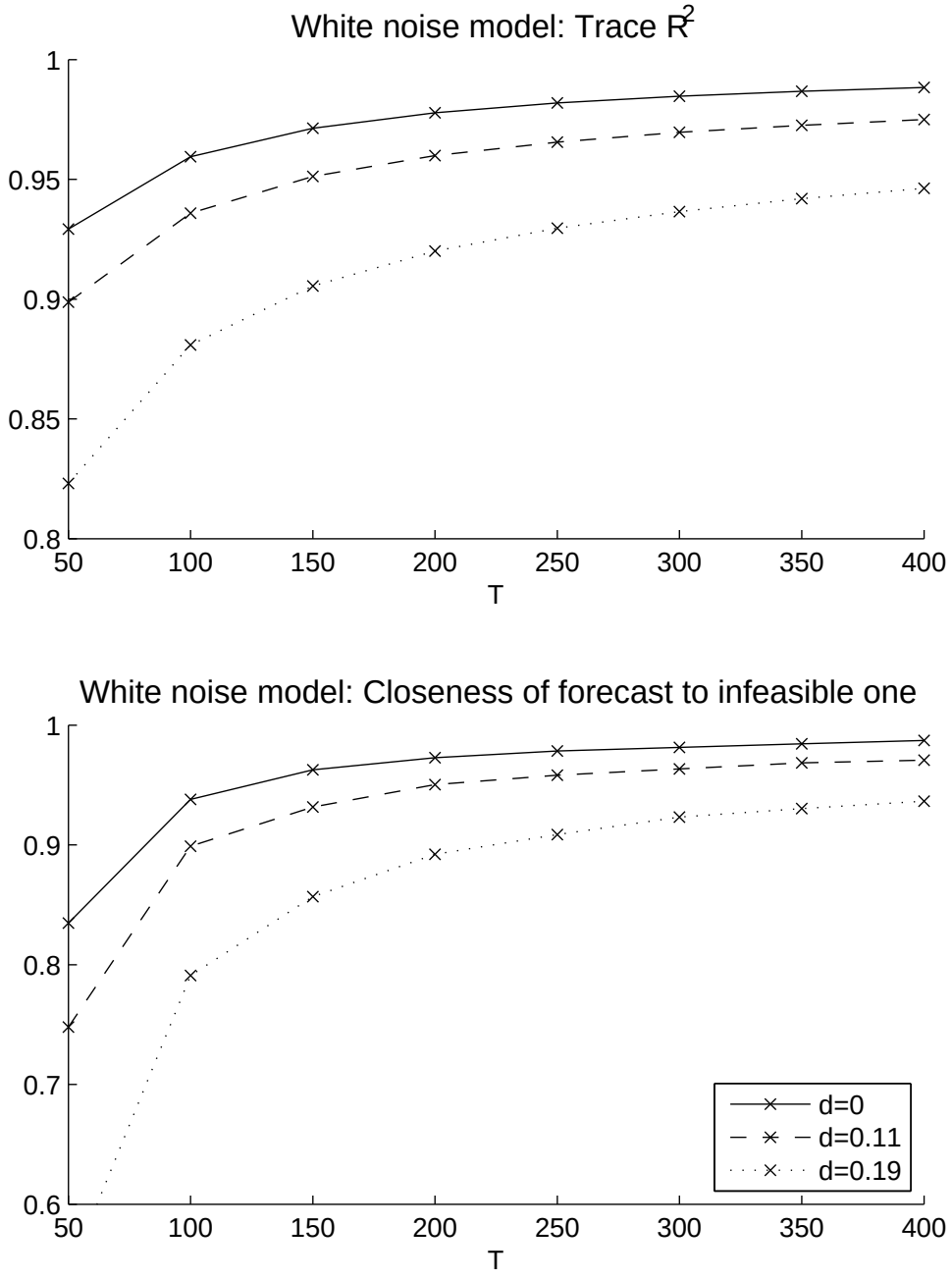

Figure 2.4: Simulation results for the white noise model, alternative rates.

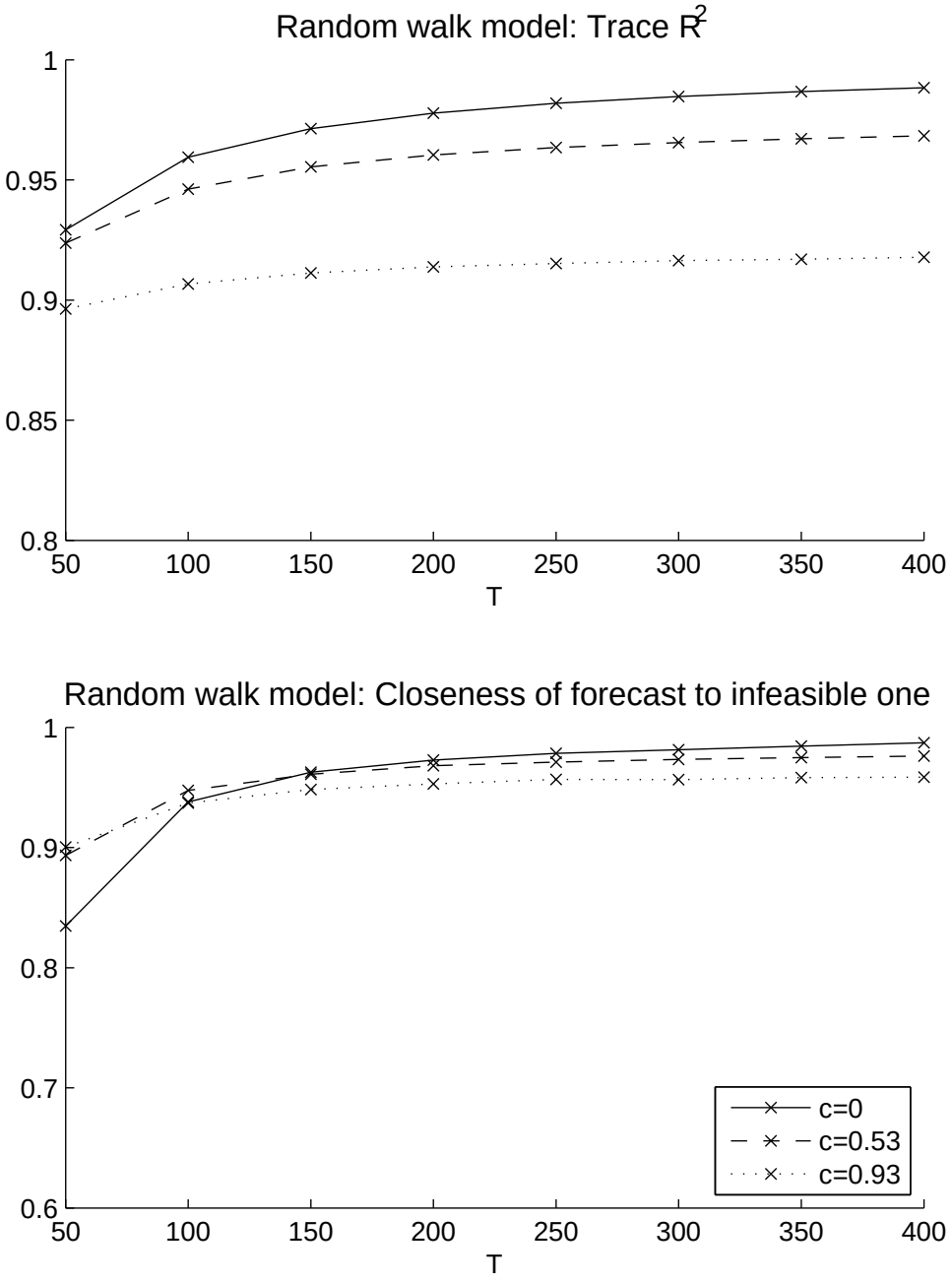

Figure 2.5: Simulation results for the random walk model, alternative rates.

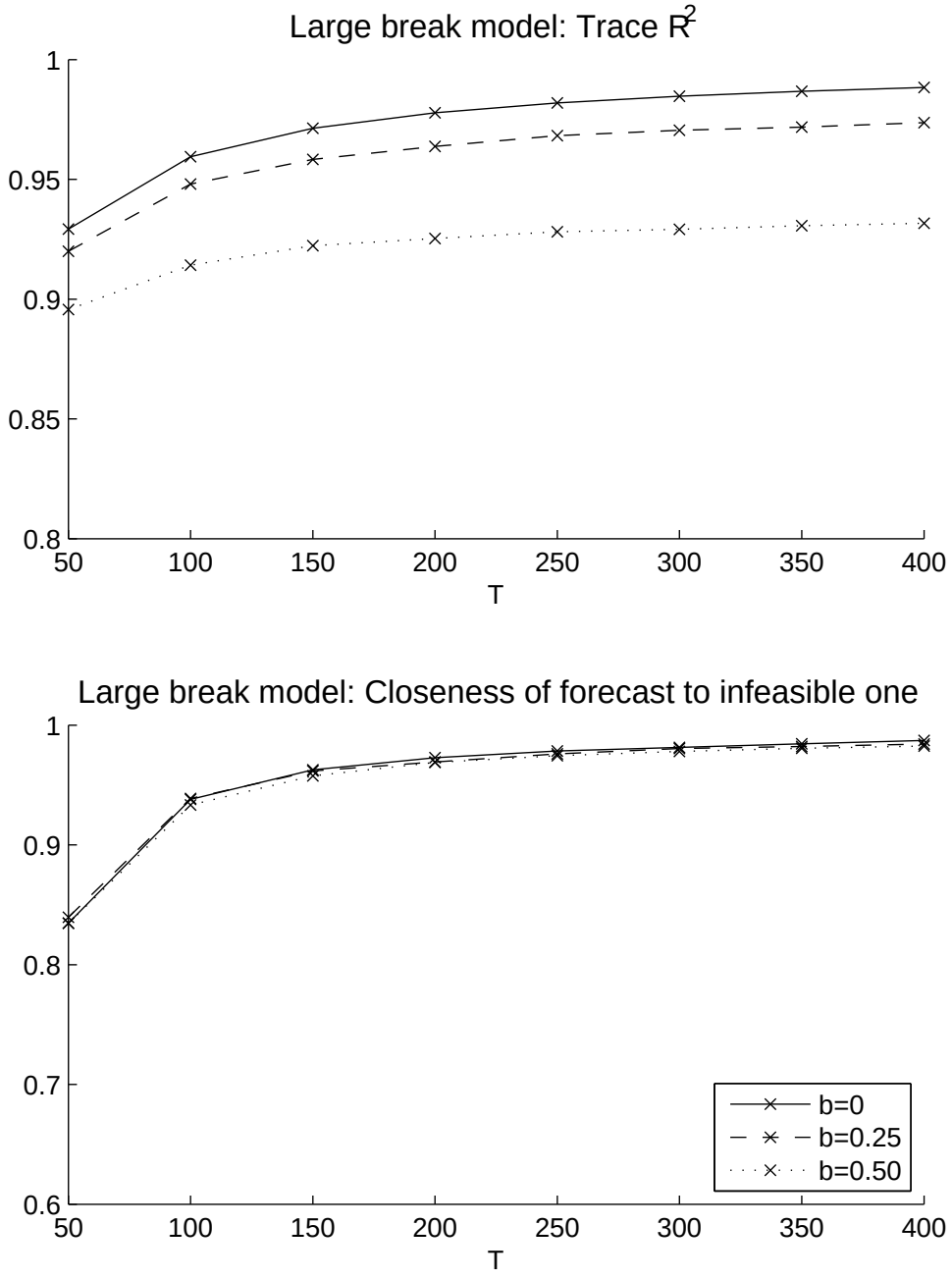

Figure 2.6: Simulation results for the large break model, alternative rates.

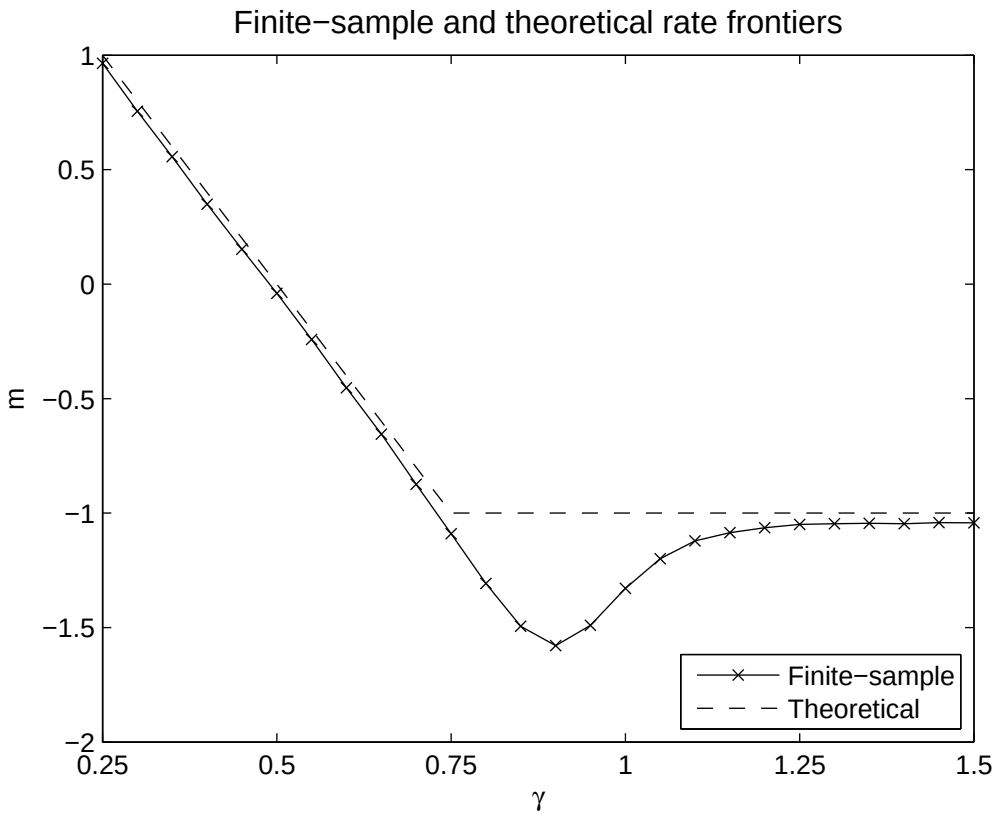

Figure 2.7: Rate frontiers for the random walk model with $c = 5$, $N = T$ and $h_{NT} = cT^{-\gamma}$. The solid line interpolates between the finite-sample rate exponent estimates $\hat{m}_\gamma$ (observations are marked with "x"), while the dashed line represents the theoretical rate exponent $m(\gamma)$.

disturbances ξ_{it} would make Theorem 2.1 tight also for the white noise model.

Second, we construct a "rate frontier" that corresponds to the predictions of Theorem 2.1 for the special case of the random walk model, which is the break process that has received most attention in the literature. Consider the explicit rate expression (2.5)–(2.6) for the random walk model under the assumptions $N = [T^\mu]$ and $h_{NT} = cT^{-\gamma}$. In the following we set $\mu = 1$ so that the rate exponent (2.6) reduces to

$$m(\gamma) = m(1, \gamma) = \max\{-1, 2 - 4\gamma\}.$$

The flat profile of the trace R^2 statistic in Figure 2.5 is fully consistent with $m(1/2) = 0$. These calculations pertain to the worst-case rate stated in Theorem 2.1. While Section 2.3.2

showed suggestive calculations for the special case $\Lambda_0 = 0$, we have not been able to prove that the convergence rate R_{NT} is sharp, in the sense that a generic DFM with random walk factor loadings that satisfies Assumptions 2.1 to 2.5 achieves the R_{NT} rate. Instead, we provide simulation evidence indicating that the independent random walk model achieves the stated bound. We maintain the simulation design described in Section 2.5.1 with $a = \beta = \rho = 0$ and $N = T$, except that we set $h_{NT} = 5T^{-\gamma}$ and vary γ over the range $0.25, 0.30, 0.35, \ldots, 1.50$. For each value of γ and each sample size $T = 200, 300, \ldots, 700$ we compute the statistic

$$\widehat{MSE}(\gamma, T) = T^{-1}(\hat{E}\|\hat{F}\|^2 - \hat{E}\|P_F\hat{F}\|^2),$$

where $\hat{E}$ denotes the average over 500 Monte Carlo repetitions. This statistic is a close analog of the mean square error that is the object of study in Theorem 2.1. Our theoretical results suggest that $\widehat{MSE}(\gamma, T)$ should grow or decay at rate $T^{m(\gamma)}$. We verify this by regressing, for each γ,

$$\log \widehat{MSE}(\gamma, T) = \text{constant}_\gamma + m_\gamma \log T,$$

using our six observations $T = 200, \ldots, 700$. Figure 2.7 plots the estimates $\hat{m}_\gamma$ against γ along with the theoretical values $m(\gamma)$. The estimated rate frontier is strikingly close to the theoretical one, although some finite-sample issues remain for intermediate values of γ. This corroborates our conjecture that Theorem 2.1 provides sharp rates for the independent random walk case.

2.6 Discussion and conclusions

The theoretical results of Section 2.3 and the simulation study of Section 2.5 point towards a considerable amount of robustness of the principal components estimator of the factors when the factor loading matrix varies over time. Although we have not proved that the

consistency rate function presented in Section 2.3.1 is tight in a formal sense, inspection of our proof, as well as calculations for special cases and Monte Carlo evidence, do not suggest any room for improvement, particularly for the random walk and large break models. In this sense our rate function represents an upper bound on the parameter instability that can be tolerated by the principal components estimator. The amount of such instability is quite large when calibrated to values of N and T typically used in applied work, which is reassuring for the nascent empirical research agenda that allows for structural instability in estimation of DFMs (Korobilis, 2013; Eickmeier et al., 2015).

Our evidence concerning the robustness of the principal components estimator raises a tension with the results in Breitung & Eickmeier (2011), who stress the harmful effect of undetected factor loading breaks on rank estimation. Our simulations show that diffusion index forecasting using principal components estimates can be effective even when the rank of the factor space is not estimated consistently. Indeed, we conjecture (but do not prove) that the principal components estimator and feasible diffusion index regression will be consistent under sequences of breaks for which the Breitung & Eickmeier (2011) test rejects. Furthermore, our simulations indicate that the direction of the rank estimation bias depends on the type of structural instability. Sorting out the relative importance of these countervailing forces for the sampling distribution of forecasts would be of independent interest and would also return the large-dimensional discussion here to the bias-variance tradeoffs associated with ignoring breaks tackled in a low-dimensional setting by Pesaran & Timmermann (2005, 2007).

In some applications, such as with data on asset returns, accurate estimation of the number of factors is of direct concern. Our results suggest that the allowable amount of structural instability in these cases is smaller than for forecasting purposes. More work is needed to establish necessary conditions for consistent rank estimation, and if necessary, to develop rank estimators that are more robust to different types of instability.

Chapter 3

Estimation of Smooth Impulse Response Functions

3.1 Introduction

An impulse response function (IRF) measures the dynamic response of a variable to an initial shock to another variable. In macroeconomics, IRFs are crucial tools for understanding economic dynamics and the causes of business cycles, cf. the survey by Ramey (2016). In applied microeconomics, IRFs conveniently summarize dynamic treatment effects.[1] In empirical settings where both the response and shock variables are observed, it is possible to estimate IRFs in a model-free way by simple regression methods (Cochrane & Piazzesi, 2002; Jordà, 2005). While such methods have low bias, they may produce jagged and highly variable IRF estimates in small samples. In many applications, researchers have *a priori* reasons to believe that the true IRF is smooth, but no established econometric method can exploit such smoothness without making strong parametric assumptions.

In this paper I propose a smooth impulse response function (SmIRF) estimator that

[1] Applications include the dynamic responses of worker earnings to separations (Jacobson, LaLonde & Sullivan, 1993), of consumption to stimulus payments (Broda & Parker, 2014), and of life outcomes to childhood teacher quality (Chetty, Friedman & Rockoff, 2014).

smooths out an initial non-smooth IRF estimate. The smoothing procedure can be applied to any uniformly consistent and asymptotically normal initial IRF estimator, for example from time series or panel regressions of outcomes on an observed shock. The degree of smoothing can be chosen to minimize a data-dependent estimate of the mean squared error (MSE) of the SmIRF estimator, thus optimally trading off bias and variance. The SmIRF estimator is a member of a computationally convenient class of shrinkage estimators that can flexibly impose a variety of smoothness, short-run, and long-run restrictions. I show that the SmIRF estimator dominates the non-smooth initial estimator in terms of MSE under realistic conditions. Finally, I propose novel procedures for constructing asymptotically uniformly valid confidence sets based on the shrinkage estimator.

Figure 3.1 illustrates how the SmIRF estimator smooths out an initial jagged IRF estimate in order to increase its precision. The response variable in the figure is a measure of U.S. financial sector balance sheet distress, while the shocks are monetary policy surprises identified from high-frequency financial data; the specification includes additional lagged control variables as in Ramey (2016). The IRF estimated using the regression-based Jordà (2005) "local projection" method is very jagged, and most of its spikes are likely due to sampling noise or outliers. The SmIRF estimator modifies the regression-based estimate by penalizing jagged IRFs, effectively averaging the initial IRF estimate across nearby impulse response horizons, and thus reducing variance. The increase in bias caused by moderate amounts of smoothing is small if the true IRF is smooth.

The SmIRF estimator is a function of a scalar smoothing parameter that can be chosen to optimally trade off bias and variance in a data-dependent way. This is done by minimizing an unbiased estimator of the risk (here: MSE) of the SmIRF estimator, as in Stein (1981). The unbiased risk estimate (URE) is easy to compute, as it depends only on the initial non-smooth IRF estimate and its estimated asymptotic variance. Minimizing the URE makes SmIRF adaptive: It only substantially smooths the IRF when the data confirms the

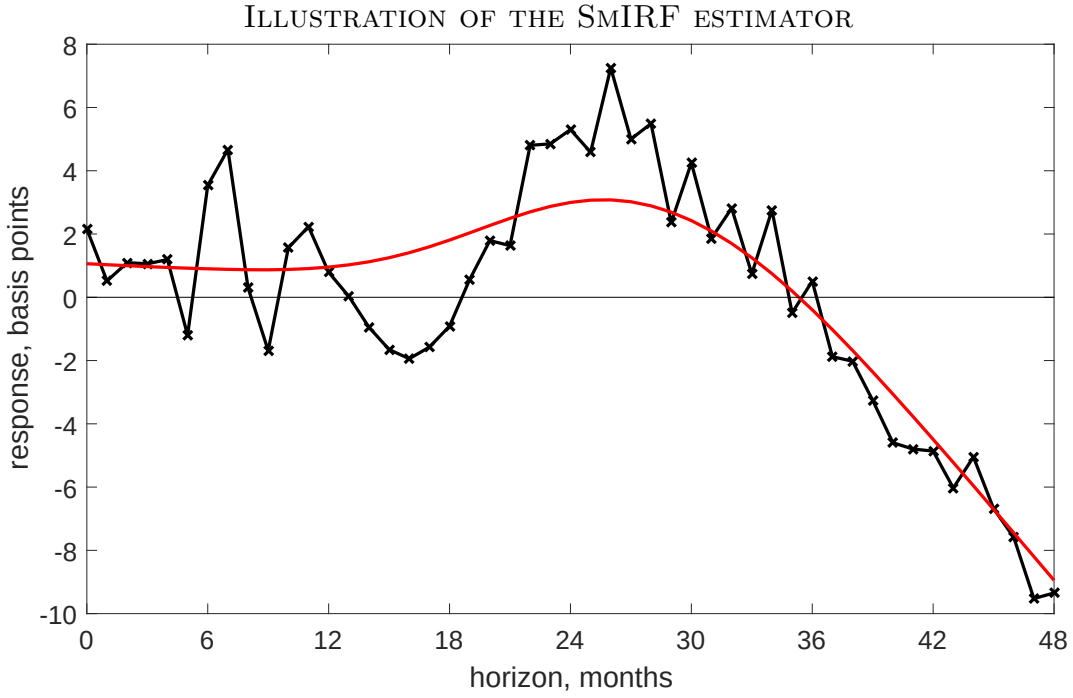

Figure 3.1: Local projection (jagged line, crosses) and SmIRF (smooth curve) estimators of the IRF of the excess bond premium to a 1-standard-deviation monetary policy shock, monthly U.S. data. Response: Gilchrist & Zakrajšek (2012) excess bond premium. Shock: Gertler & Karadi (2015) monetary policy shock. Controls: quadr. time trend, 2 lags of response, shock, log indu. prod., log cons. price index, 1-year Treas. rate. Sample: 1991:1–2012:6. Details in Appendix A.3.2.

smoothness hypothesis. Consequently, I prove that a version of the SmIRF estimator with data-dependent smoothing parameter dominates the initial non-smooth estimator in terms of MSE under realistic conditions.

The SmIRF estimator is a member of a flexible and computationally convenient class of shrinkage estimators. These shrinkage estimators penalize large squared values of user-specified linear combinations of the impulse response parameters. Estimators in this class are not only able to impose smoothness, but also short-run and long-run approximate restrictions on the impulse responses. An analytically tractable subclass of estimators are the projection shrinkage estimators, which shrink the initial IRF estimate towards a linear-in-parameters function of the response horizon, such as a polynomial.

The main theoretical contribution of this paper is to develop novel joint and pointwise

shrinkage confidence bands for conducting inference about the smoothed IRFs. The bands are obtained by numerically inverting test statistics that are functions of the shrinkage estimator, using simulated critical values.[2] These confidence sets have correct coverage in a finite-sample normal location model with arbitrary known covariance matrix. In the case of projection shrinkage, the finite-sample normal results translate into asymptotic uniform coverage when the distribution of the initial IRF estimator is unknown. The proposed confidence sets can be constructed so that they always contain the shrinkage estimator. Simulation evidence suggests that, if the true IRF is smooth, the shrinkage confidence sets often have smaller volume than the usual Wald confidence sets centered at the initial IRF estimate, and the shrinkage sets never perform substantially worse than the Wald sets.

Literature. The SmIRF estimator imposes smoothness in a more robust and transparent manner than parametric procedures such as Vector Autoregressions (VARs). This paper is not concerned with applications in which VAR-type methods are used to achieve identification due to the shock being unobserved (Stock & Watson, 2016). VARs and similar parsimonious models generate smooth IRFs by extrapolating long-run responses from short-run features of the data. While such models are efficient if the parametric assumptions hold, misspecification biases can be large, as discussed by Jordà (2005) and Ramey (2016, Sec. 2.4, 3.5). My paper demonstrates that it is not necessary to impose a rigid parametric structure to accurately estimate smooth IRFs. With a data-dependent smoothing parameter, the SmIRF estimator adapts to the smoothness of the true IRF; in contrast, robust parametric analysis requires specification tests that are not directly tied to IRFs and are often not uniformly consistent (Leeb & Pötscher, 2005).

The SmIRF estimator is related to and inspired by Shiller's (1973) smoothness prior

[2]I thank Isaiah Andrews for suggesting this strategy and Adam McCloskey for stimulating discussions.

estimator for distributed lag regressions, but I go further in providing methods for adaptive, MSE-optimal inference. The SmIRF estimator uses the same penalty for estimating jagged IRFs as the Shiller estimator. Unlike the latter, the SmIRF estimator nests the popular Jordà (2005) local projection estimator in the case of time series regression. In contrast to Shiller's focus on subjective Bayesian estimation, I provide methods for optimally selecting the degree of smoothing and for constructing confidence sets with guaranteed frequentist coverage. My procedure for selecting the smoothing parameter is more precisely theoretically founded and widely applicable than procedures designed for the Shiller estimator, cf. the survey by Hendry, Pagan & Sargan (1984, pp. 1060–1062). Moreover, I consider general shrinkage estimators that do not use the Shiller penalty.

Shrinkage estimation has recently received attention in economics outside of the domain of IRF estimation. Fessler & Kasy (2016) use an Empirical Bayes estimator to flexibly impose linear restrictions from economic theory in a manner similar to the projection shrinkage estimator in the present paper; however, they do not construct valid frequentist confidence sets. Hansen (2016a) introduces a shrinkage IV estimator which MSE-dominates two-stage least squares. Giannone, Lenza & Primiceri (2015) and Hansen (2016c) shrink VAR estimates to improve forecasting performance, assuming that the unrestricted VAR model is well-specified. Additionally, high-dimensional predictive/forecasting methods often rely on shrinkage to ensure good out-of-sample performance (Stock & Watson, 2012b; Belloni, Chernozhukov & Hansen, 2014).

The theoretical framework in this paper is formally similar to nonparametric regression, with the crucial difference that the regression errors may be heteroskedastic and cross-correlated. The impulse response horizons can be viewed as equally spaced design points, the initial IRF estimator as observed data, and the initial IRF estimation errors as regression errors. Viewed in this way, the shrinkage IRF estimators are similar to spline smoothing (Wahba, 1990); however, much of the theory for nonparametric regression has been devel-

oped under the assumption of independent and identically distributed errors, which does not apply in the IRF estimation context. Many papers get rid of serial correlation by transforming the data to a different coordinate system, but this transformation can drastically change the economic interpretation of the shrinkage estimator.[3] Lest the economic analysis be dictated by statistical convenience, my analysis directly deals with correlated errors.

My theoretical results build on developments in the shrinkage literature that followed Stein (1956) and James & Stein (1961), see Lehmann & Casella (1998, Ch. 5.4–5.7). The general class of shrinkage estimators I consider is akin to generalized ridge regression, cf. the survey by Vinod (1978). The subclass of projection shrinkage estimators has been analyzed in the abstract by Bock (1975), Oman (1982), and Casella & Hwang (1987), none of whom consider IRF estimation. My URE criterion is similar to those derived by Mallows (1973) and Berger (1985, Ch. 5.4.2), but my derivations rely on asymptotic rather than finite-sample normality, as in the model selection analysis of Claeskens & Hjort (2003) and Hansen (2010). The proof that the SmIRF estimator MSE-dominates the initial non-smooth estimator relies heavily on abstract results in Hansen (2016b). My proofs of asymptotic uniform validity of the projection shrinkage confidence sets employ the abstract drifting parameter techniques of Andrews, Cheng & Guggenberger (2011) and McCloskey (2015).

Despite their general applicability, the test inversion shrinkage confidence sets appear to be novel. Brown, Casella & Hwang (1995) and Tseng & Brown (1997) invert tests to obtain confidence sets with small prior expected volume, but the tests do not have direct connections to shrinkage estimation. My procedure allows for arbitrary dependence between the initial impulse response estimators at different horizons, whereas previous papers on shrinkage confidence sets tend to focus on the independent case, cf. the survey by Casella & Hwang (2012). In contrast to Beran (2010), my confidence sets are valid even when the

[3]For example, a penalty for estimating jagged IRFs may not resemble a jaggedness penalty once the estimator is transformed to another coordinate system.

number of parameters of interest (i.e., impulse responses) is small. However, unlike other papers, I do not provide analytic conditions for my confidence sets to beat the usual Wald sets in terms of expected volume, although I present encouraging simulation evidence.

Outline. The paper is organized as follows. Section 3.2 is a user's guide to SmIRF and other shrinkage estimators, the URE, and confidence set construction. Section 3.3 presents theoretical results on the MSE of shrinkage estimators. Section 3.4 derives valid shrinkage confidence sets. Section 3.5 contains a simulation study. Section 3.6 lists topics for future research. Appendix A.3 defines notation and data, gives technical details, and provides additional simulation results, while proofs are relegated to Appendix B.3.

3.2 Overview and examples

This section is a user's guide to estimating and doing inference on smooth impulse response functions (IRFs). First, I define IRFs, the object of interest. Second, I introduce the smooth impulse response function (SmIRF) estimator. Third, I show how to pick the smoothing parameter in a data-dependent way by minimizing an unbiased estimate of the mean squared error (MSE), thus optimally trading off bias and variance. Finally, I give algorithms for constructing joint and pointwise confidence bands. I illustrate the new methods with the empirical example from Figure 3.1 in the Introduction.

3.2.1 Impulse response functions

I start off by defining IRFs. I then review estimation of IRFs using approximately unbiased regression methods when both the outcome and shock variables are observed. Such unbiased estimators are often jagged and have unnecessarily high variance, as they do not impose smoothness on the IRF.

An IRF measures the average dynamic response of variable y_t to an initial shock or impulse to variable x_t (in the case of panel data, add an additional unit subscript j). The averaging can be across time, or across time and cross-section units, depending on whether the application involves time series or panel data. While nonlinearities can be economically interesting, in this paper I focus on average linear relationships. Hence, an IRF, broadly construed, is a vector $\beta = (\beta_0, \beta_1, \ldots, \beta_{n-1})'$ of impulse responses at horizons $i = 0, 1, \ldots, n-1$, where $n-1$ is the largest response horizon of interest, and $\beta_i = \partial y_{t+i}/\partial x_t$.

When both the outcome and shock variables are observed, asymptotically unbiased IRF estimates can be obtained by regressing current and future outcomes on the shock.[4] The regression may control for covariates or lagged outcomes. A prototypical specification is

$$y_{t+i} = \beta_i x_t + \text{controls} + \varepsilon_{t+i|t}, \tag{3.1}$$

where $\varepsilon_{t+i|t}$ is the forecast error at time t for forecasting i periods ahead. Running the above regression separately for each horizon $i = 0, 1, \ldots, n-1$, we obtain coefficient estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{n-1})'$ which constitute Jordà's (2005) "local projection" IRF estimator. This procedure has a panel regression counterpart that is often used in event studies – simply add unit subscripts j to the variables in (3.1). Even if the underlying true IRF is nonlinear, the coefficient estimates $\hat{\beta}$ capture the least-squares linear predictor of current and future outcomes y_t based on the shock x_t and controls.

Regression-based IRF estimators are often jagged and highly variable in moderate samples, especially if the regression includes many controls. This is illustrated in Figure 3.1 in the Introduction. If the IRF is estimated from horizon-by-horizon regressions, there is nothing

[4]This paper does not consider settings in which x_t is unobserved, unlike Structural Vector Autoregression (SVAR) analysis which seeks to jointly identify IRFs and shocks. In principle, the SmIRF estimator can be used to further smooth an SVAR-estimated IRF, but if identification is predicated on the assumptions of the SVAR model, it makes sense to impose smoothness directly in that model.

constraining the function to look smooth. This problem is shared by other IRF estimators in the recent literature, such as the propensity score weighted IRF estimator for discrete policy treatments in Angrist, Jordà & Kuersteiner (2013) and Ramey's (2016) instrumental variables (IV) extension of the local projection estimator.[5]

3.2.2 SmIRF and other shrinkage estimators

Below I introduce the SmIRF estimator and related estimators which smooth an initial IRF estimate by shrinking it towards a smoother shape. I argue that smoothing decreases the variance of the estimator, whereas the accompanying increase in bias is likely to be small in many applications. The SmIRF estimator belongs to a class of general shrinkage estimators that flexibly impose smoothness and other approximate restrictions. The analytically convenient subclass of projection shrinkage estimators shrinks the initial IRF estimate towards a polynomial, or any other linear-in-parameters function of the response horizon.

SmIRF estimator. In many applications, there is *a priori* reason to believe the true IRF to be smooth, and this knowledge can be exploited to improve the precision of the estimator. Smoothness of the IRF is a reasonable hypothesis in many economic applications, e.g., due to adjustment costs, consumption smoothing, information frictions, staggered decisions, or strategic complementarity. Loosely speaking, a low-bias, high-variance estimator can be smoothed by averaging the IRF across nearby response horizons, cf. Figure 3.1 in the Introduction. Such averaging decreases the variance of the estimator and reduces the influence of outlier data points. Averaging may increase the bias of the estimator, but this effect will be minimal for moderate amounts of smoothing if the true IRF is in fact smooth.

---

[5]Ramey (2016, Sec. 2.4) proposes regressing outcome y_{t+i} on policy variable x_t using a third variable z_t as IV, separately for each i. This amounts to a proportional scaling of the Jordà local projection IRF of y_t to z_t, because the "first stage" is the same at all horizons i.

The SmIRF estimator transforms an initial non-smooth IRF estimate into a smooth estimate by penalizing jagged functions. The initial estimator can be obtained from any type of data set using any method, as long as the initial estimator is uniformly consistent and asymptotically normal with consistently estimable asymptotic variance, in a sense made precise in Sections 3.3 and 3.4. These properties hold for many regression-based methods, such as the Jordà (2005) local projection estimator, under standard regularity conditions.

Given the initial non-smooth IRF estimator $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_{n-1})'$ and a scalar smoothing parameter $\lambda \geq 0$, the SmIRF estimator is defined as

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^n}{\arg\min} \sum_{i=0}^{n-1} (\beta_i - \hat{\beta}_i)^2 + \lambda \sum_{i=2}^{n-1} \{(\beta_i - \beta_{i-1}) - (\beta_{i-1} - \beta_{i-2})\}^2. \tag{3.2}$$

The SmIRF estimator trades off fidelity to the initial estimate with a penalty equal to the smoothing parameter λ times the sum of squared second differences of the IRF.[6] λ governs the degree to which the initial IRF estimate is smoothed out. If $\lambda = 0$, the SmIRF estimator equals the non-smooth initial estimate. As $\lambda \to \infty$, the SmIRF estimator converges to the straight line that best fits the initial IRF estimate. For $0 < \lambda < \infty$, the SmIRF estimator shrinks the initial estimate towards a straight line. Provided $\lambda > 0$, the SmIRF impulse response estimate $\hat{\beta}_i(\lambda)$ at horizon i is a function of the initial impulse response estimates at horizons other than i – the intended averaging effect of the smoothing procedure.

The jaggedness penalty in the definition of the SmIRF estimator is familiar to time series econometricians. Inspection of the definition (3.2) reveals that the SmIRF estimator $\hat{\beta}(\lambda)$ is just the Hodrick & Prescott (1997) trend of the artificial "time series" $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{n-1})$. Hence, the SmIRF estimator is easy to compute using standard software (see also formulas

[6]For local projection (3.1), $\hat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^n} ((T-h)\hat{\sigma}_{\tilde{x}}^2)^{-1} \sum_{t=1}^{T-n} \sum_{i=0}^{n-1} (y_{t+i} - \beta_i \tilde{x}_t)^2 + \lambda \sum_{i=2}^{n-1} \{(\beta_i - \beta_{i-1}) - (\beta_{i-1} - \beta_{i-2})\}^2$, where $\hat{\sigma}_{\tilde{x}}^2$ is the sample variance of $\tilde{x}_t$, the residuals after regressing x_t on controls. Hence, SmIRF trades off the sum of squared forecast errors across i with the jaggedness penalty.

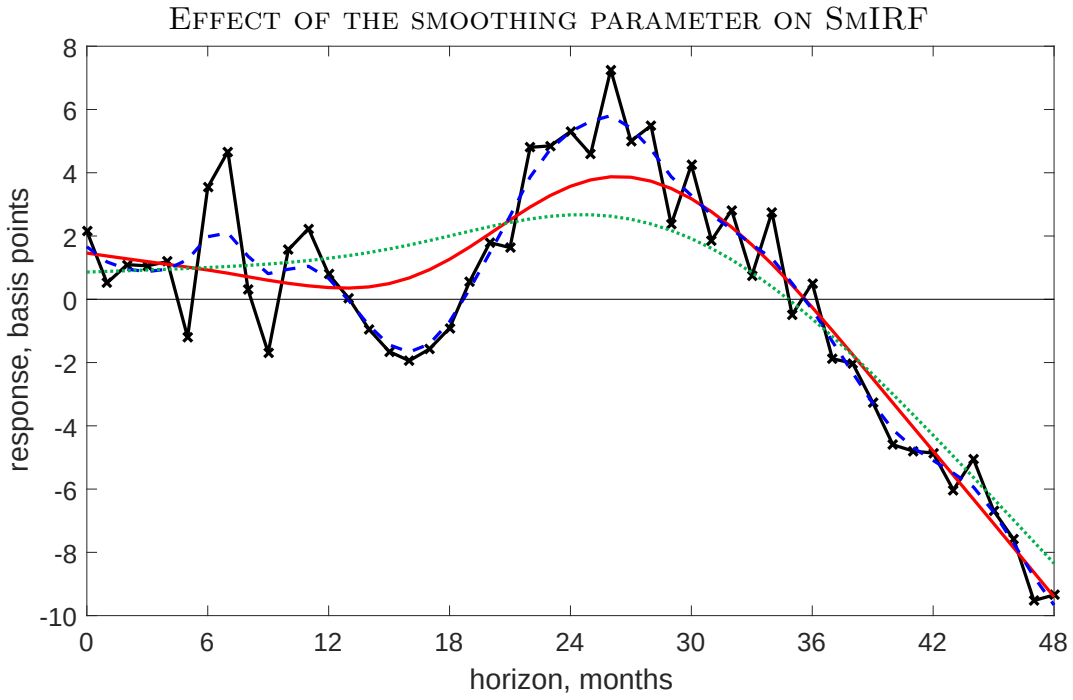

Figure 3.2: SmIRF estimator for $\lambda = 0$ (solid line, crosses), 2 (dashed), 298 (solid, no crosses), and 2000 (dotted). $\lambda = 298$ is optimal in the sense of Section 3.2.3. See caption for Figure 3.1.

below). As mentioned in the Introduction, the quadratic second difference penalty was used by Shiller (1973) to produce smooth distributed lag regression estimates. As with many penalized estimators, the SmIRF estimator can be interpreted as a Bayesian posterior mean, in this case using Shiller's smoothness prior.

Figure 3.2 illustrates how larger values of the smoothing parameter λ impose increasing amounts of smoothness on the SmIRF estimator. The optimal amount of smoothness to impose depends on how fast the bias increases as the IRF estimator is smoothed further, which in turn depends on the unknown smoothness of the true IRF. Section 3.2.3 below shows how to select λ in a data-dependent way to optimally trade off bias and variance.[7]

[7]In the setting of Footnote 6, Shiller (1973, p. 779) suggests the rule of thumb $\lambda = n/\sqrt{8S\sigma_{\tilde{x}}^2}$ if the IRF is expected to be symmetrically tent-shaped with cumulative impulse response S across the n horizons.

General shrinkage class. The SmIRF estimator defined above is a special case of the class of general shrinkage estimators given by

$$\hat{\beta}_{M,W}(\lambda) = \underset{\beta \in \mathbb{R}^n}{\arg\min} \ \|\beta - \hat{\beta}\|_W^2 + \lambda \|M\beta\|^2 = \Theta_{M,W}(\lambda)\hat{\beta}, \tag{3.3}$$

where $\|v\|_W^2 = v'Wv$ and $\|v\|^2 = v'v$ for any vector $v \in \mathbb{R}^n$. M is an $m \times n$ matrix, W is an $n \times n$ symmetric positive definite weight matrix, and

$$\Theta_{M,W}(\lambda) = (I_n + \lambda W^{-1} M'M)^{-1}. \tag{3.4}$$

The rows of the matrix M determine which linear combinations of the impulse responses to penalize. The weight matrix W down-weights certain impulse responses relative to others in the fit to the initial IRF estimate.

The SmIRF estimator (3.2) obtains when $W = I_n$ and

$$M = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{(n-2)\times n}, \tag{3.5}$$

the second difference matrix. If M equals the above second difference matrix with the first row deleted, the impact impulse response β_0 will not be penalized, which is desirable if the impact response is of special interest. Being essentially a version of ridge regression (Vinod, 1978), general shrinkage estimators are easy to compute, as the explicit expression (3.3) shows.

General shrinkage estimators (3.3) can flexibly impose not only smoothness, but also short-run and long-run approximate restrictions on the IRFs. For example, if M equals the

first unit vector, the contemporaneous impulse response is shrunk towards zero. If M equals the row vector consisting of ones, the shrinkage estimator shrinks the cumulative sum of the IRF towards 0. It is computationally straight-forward to introduce multiple penalty terms with separate shrinkage parameters λ_k, although I do not consider that possibility here.

PROJECTION SHRINKAGE CLASS. An analytically convenient subclass of estimators are the projection shrinkage estimators. These estimators are obtained from the general class (3.3) by choosing weight matrix $W = I_n$ and penalty matrix $M = P$, where P is an $n \times n$ orthogonal projection matrix, i.e., a symmetric matrix satisfying $P^2 = P$ (idempotence). As shown in Appendix A.3.3, the projection shrinkage estimator $\hat{\beta}_P(\lambda) := \hat{\beta}_{P,I_n}(\lambda)$ can be written

$$\begin{aligned}\hat{\beta}_P(\lambda) &= \underset{\beta\in\mathbb{R}^n}{\arg\min}\left\{\|P\beta - P\hat{\beta}\|^2 + \|(I_n - P)\beta - (I_n - P)\hat{\beta}\|^2 + \lambda\|P\beta\|^2\right\} \qquad (3.6)\\ &= \frac{1}{1+\lambda}P\hat{\beta} + (I_n - P)\hat{\beta}.\end{aligned}$$

In words, the projection shrinkage estimator shrinks towards zero the projection of the initial IRF estimate $\hat{\beta}$ onto the space spanned by the matrix P, while the projection of $\hat{\beta}$ onto the orthogonal complement of this space is unchanged.

Projection shrinkage estimators can be designed to shrink the initial IRF estimate towards a polynomial, or any other linear-in-parameters function of the response horizon. Suppose we have a prior belief that the true IRF is likely to look similar to the function $l(i) = \sum_{k=0}^{p} a_k b_k(i)$ of the response horizon i, for some unknown constants $a_0, \ldots, a_p$, and some user-specified basis functions $b_0(\cdot), \ldots, b_p(\cdot)$. For example, we can consider polynomials with $b_k(i) = i^k$. If we set $P = I_n - L(L'L)^{-1}L'$, where L is the $n \times (p+1)$ matrix whose $(k+1)$-th column equals $(b_k(0), b_k(1), \ldots, b_k(n-1))'$, then the term $\lambda\|P\beta\|^2$ in the projection shrinkage objective function (3.6) penalizes deviations of the IRF from functions of the form $l(i)$ (for some

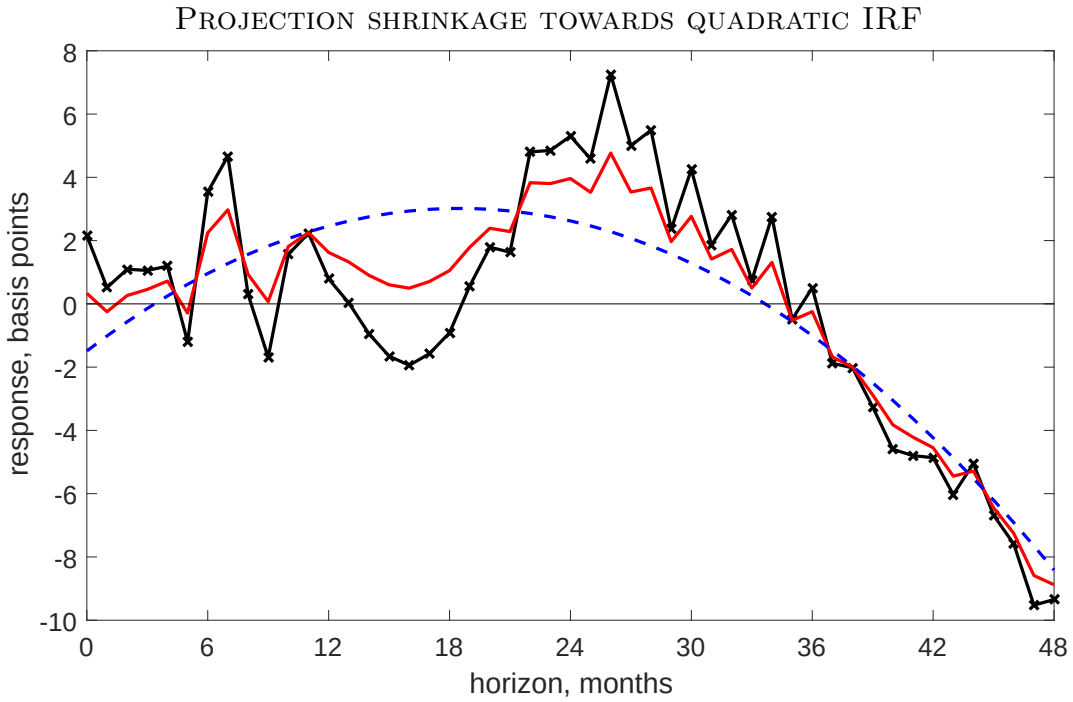

Figure 3.3: Initial IRF estimate (jagged line, crosses), best-fitting quadratic IRF (dashed), and projection shrinkage estimator with $\lambda = 1$ (solid, no crosses). See caption for Figure 3.1.

constants $a_0, \ldots, a_p$). Hence, the projection shrinkage estimator $\hat{\beta}_P(\lambda)$ shrinks the initial IRF estimate $\hat{\beta}$ towards the IRF $\{\sum_{k=0}^{p} \hat{a}_k b_k(i)\}_{0 \le i \le n-1}$, where $\hat{a}_k$ are the least-squares coefficients in a regression of $\hat{\beta}_0, \ldots, \hat{\beta}_{n-1}$ on the columns of L.

In Figure 3.3 the initial IRF estimate is shrunk towards a quadratic function. This procedure does not produce as smooth-looking IRFs as the basic SmIRF estimator (3.2). Nevertheless, it achieves the same goal of reducing the variance of the initial IRF estimator by using global features of the IRF to discipline the estimate at each response horizon.

3.2.3 Unbiased risk estimate

I now propose a criterion for selecting the smoothing parameter in a data-dependent way to minimize the MSE of a general shrinkage estimator, such as SmIRF. Let $\beta^\dagger$ denote the true IRF, i.e., the value that the initial estimator $\hat{\beta}$ is approximately unbiased for. To trade off

bias and variance, we would like to choose $\lambda \geq 0$ to minimize the weighted MSE criterion

$$R_{M,W,\tilde{W}}(\lambda) = T\,E\left[\|\hat{\beta}_{M,W}(\lambda) - \beta^{\dagger}\|^2_{\tilde{W}}\right],$$

where $\tilde{W}$ is a user-specified $n \times n$ symmetric positive definite weight matrix (in most applications, $\tilde{W} = W$). The sample size is denoted T, but the underlying data need not be time series data. The above expectation averages over the unknown sampling distribution of the initial estimator $\hat{\beta}$, so $R_{M,W,\tilde{W}}(\lambda)$ cannot be used to choose λ in practice.

To estimate the MSE of general shrinkage estimators, I assume we have available an initial IRF estimator $\hat{\beta}$ and a consistent estimator $\hat{\Sigma}$ of its asymptotic variance. As formally defined in Section 3.3.1, $\hat{\beta}$ must be approximately unbiased and asymptotically normal, and $\hat{\Sigma}$ must be consistent for the asymptotic variance $\Sigma = \lim_{T\to\infty} E[T(\hat{\beta} - \beta^{\dagger})(\hat{\beta} - \beta^{\dagger})']$. It is well known how to construct such estimators if $\hat{\beta}$ is obtained from time series or panel regression or similar methods.[8]

The unbiased risk estimate (URE) is an asymptotically uniformly unbiased estimator of the true MSE, up to a constant that does not depend on λ, as shown in Section 3.3.1:

$$\hat{R}_{M,W,\tilde{W}}(\lambda) = T\|\hat{\beta}_{M,W}(\lambda) - \hat{\beta}\|^2_{\tilde{W}} + 2\,\mathrm{tr}\left\{\tilde{W}\Theta_{M,W}(\lambda)\hat{\Sigma}\right\}, \tag{3.7}$$

where "tr" denotes the trace of a matrix. The URE depends on the data only through $\hat{\beta}$ and $\hat{\Sigma}$. The first term in expression (3.7) measures the in-sample fit of the shrinkage estimator relative to the initial IRF estimate. The second term penalizes small values of $\lambda \geq 0$, since such values lead to a high-variance shrinkage estimator for which the in-sample

[8] In the case of panel regression, $\hat{\Sigma}$ will be a clustered variance estimator; in the case of time series regression, a heteroskedasticity and autocorrelation consistent (HAC) estimator. In some applications, the shock x_t is obtained from a preliminary estimation procedure, e.g., as a residual. $\hat{\Sigma}$ should then reflect the additional estimation uncertainty due to the generated regressor (Pagan, 1984).

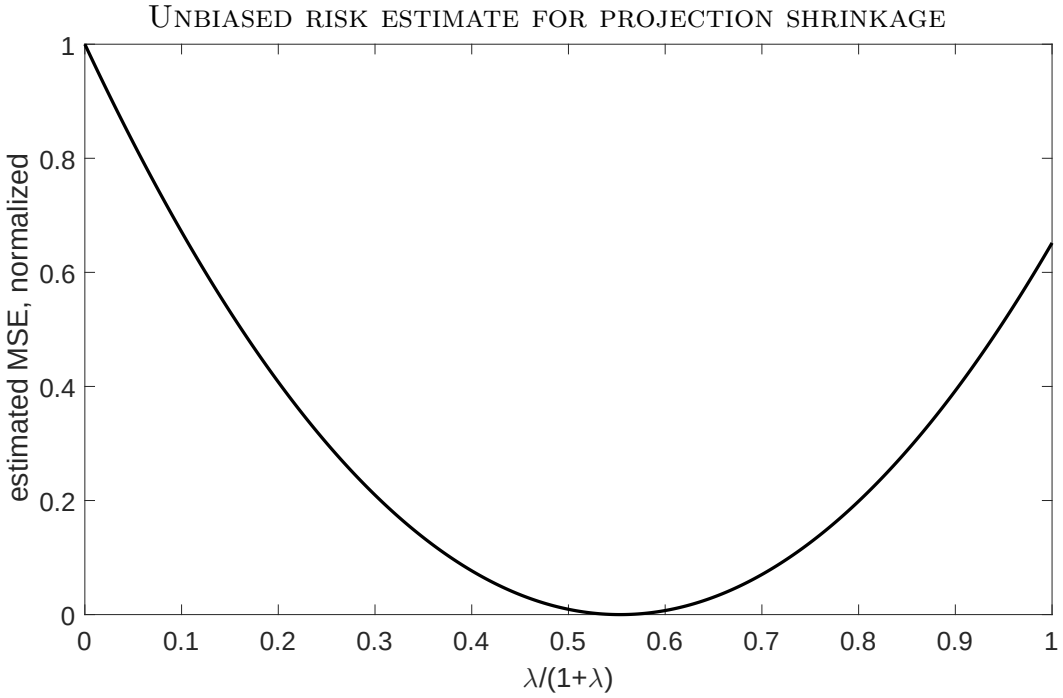

Figure 3.4: URE criterion to optimally select the smoothing parameter λ for the projection shrinkage estimator in Figure 3.3. The horizontal axis plots $\lambda/(1+\lambda)$, not λ. MSE estimates on the vertical axis are normalized to $[0, 1]$. HAC: Newey-West, 24 lags. See caption for Figure 3.1.

fit relative to $\hat{\beta}$ is an overoptimistic measure of out-of-sample performance relative to the truth $\beta^\dagger$. Similar ideas underlie many model selection criteria (Claeskens & Hjort, 2008; Hansen, 2010). Appendix A.3.4 shows that the URE can be rewritten as a sum of unbiased estimates of the variance and the squared bias of the shrinkage estimator.

Figure 3.4 plots the URE corresponding to the projection shrinkage estimator in Figure 3.3. It is straight-forward to compute the URE (3.7) over a fine grid of λ values for plotting purposes. The minimizing λ value can be computed using one-dimensional numerical optimization. The shape of the URE criterion is informative about how sensitively the MSE performance of the estimator depends on the smoothing parameter.

I propose selecting the shrinkage parameter that minimizes the URE. For penalty matrix M and weight matrices $W, \tilde{W}$, the optimal shrinkage estimator is $\hat{\beta}_{M,W,\tilde{W}} := \hat{\beta}_{M,W}(\hat{\lambda}_{M,W,\tilde{W}})$, where $\hat{\lambda}_{M,W,\tilde{W}}$ is the URE-minimizing λ. Simulations in Section 3.5 indicate that this esti-

mator often has lower weighted MSE than the initial IRF estimator in realistic settings.

The minimum-URE estimator has a simple form with provably desirable MSE properties in the case of projection shrinkage and $W = \tilde{W} = I_n$. Appendix A.3.3 shows that, in this case, the URE is a quadratic function in $\lambda/(1+\lambda)$. The minimum-URE projection shrinkage estimator (restricting $\lambda \geq 0$) equals

$$\hat{\beta}_{P,I_n,I_n} = \left(1 - \frac{\operatorname{tr}(\hat{\Sigma}_P)}{T\|P\hat{\beta}\|^2}\right)_+ P\hat{\beta} + (I_n - P)\hat{\beta} = \hat{\beta} - \min\left\{\frac{\operatorname{tr}(\hat{\Sigma}_P)}{T\|P\hat{\beta}\|^2}, 1\right\} P\hat{\beta}, \tag{3.8}$$

where $\hat{\Sigma}_P = P\hat{\Sigma}P$ and $x_+ = \max\{x, 0\}$ for any $x \in \mathbb{R}$. Expression (3.8) illustrates that choosing λ to minimize the URE makes shrinkage estimation adaptive: The amount of shrinkage applied to the initial estimate $\hat{\beta}$ depends on the extent to which the data is compatible with the shrinkage hypothesis, in this case through the ratio $T\|P\hat{\beta}\|^2/\operatorname{tr}(\hat{\Sigma}_P)$. As a consequence, Section 3.3.2 proves that the optimal projection shrinkage estimator uniformly dominates the initial IRF estimator in terms of MSE under realistic conditions. An additional attractive feature of the optimal shrinkage estimator is that it depends on the asymptotic variance estimate $\hat{\Sigma}$ only through the scalar $\operatorname{tr}(\hat{\Sigma}_P)$.[9]

3.2.4 Confidence bands

Confidence bands for general shrinkage estimators can be constructed by a test inversion procedure that takes into account the shrinkage bias and the randomness induced by a data-dependent shrinkage parameter.[10] Here I provide recipes for constructing joint and pointwise confidence bands. As discussed in detail in Sections 3.4 and 3.5, the proposed

[9]This fact is especially helpful if $\hat{\beta}$ is obtained from a time series regression, as $\hat{\Sigma}$ will then typically be a HAC estimator with limited accuracy in small samples (Müller, 2014).

[10]If the shrinkage parameter λ has been picked based on introspection before seeing the data, as in Footnote 7, a quick-and-dirty confidence band can be constructed using $\operatorname{Var}(\sqrt{T}\hat{\beta}_{M,W}(\lambda)) \approx \Theta_{M,W}(\lambda)\hat{\Sigma}\Theta_{M,W}(\lambda)'$ for fixed λ. This procedure is only accurate for small λ because it ignores the shrinkage bias.

shrinkage bands do not have uniformly smaller area than the usual Wald confidence bands centered at the initial IRF estimate $\hat{\beta}$. Nevertheless, simulation evidence indicates that the bands perform well when the true IRF is smooth, while apparently never doing substantially worse than the standard bands. Unlike the standard bands, the shrinkage bands can be constructed so they always contain the shrinkage estimator.

POINTWISE BANDS. Pointwise confidence bands guarantee a pre-specified asymptotic coverage probability in repeated experiments, considering each impulse response separately. Pointwise bands are commonly used in applied macroeconomics and panel event studies.

The pointwise shrinkage confidence bands are based on simulated critical values obtained horizon by horizon. Suppose we seek a confidence set for the linear combination $s'\beta^\dagger$ of the IRF parameters, where s is a user-specified selection vector (most commonly a unit vector). First, for any $\eta \in \mathbb{R}^n$ and $n \times n$ symmetric positive definite matrix Σ, define[11]

$$\hat{\theta}_{M,W,\tilde{W}}(\eta, \Sigma) = \Theta_{M,W}(\hat{\lambda}_{M,W,\tilde{W}}(\eta, \Sigma))\eta, \tag{3.9}$$

$$\hat{\lambda}_{M,W,\tilde{W}}(\eta, \Sigma) = \arg\min_{\lambda \geq 0} \left(\|\{\Theta_{M,W}(\lambda) - I_n\}\eta\|^2_{\tilde{W}} + 2\,\mathrm{tr}\{\tilde{W}\Theta_{M,W}(\lambda)\Sigma\} \right). \tag{3.10}$$

Second, for any $\theta \in \mathbb{R}^n$ and Σ, let $\zeta = (s'\Sigma s)^{-1}\Sigma s$, and define $q_{s,1-\alpha,M,W,\tilde{W}}(\theta, \Sigma)$ to be the $1-\alpha$ quantile of the distribution of

$$\{s'\hat{\theta}_{M,W,\tilde{W}}(\zeta u + \theta, \Sigma) - s'\theta\}^2, \tag{3.11}$$

where $u \sim N(0, s'\Sigma s)$. This quantile can be approximated arbitrarily well for given η and Σ by repeatedly simulating draws of u. The simulations run faster for projection shrinkage

[11] If the minimum (3.10) is attained at $\lambda = \infty$, set $\hat{\theta}_{M,W,\tilde{W}}(\eta, \Sigma) = \lim_{\lambda \to \infty} \Theta_{M,W}(\lambda)\eta$. The limit equals $(I_n - W^{-1}M'(MW^{-1}M')^{-1}M)\eta$ if M has full row rank, and $(I_n - M)\eta$ if M is an orthogonal projection.

estimators because the minimizer (3.10) is available in closed form, cf. Appendix A.3.3.

A shrinkage confidence set with asymptotic $1-\alpha$ coverage rate for $s'\beta^\dagger$ is given by

$$\hat{\mathcal{C}}_{s,1-\alpha} = \left\{\mu \in \mathbb{R} \colon T(s'\hat{\beta}_{M,W,\tilde{W}} - \mu)^2 \leq q_{s,1-\alpha,M,W,\tilde{W}}(\sqrt{T}(\hat{\zeta}\mu + \hat{\nu}), \hat{\Sigma})\right\},$$

where $\hat{\zeta} = (s'\hat{\Sigma}s)^{-1}\hat{\Sigma}s$ and $\hat{\nu} = (I_n - \hat{\zeta}s')\hat{\beta}$. To approximate the set, consider a fine grid of μ values containing $s'\hat{\beta}_{M,W,\tilde{W}}$, and evaluate the above inequality at each point in the grid. While more computationally intensive than the usual Wald interval, the grid search is fast for the case of projection shrinkage and not too onerous in the general case. In the case of projection shrinkage, Appendix A.3.3 shows that the grid search can be confined to a bounded interval: $\hat{\mathcal{C}}_{s,1-\alpha} \subset [s'\hat{\beta}_{P,I_n,I_n} - \hat{\xi}, s'\hat{\beta}_{P,I_n,I_n} + \hat{\xi}]$, where $\hat{\xi} = \sqrt{(s'\hat{\Sigma}s/T)z_{1,1-\alpha}} + \|s\|\sqrt{\mathrm{tr}(\hat{\Sigma})/T}$ and $z_{1,1-\alpha}$ is the $1-\alpha$ quantile of a $\chi^2(1)$ distribution.

The proposed set $\hat{\mathcal{C}}_{s,1-\alpha}$ always contains the shrinkage point estimate $s'\hat{\theta}_{M,W,\tilde{W}}$, but an alternative shrinkage set without this guarantee is often faster to compute. The alternative procedure involves a user-specified tuning parameter $\delta \in [0, \alpha]$. Define the usual Wald interval for $s'\beta^\dagger$ at level $1-\delta$: $\hat{\mathcal{I}}_{s,1-\delta} = [s'\hat{\beta} - \hat{c}_{1-\delta}/\sqrt{T}, s'\hat{\beta} + \hat{c}_{1-\delta}/\sqrt{T}]$, where $\hat{c}_{1-\delta} = \sqrt{(s'\hat{\Sigma}s)z_{1,1-\delta}}$. Define also $\tilde{q}_{s,1-\alpha,M,W,\tilde{W}}(\theta, \Sigma, c)$ to be the $1-\alpha$ quantile of the statistic (3.11) when u has a truncated normal distribution with mean 0, variance parameter $s'\Sigma s$ and truncation interval $|u| < c$ (this quantile can be computed by simulation). Finally, define the alternative level $1-\alpha$ shrinkage set[12]

$$\hat{\mathcal{C}}_{s,1-\alpha,1-\delta} = \left\{\mu \in \hat{\mathcal{I}}_{s,1-\delta} \colon T(s'\hat{\beta}_{M,W,\tilde{W}} - \mu)^2 \leq \tilde{q}_{s,\frac{1-\alpha}{1-\delta},M,W,\tilde{W}}(\sqrt{T}(\hat{\zeta}\mu + \hat{\nu}), \hat{\Sigma}, \hat{c}_{1-\delta})\right\}.$$

The alternative shrinkage set is always contained in the $(1-\delta)$-level Wald interval $\hat{\mathcal{I}}_{s,1-\delta}$,

[12]The right-hand side of the inequality in the definition of $\hat{\mathcal{C}}_{s,1-\alpha,1-\delta}$ refers to the $\frac{1-\alpha}{1-\delta}$ quantile, as the use of the auxiliary confidence interval $\hat{\mathcal{I}}_{s,1-\delta}$ necessitates an adjustment of the critical value, cf. Section 3.4.

which limits the worst-case length of the confidence set but implies that the set does not always contain the shrinkage estimate. A higher value for the tuning parameter δ yields a smaller worst-case length but a higher probability of not containing the shrinkage estimate. I suggest the default value $\delta = \alpha/10$ as in McCloskey (2015, Sec. 3.5).

Figures 3.5 and 3.6 show that the pointwise projection shrinkage band can be narrower at most horizons than the usual pointwise band centered at the local projection estimator. Figure 3.5 draws a pointwise 90% confidence band based on the alternative shrinkage set with $\delta = 0.01$. While the shrinkage band is not very different from the usual Wald band in Figure 3.6, the former is slightly narrower at most horizons. Although not guaranteed, here the alternative shrinkage set actually does contain the shrinkage estimator at all horizons.

JOINT BANDS. A joint confidence band of asymptotic level $1-\alpha$ covers the true IRF at *all* horizons with probability $1-\alpha$ in repeated experiments, for large sample sizes. Joint bands are needed when testing whether the entire IRF is zero, or for hypothesis tests concerning the overall shape of the IRF. Sims & Zha (1999) and Inoue & Kilian (2016) recommend the use of joint bands instead of pointwise bands for macroeconomic applications.

I construct joint shrinkage bands by inverting a statistic with simulated critical values, as in the pointwise case. For any $\theta = (\theta_0, \ldots, \theta_{n-1})' \in \mathbb{R}^n$ and $n \times n$ symmetric positive definite Σ, let $q_{1-\alpha,M,W,\tilde{W}}(\theta, \Sigma)$ be the $1-\alpha$ quantile of the distribution of

$$\sup_{0 \le i \le n-1} \left| \Sigma_{ii}^{-1/2} \left(\hat{\theta}_{i,M,W,\tilde{W}}(\theta + U, \Sigma) - \theta_i \right) \right|,$$

where $\hat{\theta}_{i,M,W,\tilde{W}}(\eta, \Sigma)$ is the $(i+1)$-th element of (3.9), Σ_{ii} is the i-th diagonal element of Σ, and $U \sim N(0, \Sigma)$. This quantile can be computed by repeated simulation of U. Finally,

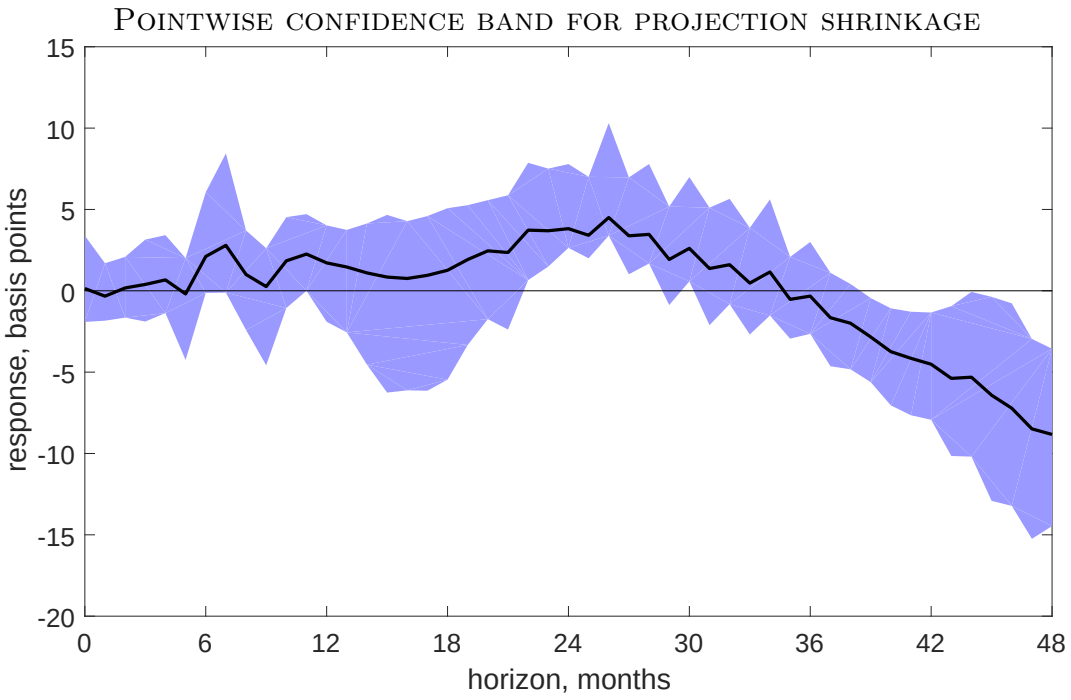

Figure 3.5: Projection shrinkage point estimate (thick line) and 90% pointwise confidence band $\hat{\mathcal{C}}_{s,0.9,1-\delta}$ for $\delta = 0.01$ (shaded). HAC: Newey-West, 24 lags. See caption for Figure 3.1.

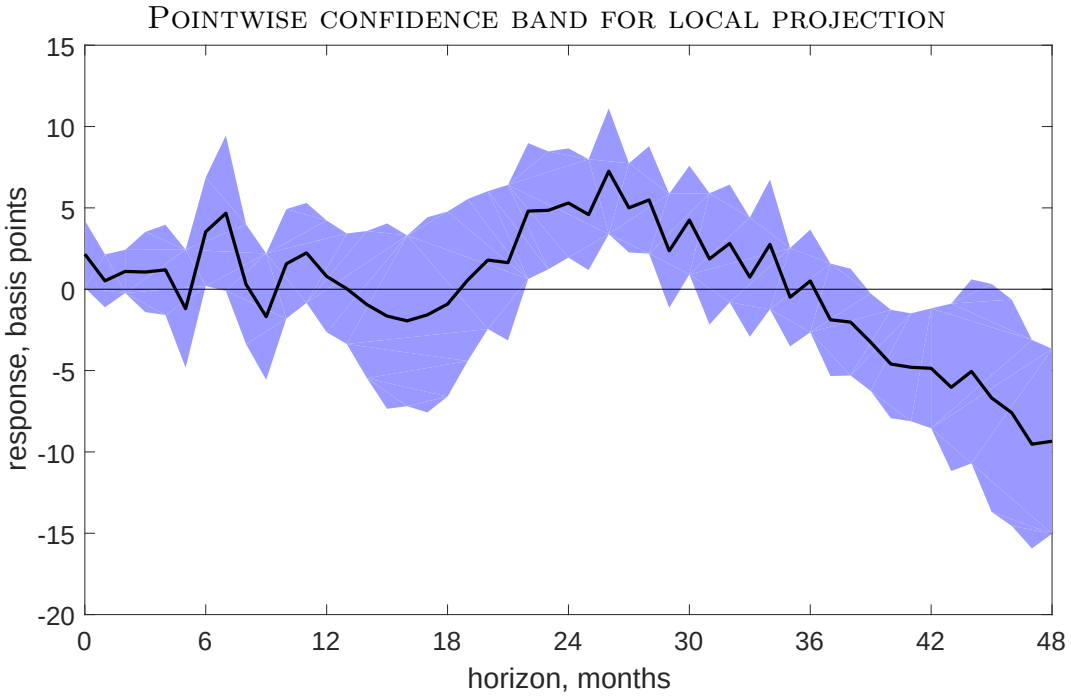

Figure 3.6: Local projection point estimate (thick line) and 90% pointwise confidence band (shaded). HAC: Newey-West, 24 lags. See caption for Figure 3.1.

define the joint level $1-\alpha$ shrinkage confidence set[13]

$$\hat{\mathcal{C}}_{1-\alpha} = \left\{ (\beta_0, \ldots, \beta_{n-1})' \in \mathbb{R}^n \colon \sup_{0 \leq i \leq n-1} \sqrt{T} \left| \hat{\Sigma}_{ii}^{-1/2} \left(\hat{\beta}_{i,M,W,\tilde{W}} - \beta_i \right) \right| \leq q_{1-\alpha,M,W,\tilde{W}}(\sqrt{T}\beta, \hat{\Sigma}) \right\}.$$

The joint shrinkage band can be computed numerically by an accept/reject procedure, as with the "shotgun plots" in Inoue & Kilian (2016): Simulate draws of $\beta = (\beta_0, \ldots, \beta_{n-1})'$ from some proposal distribution and retain them if they satisfy the inequality in the definition of $\hat{\mathcal{C}}_{1-\alpha}$. If the proposal distribution has full support, this procedure will exhaust the joint confidence band as the number of draws tends to infinity. I suggest using the proposal distribution $\hat{\beta}_{M,W,\tilde{W}} + \sqrt{z_{1,1-\alpha}/T}\hat{\Sigma}^{1/2}\tilde{U}$, where $\tilde{U}$ is an n-dimensional vector consisting of i.i.d. t-distributed elements with few degrees of freedom, e.g., 5.

Figure 3.7 depicts draws from a joint confidence band around the projection shrinkage estimator. Even at the 68% confidence level used by Inoue & Kilian (2016), the joint band is wide and contains a variety of differently shaped IRFs. The uncertainty about the shape of the tail of the IRF is particularly high.

3.3 Mean squared error optimality

In this section I present theoretical results on the MSE of shrinkage estimators. First, I give conditions under which the URE is asymptotically unbiased for the true MSE of general shrinkage estimators. Then I show that projection shrinkage estimators can achieve uniformly lower asymptotic MSE than the unconstrained estimator when the smoothing parameter is chosen to minimize the URE.

I assume that the initial non-smooth IRF estimator $\hat{\beta}$ is consistent for the true IRF and

[13]Inoue & Kilian (2016) construct joint confidence bands based on the Wald set. I prefer the weighted supremum metric in the definition of $\hat{\mathcal{C}}_{1-\alpha}$, since the Euclidean norm used by the Wald set allows large deviations from the point estimate at some horizons, provided the deviations are small at other horizons.

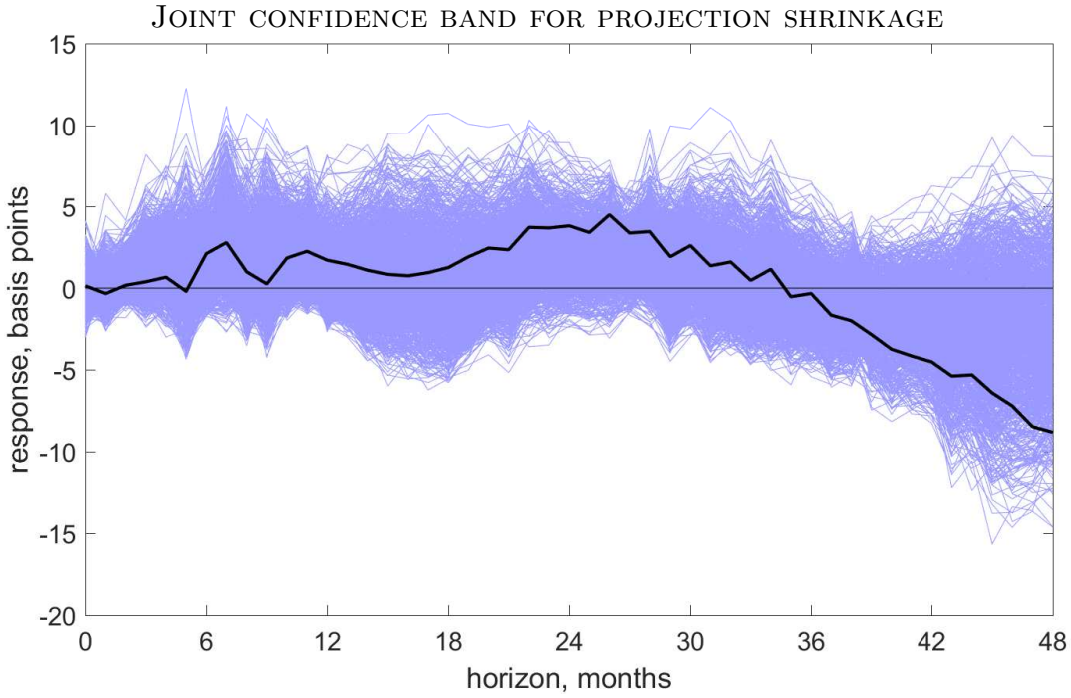

Figure 3.7: Projection shrinkage point estimate (thick line) and draws from 68% joint confidence band (thin lines). 10,000 t-distributed proposal draws, 5 d.f.; 1,251 draws accepted. HAC: Newey-West, 24 lags. See caption for Figure 3.1.

asymptotically normal, with consistently estimable asymptotic variance.

Assumption 3.1. *The distribution of the data for sample size T is indexed by a parameter $\beta_T^\dagger \in \mathbb{R}^n$. The estimators $\hat{\beta} \in \mathbb{R}^n$ and $\hat{\Sigma} \in \mathbb{S}^n$ satisfy $\sqrt{T}(\hat{\beta} - \beta_T^\dagger) \xrightarrow{d} N(0, \Sigma)$ and $\hat{\Sigma} \xrightarrow{p} \Sigma$ for some $\Sigma \in \mathbb{S}_+^n$, and the sequence $\{T\|\hat{\beta} - \beta_T^\dagger\|^2 + \|\hat{\Sigma}\|\}_{T \geq 1}$ is uniformly integrable.*

The assumptions on the estimators $\hat{\beta}$ and $\hat{\Sigma}$ are standard. The notation indicates that the IRF $\beta_T^\dagger$ may depend on the sample size, which is convenient when stating the results in this section. The parameter $\beta_T^\dagger$ is pseudo-true, in the sense that $\hat{\beta} - \beta_T^\dagger \xrightarrow{p} 0$, but otherwise the parameter may have no direct connection to the underlying data generating model. The uniform integrability assumption is implied by $\|\sqrt{T}(\hat{\beta} - \beta_T^\dagger)\|$ having uniformly bounded $2 + \varepsilon$ moment and $\|\hat{\Sigma}\|$ having uniformly bounded $1 + \varepsilon$ moment for sufficiently large T.[14]

[14]Uniform integrability essentially rules out cases where, for example, $\hat{\beta}$ does not have finite moments for

3.3.1 Unbiased risk estimate

I now justify the name "unbiased risk estimate" by proving that the URE criterion (3.7) is asymptotically uniformly unbiased for the MSE of a general shrinkage estimator.

I restrict attention to IRFs that are moderately smooth in an asymptotic sense, since otherwise shrinkage does not matter asymptotically. Let M be the matrix defining the penalty term in the general shrinkage estimator (3.3), and define the projection matrix $P_M = M'(MM')^{-1}M$ (if $M = P$ is itself a projection matrix, set $P_M = P$). I assume below that $\lim_{T\to\infty} \|\sqrt{T}P_M\beta_T^\dagger\| < \infty$, following Hansen's (2016b) insight that only such local asymptotic sequences generate a nontrivial role for shrinkage asymptotically. If instead $\liminf_{T\to\infty} \|\sqrt{T}P_M\beta_T^\dagger\| = \infty$, one can show that both the true MSE $R_{M,W,\tilde{W}}(\lambda)$ and the URE $\hat{R}_{M,W,\tilde{W}}(\lambda)$ tend to infinity asymptotically for any $\lambda > 0$, which is an uninteresting conclusion from an applied perspective.

The following proposition states the asymptotic uniform unbiasedness of the URE criterion, up to a constant. Since this constant does not depend on λ, it is irrelevant for the purposes of selecting the smoothing parameter.

Proposition 3.1. *Let Assumption 3.1 hold. Assume that either: (a) $M \in \mathbb{R}^{m\times n}$ has full row rank, or (b) $M = P \in \mathbb{R}^{n\times n}$ is an orthogonal projection matrix. Define $P_M = M'(MM')^{-1}M$ in case (a) or $P_M = P$ in case (b). Assume that $\sqrt{T}P_M\beta_T^\dagger \to h \in \mathbb{R}^n$. Let $W, \tilde{W} \in \mathbb{S}_+^n$. Then there exists a random variable $\hat{C}$ that does not depend on λ such that*

$$\lim_{T\to\infty} \sup_{\lambda\geq 0} \left| R_{M,W,\tilde{W}}(\lambda) - E\left(\hat{R}_{M,W,\tilde{W}}(\lambda) + \hat{C}\right)\right| = 0.$$

---

any finite T. The assumption can probably be relaxed at the expense of a more complicated statement in Proposition 3.1, and a trimmed-risk statement in Proposition 3.2 similar to Hansen (2016b, Thm. 1).

3.3.2 Risk of projection shrinkage

The proposition below gives conditions under which projection shrinkage estimators have small MSE relative to the initial IRF estimator. Given an orthogonal projection matrix $P \in \mathbb{R}^{n\times n}$, the MSE dominance result applies to the class of shrinkage estimators

$$\hat{\beta}_P(\tau) = \hat{\beta} - \min\left\{\frac{\tau}{T\|P\hat{\beta}\|^2}, 1\right\} P\hat{\beta}, \quad \tau \geq 0,$$

where I have abused notation slightly relative to definition (3.6) (note that here I use τ instead of λ for the argument). The optimal projection shrinkage estimator (3.8) is a member of the above class with $\tau = \mathrm{tr}(\hat{\Sigma}_P)$. The following proposition is a minor extension of results in Oman (1982) and Hansen (2016b). For the same reason as in the previous subsection, I restrict attention to a $1/\sqrt{T}$ neighborhood of the shrinkage space.

Proposition 3.2 (Oman, 1982; Hansen, 2016b)**.** *Let Assumption 3.1 hold, and assume* $\sqrt{T}P\beta_T^\dagger \to h \in \mathbb{R}^n$. *Let* $\hat{\tau} \geq 0$ *be a scalar random variable satisfying* $\hat{\tau} \xrightarrow{p} \tau \leq 2(\mathrm{tr}(\Sigma) - 2\rho(\Sigma))$ *and such that the sequence* $\{T\|\hat{\beta} - \beta_T^\dagger\|^2 + \hat{\tau}\}_{T\geq 1}$ *is uniformly integrable. Define* $\Sigma_P = P\Sigma P$. *Then*

$$\limsup_{T\to\infty} E\left(T\|\hat{\beta}_P(\hat{\tau}) - \beta_T^\dagger\|^2\right) \leq \mathrm{tr}(\Sigma) - \tau\frac{2(\mathrm{tr}(\Sigma_P) - 2\rho(\Sigma_P)) - \tau}{\mathrm{tr}(\Sigma_P) + \|h\|^2}.$$

The result shows that if $\mathrm{plim}_{T\to\infty}\hat{\tau} \leq 2(\mathrm{tr}(\Sigma_P) - 2\rho(\Sigma_P))$, the limiting MSE of the shrinkage estimator $\hat{\beta}_P(\hat{\tau})$ is less than that of the initial IRF estimator $\hat{\beta} = \hat{\beta}_P(0)$, uniformly in a $1/\sqrt{T}$ neighborhood of the shrinkage space $\mathrm{span}(I_n - P)$. The first term on the right-hand side above equals the limiting MSE of the initial estimator. The fraction on the right-hand side above is maximized at $\tau = \mathrm{tr}(\Sigma_P) - 2\rho(\Sigma_P)$; however, it does not follow that this is the MSE-optimal probability limit for $\hat{\tau}$, as the right-hand side is only an upper bound.

The proposition implies conditions under which the URE-minimizing projection shrinkage

estimator dominates the initial IRF estimator.

Corollary 3.1. *Let Assumption 3.1 hold, and assume* $\sqrt{T}P\beta_T^\dagger \to h \in \mathbb{R}^n$. *If* $\mathrm{tr}(\Sigma_P) > 4\rho(\Sigma_P)$, *the URE-minimizing projection shrinkage estimator* $\hat{\beta}_P(\hat{\tau})$ *with* $\hat{\tau} = \mathrm{tr}(\hat{\Sigma}_P)$ *has smaller limiting MSE than the initial IRF estimator* $\hat{\beta}$, *uniformly in* h.

The sufficient condition $\mathrm{tr}(\Sigma_P) > 4\rho(\Sigma_P)$ in Corollary 3.1 requires that the number of response horizons of interest is sufficiently high. Since $\mathrm{tr}(\Sigma_P) \leq \mathrm{rk}(P)\rho(\Sigma_P)$, a necessary condition for $\mathrm{tr}(\Sigma_P) > 4\rho(\Sigma_P)$ is that $\mathrm{rk}(P) > 4$. If the projection shrinkage estimator is used to shrink the initial IRF estimate towards a polynomial of order p, as explained in Section 3.2.2, then $\mathrm{rk}(P) = n - p$, so the necessary condition is $n > p + 4$. The higher the order of the polynomial, the harder is it to MSE-dominate the initial IRF estimator for given number of horizons n, because even the shrunk estimate will have high variance. The more response horizons the researcher cares about in the MSE criterion, the more is shrinkage likely to be beneficial. See Oman (1982) for a general discussion of $\mathrm{tr}(\Sigma_P)$ versus $\rho(\Sigma_P)$.

3.4 Confidence sets

Here I develop novel methods for computing joint and marginal confidence sets based on the shrinkage estimators in Section 3.2. For expositional convenience, I start off by constructing valid confidence sets in an idealized finite-sample model in which the initial IRF estimate is exactly normally distributed with arbitrary known covariance structure. Then I show that, for the case of projection shrinkage, the finite-sample results imply that the confidence sets achieve asymptotic uniform coverage under weak assumptions.

3.4.1 Finite-sample normal model

In this subsection I construct valid test inversion confidence sets in a finite-sample normal model with arbitrary known covariance matrix. The finite-sample normal model is the

appropriate limit experiment for shrinkage estimators, in a sense that is made formal in the next subsection for the special case of projection shrinkage. Results in this section motivate the use of the confidence sets introduced in Section 3.2.4.

MODEL. Assume that we observe a single draw $\hat{\theta} \sim N(\theta^\dagger, \Sigma)$, where $\theta^\dagger \in \mathbb{R}^n$ is the unknown parameter of interest, and $\Sigma \in \mathbb{S}_+^n$ is an arbitrary known covariance matrix. This idealized model has received extensive attention in the literature on shrinkage confidence sets, although typically under the additional assumption that Σ is spherical (Casella & Hwang, 2012). To map into the notation of the previous sections, think of $\hat{\theta} = \sqrt{T}\hat{\beta}$ and $\theta^\dagger = \sqrt{T}\beta^\dagger$, where the conditions in Assumption 3.1 of asymptotic normality and consistently estimable asymptotic variance are replaced with exact finite-sample normality with known covariance matrix.

I consider a general shrinkage estimator and URE constructed from $\hat{\theta}$ by analogy with Sections 3.2.2 and 3.2.3. Define the shrinkage estimator

$$\hat{\theta}_{M,W}(\lambda) = \Theta_{M,W}(\lambda)\hat{\theta}, \quad \lambda \geq 0,$$

where $\Theta_{M,W}(\lambda)$ is given by (3.4), and also

$$\hat{\lambda}_{M,W,\tilde{W}} = \underset{\lambda \geq 0}{\arg\min} \left(\|\hat{\theta}_{M,W}(\lambda) - \hat{\theta}\|_{\tilde{W}}^2 + 2\,\mathrm{tr}\{\tilde{W}\Theta_{M,W}(\lambda)\Sigma\} \right). \tag{3.12}$$

Denote the minimum-URE shrinkage estimator by $\hat{\theta}_{M,W,\tilde{W}} = \hat{\theta}_{M,W}(\hat{\lambda}_{M,W,\tilde{W}})$.[15] If $W = \tilde{W} = I_n$ and $M = P$ is an orthogonal projection matrix, we have

$$\hat{\theta}_{P,I_n,I_n} = \hat{\theta} - \min\left\{ \frac{\mathrm{tr}(P\Sigma)}{\|P\hat{\theta}\|^2}, 1 \right\} P\hat{\theta}, \tag{3.13}$$

[15] If the minimum in (3.12) is attained at $\lambda = \infty$, $\hat{\theta}_{M,W,\tilde{W}}$ is defined as a limit, cf. Footnote 11.

which is the analogue of the optimal projection shrinkage estimator (3.8).

In the following I invert tests based on the shrinkage estimator $\hat{\theta}_{M,W,\tilde{W}}$ to construct finite-sample valid confidence sets for $\theta^\dagger$ or for a linear combination $s'\theta^\dagger$. When constructing joint confidence sets for the vector $\theta^\dagger$, a variety of shrinkage confidence sets have been shown to dominate the usual Wald ellipse centered at $\hat{\theta}$, for the special case $\Sigma = I_n$ (Casella & Hwang, 2012). Simulations in Section 3.5 suggest that shrinkage confidence sets are competitive when based on general shrinkage estimators and with non-diagonal Σ, although I have not proved analytic dominance results. Marginal shrinkage confidence sets for the scalar $s'\theta^\dagger$ cannot have uniformly smaller expected length than the usual Wald interval centered at $s'\hat{\theta}$, as the latter is uniquely minimax up to Lebesgue null sets (Joshi, 1969). Nevertheless, the simulations in Section 3.5 indicate that marginal shrinkage confidence sets often outperform the usual confidence interval when the true IRF is smooth without much worse expected length in the non-smooth case.[16]

Joint confidence sets. To construct a joint confidence set for $\theta^\dagger$, I invert the test statistic $g(\hat{\theta}_{M,W,\tilde{W}} - \theta, \Sigma)$, where $g\colon \mathbb{R}^n \times \mathbb{S}^n_+ \to \mathbb{R}_+$ is a continuous function. For example, the choice $g(\theta, \Sigma) = \|\operatorname{diag}(\Sigma)^{-1/2}\theta\|_\infty$, where $\operatorname{diag}(\Sigma)$ equals Σ with all non-diagonal elements set to zero, yields the confidence set presented in Section 3.2.4. Let $q_{1-\alpha,M,W,\tilde{W}}(\theta, \Sigma)$ be the $1-\alpha$ quantile of $g(\hat{\theta}_{M,W,\tilde{W}}(\theta + U, \Sigma) - \theta, \Sigma)$, $U \sim N(0, \Sigma)$, cf. definition (3.9) (this coincides with the definition in Section 3.2.4 for the choice of $g(\cdot, \cdot)$ used there). Then, by definition,

$$\hat{\mathcal{C}}^\theta_{1-\alpha} = \{\theta \in \mathbb{R}^n \colon g(\hat{\theta}_{M,W,\tilde{W}} - \theta, \Sigma) \leq q_{1-\alpha,M,W,\tilde{W}}(\theta, \Sigma)\}$$

[16] In the model considered here, the Wald interval has uniformly shortest expected length among unbiased confidence sets (Lehmann & Romano, 2005, Ch. 5.5). The reason the marginal shrinkage sets can achieve smaller expected length than the Wald interval for some parameter values is that the shrinkage sets, while similar, do not attain their highest coverage rate at $s'\theta = s'\theta^\dagger$. I thank Adam McCloskey for a discussion.

is a confidence set for $\theta^\dagger$ with $1-\alpha$ coverage probability.

The proposed shrinkage confidence set has several attractive features, although I have not proved any formal optimality properties. First, its construction is based directly on the general shrinkage estimator, which is a desirable and economically intuitive estimator from the perspective of point estimation, as argued in Section 3.2. Second, the set is guaranteed to contain the shrinkage estimator $\hat{\theta}_{M,W,\tilde{W}}$, unlike, say, the usual Wald ellipse centered at $\hat{\theta}$. Third, one can show that in the projection shrinkage case (3.13), the set $\hat{\mathcal{C}}$ coincides with the Wald ellipse almost surely in the limit under any sequence of probability measures with $\|P\theta\| \to \infty$.[17] A drawback of the shrinkage confidence set is that it requires extensive simulation to compute numerically, as described in Section 3.2.4.

Marginal confidence sets. To construct computationally cheap confidence sets for the linear combination $s'\theta^\dagger$, I argue that conditional inference is appropriate. Define $\tilde{P} = \zeta s'$, where $\zeta = (s'\Sigma s)^{-1}\Sigma s$. Since $I_n - \tilde{P}$ is idempotent with rank $n-1$, there exist full-rank matrices $A, B \in \mathbb{R}^{n\times(n-1)}$ such that $I_n - \tilde{P} = AB'$. The map $\psi\colon \mathbb{R}^n \to \mathbb{R}\times\mathbb{R}^{n-1}$ given by $\psi(\theta) = (s'\theta, B'\theta)$ is then a reparametrization of the mean parameter vector,[18] and $\hat{\nu}^\theta = (I_n - \tilde{P})\hat{\theta}$ is an S-ancillary statistic for $s'\theta$: The distribution of $\hat{\nu}^\theta$ depends on the unknown parameters $(s'\theta, B'\theta)$ only through $B'\theta$, whereas the conditional distribution of $\hat{\theta}$ given $\hat{\nu}^\theta$ does not depend on $B'\theta$ (Lehmann & Romano, 2005, p. 398). These considerations suggest that one should condition on $\hat{\nu}^\theta$ when doing inference about $s'\theta$.

I obtain a confidence set for $\mu^\dagger = s'\theta^\dagger$ by inverting the statistic $(s'\hat{\theta}_{M,W,\tilde{W}} - \mu)^2$, conditioning on $\hat{\nu}^\theta = (I_n - \tilde{P})\hat{\theta}$. Let $q_{s,1-\alpha,M,W,\tilde{W}}(\theta,\Sigma)$ be the quantile function defined in Section 3.2.4. Since the jointly Gaussian random variables $s'\hat{\theta}$ and $\hat{\nu}^\theta$ are orthogonal and

[17] The argument is similar to the proof of Proposition 3.3 below.

[18] This follows from the matrix $(s, B)'$ being non-singular: The kernel of B' is $\text{span}(\zeta)$, but $s'\zeta = 1$.

thus independent, the distribution of $\hat{\theta} = \zeta(s'\hat{\theta}) + (I_n - \tilde{P})\hat{\theta}$ conditional on $\hat{\nu}^\theta = \nu$ equals the distribution of $\zeta(u + \mu^\dagger) + \nu$, where $u \sim N(0, s'\Sigma s)$. Hence,

$$\hat{\mathcal{C}}^\theta_{s,1-\alpha} = \{\mu \in \mathbb{R} \colon (s'\hat{\theta}_{M,W,\tilde{W}} - \mu)^2 \leq q_{s,1-\alpha,M,W,\tilde{W}}(\zeta\mu + \hat{\nu}^\theta, \Sigma)\}$$

is a confidence set for $\mu^\dagger = s'\theta^\dagger$ with conditional coverage $1-\alpha$, and thus also valid unconditional coverage: For $\nu \in \text{span}(I_n - \tilde{P})$,

$$\begin{aligned}
&\text{Prob}(\mu^\dagger \in \hat{\mathcal{C}}^\theta_{s,1-\alpha} \mid \hat{\nu}^\theta = \nu) \\
&\quad = \text{Prob}\left(\{s'\hat{\theta}_{M,W,\tilde{W}}(\zeta u + \zeta\mu^\dagger + \nu, \Sigma) - s'(\zeta\mu^\dagger + \nu)\}^2 \leq q_{s,1-\alpha,M,W,\tilde{W}}(\zeta\mu^\dagger + \nu, \Sigma)\right) \\
&\quad = 1 - \alpha.
\end{aligned}$$

The first equality uses independence of $s'\hat{\theta}$ and $\hat{\nu}^\theta$, the definition (3.9) of $\hat{\theta}_{M,W,\tilde{W}}(\eta, \Sigma)$, $s'\zeta = 1$, and $s'\nu = 0$. The second equality uses the definition of $q_{s,1-\alpha,M,W,\tilde{W}}(\cdot,\cdot)$.

An alternative confidence set is obtained by intersecting a Wald interval with the above confidence set, using adjusted critical values. Let $\delta \in [0, \alpha]$, and define the $1-\delta$ level Wald confidence interval $\hat{\mathcal{I}}^\theta_{s,1-\delta} = [s'\hat{\theta} - c_{1-\delta}, s'\hat{\theta} + c_{1-\delta}]$, where $c_{1-\delta} = \sqrt{(s'\Sigma s) z_{1,1-\delta}}$. Let $\tilde{q}_{s,1-\alpha,M,W,\tilde{W}}(\theta, \Sigma, c)$ denote the quantile function defined in Section 3.2.4. Then

$$\hat{\mathcal{C}}^\theta_{s,1-\alpha,1-\delta} = \{\mu \in \hat{\mathcal{I}}^\theta_{s,1-\delta} \colon (s'\hat{\theta}_{M,W,\tilde{W}} - \mu)^2 \leq \tilde{q}_{s,\frac{1-\alpha}{1-\delta},M,W,\tilde{W}}(\zeta\mu + \hat{\nu}^\theta, \Sigma, c_{1-\delta})\}$$

is a $1-\alpha$ level conditional (and thus unconditional) confidence set: For $\nu \in \text{span}(I_n - \tilde{P})$,

$$\begin{aligned}
&\text{Prob}(\mu^\dagger \in \hat{\mathcal{C}}^\theta_{s,1-\alpha,1-\delta} \mid \hat{\nu}^\theta = \nu) \\
&\quad = \text{Prob}\left((s'\hat{\theta}_{M,W,\tilde{W}} - \mu^\dagger)^2 \leq \tilde{q}_{s,\frac{1-\alpha}{1-\delta},M,W,\tilde{W}}(\zeta\mu^\dagger + \nu, \Sigma, c_{1-\delta}) \,\middle|\, \mu^\dagger \in \hat{\mathcal{I}}^\theta_{s,1-\delta}, \hat{\nu}^\theta = \nu\right) \\
&\qquad \times \text{Prob}(\mu^\dagger \in \hat{\mathcal{I}}^\theta_{s,1-\delta})
\end{aligned}$$

$$= \text{Prob}\Big(\{ s'\hat{\theta}_{M,W,\tilde{W}}(\zeta u + \zeta\mu^\dagger + \nu, \Sigma) - s'(\zeta\mu^\dagger + \nu) \}^2 \\ \leq \tilde{q}_{s,\frac{1-\alpha}{1-\delta},M,W,\tilde{W}}(\zeta\mu^\dagger + \nu, \Sigma, c_{1-\delta}) \,\Big|\, |u| \leq c_{1-\delta} \Big)(1-\delta)$$
$$= \frac{1-\alpha}{1-\delta}(1-\delta).$$

The first equality uses independence of $s'\hat{\theta}$ and $\hat{\nu}^\theta$. The second equality sets $u = s'\hat{\theta} - \mu^\dagger \sim N(0, s'\Sigma s)$ and uses independence, the definition (3.9) of $\hat{\theta}_{M,W,\tilde{W}}(\eta, \Sigma)$, $s'\zeta = 1$, and $s'\nu = 0$. The last equality follows from the definition of $\tilde{q}_{s,1-\alpha,M,W,\tilde{W}}(\cdot,\cdot,\cdot)$.

The alternative confidence set for $\mu^\dagger$ has known worst-case length and is easy to compute, but it is not guaranteed to contain the shrinkage estimator. Since $\hat{\mathcal{C}}^\theta_{s,1-\alpha,1-\delta} \subset \hat{\mathcal{I}}^\theta_{s,1-\delta}$, the worst-case length of the set is $2c_{1-\delta}$. When numerically computing the set by grid search, the grid can of course be confined to $\hat{\mathcal{I}}^\theta_{s,1-\delta}$. However, the set has two drawbacks relative to the pure inversion-based set $\hat{\mathcal{C}}^\theta_{s,1-\alpha}$. First, $\hat{\mathcal{C}}^\theta_{s,1-\alpha,1-\delta}$ contains the shrinkage estimator $s'\hat{\theta}_{M,W,\tilde{W}}$ if and only if the latter is contained in the Wald interval $\hat{\mathcal{I}}^\theta_{s,1-\delta}$.[19] Second, the construction of $\hat{\mathcal{C}}^\theta_{s,1-\alpha,1-\delta}$ hinges on the tuning parameter δ, as discussed in Section 3.2.4.

3.4.2 Asymptotic uniform coverage

I now show that, when applied to a class of projection shrinkage estimators, the shrinkage confidence sets introduced in the previous subsection achieve asymptotic uniform coverage under weak assumptions on the initial IRF estimator. In place of the idealized assumptions in the finite-sample normal model from the previous subsection, I assume that the initial IRF estimator is uniformly asymptotically normal with a uniformly consistently estimable asymptotic variance. This effectively strengthens Assumption 3.1 in Section 3.3.

Assumption 3.2. *Define* $\mathcal{S} = \{A \in \mathbb{S}^n_+ : \underline{c} \leq 1/\rho(A^{-1}) \leq \rho(A) \leq \bar{c}\}$ *for some fixed* $\underline{c}, \bar{c} > 0$.

[19] In the case of projection shrinkage (3.13) with $W = \tilde{W} = I_n$ and $M = P$, $\hat{\mathcal{C}}^\theta_{s,1-\alpha,1-\delta}$ always contains $s'\hat{\theta}_{P,I_n,I_n}$ if $c_{1-\delta} \geq \|s\|\sqrt{\text{tr}(P\Sigma)}$. This follows from $\|\hat{\theta}_{P,I_n,I_n} - \hat{\theta}\| \leq \sqrt{\text{tr}(P\Sigma)}$, cf. Appendix A.3.3.

The distribution of the data F_T for sample size T is indexed by three parameters $\beta \in \mathbb{R}^n$, $\Sigma \in \mathcal{S}$, and $\gamma \in \Gamma$, where Γ is some set. The estimators $(\hat{\beta}, \hat{\Sigma}) \in \mathbb{R}^n \times \mathbb{S}^n$ satisfy the following: For all subsequences $\{k_T\}_{T\geq 1}$ of $\{T\}_{T\geq 1}$ and all sequences $\{\beta_{k_T}, \Sigma_{k_T}, \gamma_{k_T}\}_{T\geq 1} \in \mathbb{R}^n \times \mathcal{S} \times \Gamma$, we have, as $T \to \infty$,

$$\sqrt{k_T}\hat{\Sigma}^{-1/2}(\hat{\beta} - \beta_{k_T}) \xrightarrow[F_{k_T}(\beta_{k_T}, \Sigma_{k_T}, \gamma_{k_T})]{d} N(0, I_n), \quad (\hat{\Sigma} - \Sigma_{k_T}) \xrightarrow[F_{k_T}(\beta_{k_T}, \Sigma_{k_T}, \gamma_{k_T})]{p} 0.$$

The assumption requires that $\hat{\beta}$ is asymptotically normal and $\hat{\Sigma}$ consistent under drifting sequences of parameters. This is a type of asymptotic regularity condition on the estimators. While the parameter space for the IRF β is unrestricted, the parameter space for the asymptotic variance Σ of $\hat{\beta}$ is restricted to a compact subset of the space of positive definite matrices, thus assuming away near-singular cases. The parameter γ in the assumption captures all aspects of the distribution of the data that are not controlled by the parameters β and Σ. If $\hat{\beta}$ is obtained from a time series or panel regression, Assumption 3.2 will typically be satisfied under mild assumptions on the moments of the regressors and residuals. Finite-sample normality is not required.

I consider a class of estimators that contains the optimal projection shrinkage estimator. Given the initial IRF estimator $\hat{\beta}$ an $n \times n$ orthogonal projection matrix P and a function $f(\cdot, \cdot)$ satisfying Assumption 3.3 below, define the shrinkage estimator

$$\hat{\beta}_P = \hat{\beta} - f(T\|P\hat{\beta}\|^2, \hat{\Sigma})P\hat{\beta}.$$

Assumption 3.3. *$f\colon \mathbb{R}_+ \times \mathbb{S}^n_+ \to \mathbb{R}$ is continuous, and $\lim_{x\to\infty} xf(x^2, \Sigma) \to 0$ for all $\Sigma \in \mathbb{S}^n_+$.*

The choice $f(x, \Sigma) = \min\{\mathrm{tr}(P\Sigma)/x, 1\}$ satisfies Assumption 3.3 and yields the optimal projection shrinkage estimator (3.8).

JOINT CONFIDENCE SETS. The next result states that the joint confidence set in Section 3.4.1 is asymptotically uniformly valid. Let $g\colon \mathbb{R}^n \times \mathbb{S}^n_+ \to \mathbb{R}_+$, and define the test statistic

$$\hat{S}(\beta) = g\left(\sqrt{T}(\hat{\beta}_P - \beta), \hat{\Sigma}\right).$$

Let the quantile function $q_{1-\alpha}(\theta, \Sigma)$ be defined as $q_{1-\alpha,P,I_n,I_n}(\theta, \Sigma)$ in Section 3.2.4, except that $\hat{\theta}_{P,I_n,I_n}(\eta, \Sigma)$ is substituted with $\hat{\theta}_P(\eta, \Sigma) = \eta - f(\|P\eta\|^2, \Sigma)P\eta$.

Proposition 3.3. *Let Assumptions 3.2 and 3.3 hold, and assume that $g(\cdot, \cdot)$ is continuous. Then*

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\left(\hat{S}(\beta) \leq q_{1-\alpha}(\sqrt{T}\beta, \hat{\Sigma})\right) = 1 - \alpha. \tag{3.14}$$

Thus, for sufficiently large sample sizes, the worst-case *finite-sample* coverage probability of the confidence region $\{\beta \in \mathbb{R}^n\colon \hat{S}(\beta) \leq q_{1-\alpha}(\sqrt{T}\beta, \hat{\Sigma})\}$ does not fall below $1 - \alpha$. The proof uses the drifting parameter techniques of Andrews et al. (2011) and McCloskey (2015).

MARGINAL CONFIDENCE SETS. Similarly to the joint case, I prove that the marginal confidence sets constructed in Section 3.4.2 are asymptotically uniformly valid. Suppose we wish to conduct inference on the linear combination $s'\beta$ of the true IRF β, where $s \in \mathbb{R}^n\backslash\{0\}$. Define $\hat{\zeta} = (s'\hat{\Sigma}s)^{-1}\hat{\Sigma}s$, $\hat{P} = \hat{\zeta}s'$, and $\hat{\nu} = (I_n - \hat{P})\hat{\beta}$. Define the test statistics

$$\hat{S}_{s,W}(\mu) = T(s'\hat{\beta} - \mu)^2, \quad \hat{S}_s(\mu) = T(s'\hat{\beta}_P - \mu)^2, \quad \mu \in \mathbb{R}.$$

Let the quantile functions $q_{s,1-\alpha}(\beta, \Sigma)$ and $\tilde{q}_{s,1-\alpha}(\beta, \Sigma, c)$ be defined as $q_{s,1-\alpha,P,I_n,I_n}(\theta, \Sigma)$ and $\tilde{q}_{s,1-\alpha,P,I_n,I_n}(\beta, \Sigma, c)$, respectively, in Section 3.2.4, except that $\hat{\theta}_{P,I_n,I_n}(\eta, \Sigma)$ is substituted with $\hat{\theta}_P(\eta, \Sigma) = \eta - f(\|P\eta\|^2, \Sigma)P\eta$. Finally, define $\hat{c}_{1-\delta} = \sqrt{(s'\hat{\Sigma}s)z_{1,1-\delta}}$.

Proposition 3.4. *Let Assumptions 3.2 and 3.3 hold. Then*

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\left(\hat{S}_s(s'\beta) \leq q_{s,1-\alpha}\Big(\sqrt{T}(\hat{\zeta}(s'\beta)+\hat{\nu}),\hat{\Sigma}\Big)\right) = 1-\alpha. \quad (3.15)$$

Moreover, for all $\delta \in [0,\alpha]$,

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\Big(\hat{S}_{s,W}(s'\beta) \leq \hat{c}^2_{1-\delta},$$
$$\hat{S}_s(s'\beta) \leq \tilde{q}_{s,\frac{1-\alpha}{1-\delta}}\Big(\sqrt{T}(\hat{\zeta}(s'\beta)+\hat{\nu}),\hat{\Sigma},\hat{c}_{1-\delta}\Big)\Big) = 1-\alpha. \quad (3.16)$$

3.5 Simulation study

This section illustrates the properties of the shrinkage estimators and confidence sets by simulation. First, I consider the idealized setting of the finite-sample normal model with known covariance matrix from Section 3.4.1. The simulations show that shrinkage often delivers large gains for joint estimation and confidence set construction, while the marginal shrinkage confidence sets are competitive with the standard Wald confidence interval. Second, I consider a realistic time series regression setting with data generating process (DGP) calibrated to the empirical example from Section 3.2. I find that the advantages of shrinkage carry over to this setting, despite the need to estimate the asymptotic variance of the initial IRF estimator.

3.5.1 Normal model with known covariance matrix

To focus on essentials, I first consider the finite-sample normal location model with known covariance matrix from Section 3.4.1.

DGP AND ESTIMATORS. We observe a single normal draw $\hat{\theta} \sim N(\theta^\dagger, \Sigma)$ with mean $\theta^\dagger \in \mathbb{R}^n$ and known covariance matrix Σ. Given jaggedness parameter K, the true IRF $\theta^\dagger = (\theta_0^\dagger, \dots, \theta_{n-1}^\dagger)'$ is

$$\theta_i^\dagger = \begin{cases} 1 - \frac{i}{n-1} & \text{if } K = 0, \\ \sin \frac{2\pi K i}{n-1} & \text{if } K > 0, \end{cases} \qquad i = 0, 1, \dots, n-1.$$

Hence, the true IRF is linearly decreasing from 1 to 0 if $K = 0$, while it is shaped like K full waves of a sine curve over $[0, n-1]$ when $K > 0$. Σ has the exponentially decreasing structure $\mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_k) = \sigma_i \sigma_k \kappa^{|i-k|}$, where $\sigma_i = \sigma_0(1 + i\frac{\varphi - 1}{n-1})$, so that $\varphi^2 = \mathrm{Var}(\hat{\theta}_{n-1})/\mathrm{Var}(\hat{\theta}_0)$. I consider different values for the parameters n, K, $\kappa \in [0, 1]$, $\sigma_0 > 0$, and $\varphi > 0$.

I investigate the performance of the SmIRF estimator and the optimal projection shrinkage estimator that shrinks towards a quadratic IRF. The quadratic projection shrinkage estimator is given by (3.13), where P is the projection matrix that shrinks towards a second-degree polynomial, cf. Section 3.2.2. Appendix A.3.5 contains simulation results for the SmIRF estimator. These are qualitatively similar to the projection shrinkage results.

RESULTS. Table 3.1 shows that the quadratic projection shrinkage estimator outperforms the initial IRF estimator in terms of total MSE for all DGPs considered. This performance improvement is due to the reduction in variance caused by shrinkage. The bias of the shrinkage estimator is only substantial in the DGPs with $K = 1$ or 2, i.e., for very non-quadratic IRFs.[20] Figure 3.8 illustrates in more detail how the relative MSE and squared bias of the projection shrinkage estimator depends on the jaggedness of the IRF (i.e., the parameter K). The relative MSE of the shrinkage estimator is small for IRFs that are well-approximated by a quadratic function (roughly, $K \leq 0.75$), and is still below 1 for very jagged IRFs due to the adaptivity afforded by a data-dependent smoothing parameter.

[20] I define bias as $\|E(\hat{\theta}) - \theta^\dagger\|$ and variance as $E(\|\hat{\theta} - E(\hat{\theta})\|^2)$, and similarly for the shrinkage estimator.

Simulation results: Projection shrinkage, normal model

					Joint			Marginal					
Parameters					MSE	Var	CV	MSE		Lng $\hat{\mathcal{C}}_s$		Lng $\hat{\mathcal{C}}_{s,1-\delta}$	
n	K	κ	σ_0	φ				Imp	Mid	Imp	Mid	Imp	Mid
10	0.5	0.5	0.2	3	0.65	0.65	0.80	1.31	0.55	1.08	0.86	1.08	0.86
25	0.5	0.5	0.2	3	0.35	0.35	0.61	1.19	0.29	0.94	0.82	0.93	0.82
50	0.5	0.5	0.2	3	0.20	0.20	0.44	0.82	0.16	0.87	0.81	0.85	0.80
25	0	0.5	0.2	3	0.35	0.35	0.64	1.20	0.29	0.93	0.82	0.93	0.82
25	1	0.5	0.2	3	0.84	0.66	1.15	3.60	0.56	1.49	0.85	1.40	0.85
25	2	0.5	0.2	3	0.90	0.77	0.97	1.16	0.70	1.04	0.87	1.03	0.87
25	0.5	0	0.2	3	0.16	0.16	0.41	0.58	0.13	0.84	0.81	0.84	0.80
25	0.5	0.9	0.2	3	0.82	0.82	0.85	1.57	0.78	1.19	0.90	1.19	0.90
25	0.5	0.5	0.1	3	0.36	0.35	0.60	1.28	0.31	0.96	0.82	0.95	0.82
25	0.5	0.5	0.4	3	0.35	0.35	0.65	1.15	0.30	0.94	0.82	0.93	0.82
25	0.5	0.5	0.2	1	0.34	0.34	0.57	0.70	0.29	0.88	0.82	0.88	0.82
25	0.5	0.5	0.2	5	0.35	0.35	0.71	1.88	0.28	1.07	0.82	1.00	0.82

Table 3.1: Simulation results for quadratic projection shrinkage, finite-sample normal model. Columns 1–5: DGP parameters. Column 6: Joint MSE of shrinkage estimator relative to joint MSE of $\hat{\theta}$. Column 7: Joint variance of shrinkage estimator relative to joint variance of $\hat{\theta}$. Column 8: Critical value at $\theta^\dagger$ for 90% joint shrinkage confidence set, relative to critical value of joint Wald set. Columns 9–10: Marginal MSE of SmIRF relative to $\hat{\theta}$ at horizons $i = 0$ ("Imp") and $i = 1 + [n/2]$ ("Mid"). Columns 11–14: Average length of 90% marginal shrinkage sets relative to Wald interval for sets $\hat{\mathcal{C}}_s = \hat{\mathcal{C}}_{s,1-\alpha}$ and $\hat{\mathcal{C}}_{s,1-\delta} = \hat{\mathcal{C}}_{s,1-\alpha,1-\delta}$. Length is defined as number of grid points in set, divided by total grid points (50), times length of grid. 5000 simulations per DGP, 1000 simulations to compute quantiles, $\alpha = 0.1$, $\delta = 0.01$.

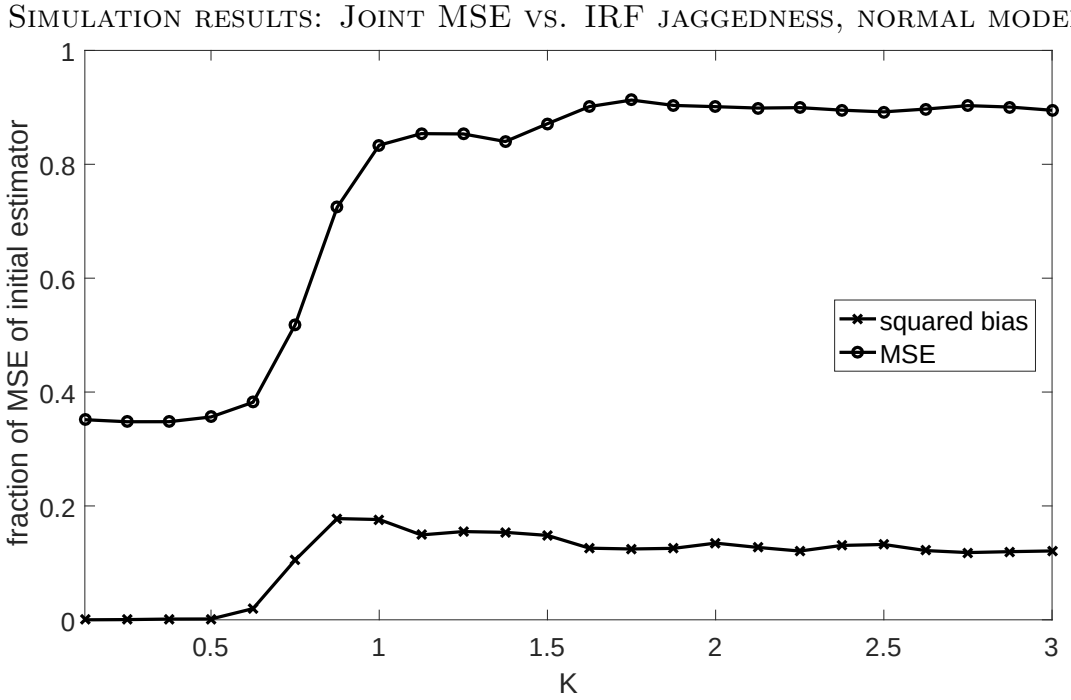

Figure 3.8: Joint MSE and squared bias of quadratic projection shrinkage estimator as functions of the jaggedness of the true IRF, finite-sample normal model. Horizontal axis shows $K = 0.125, 0.25, \ldots, 3$, with true IRF given by $\theta_i^\dagger = \sin \frac{2\pi K i}{n-1}$. Vertical axis units normalized by joint MSE of $\hat{\theta}$, for each K. Other DGP parameters: $n = 25$, $\kappa = 0.5$, $\sigma_0 = 0.2$, $\varphi = 3$.

Although shrinkage is not designed to improve individual impulse response estimates, Table 3.1 shows that shrinkage often produces more accurate estimates of a single longer-horizon impulse response. Intuitively, except at short horizons, the shrinkage estimator smooths between several nearby horizons, thus reducing variance and improving MSE if the true IRF is smooth.[21] At short horizons, such as the impact response $i = 0$, the projection shrinkage estimator tends to perform worse than the initial estimator because the number of nearby horizons is smaller, so the variance reduction is too small to outweigh the increase in bias. Nevertheless, the *absolute* MSE of the shrinkage estimator will often be small even for the impact response, as this response is typically estimated with relatively high precision.

Table 3.1 shows that joint and marginal shrinkage confidence sets are competitive with

[21]However, being admissible, $\hat{\theta}_i$ outperforms shrinkage for some non-smooth parametrizations.

the usual Wald confidence sets.[22] As an imperfect measure of the relative volume of the joint shrinkage confidence set, the table lists the critical value $q_{1-\alpha,P,I_n,I_n}(\theta^\dagger, \Sigma)$ evaluated at the true IRF, divided by the corresponding critical value $q_{1-\alpha,0,I_n,I_n}(\theta^\dagger, \Sigma)$ (which does not depend on $\theta^\dagger$) of the usual joint Wald set, cf. Section 3.2.4.[23] The relative critical value is below 1 for every DGP considered, often substantially. The marginal shrinkage sets have average length that is competitive with the usual Wald interval, often outperforming the latter at the middle response horizon. Although the Wald interval at the impact horizon $i = 0$ tends to be shorter, the average lengths of the marginal shrinkage sets are small in *absolute* terms at this horizon. There is very little difference between the average lengths of the pure inversion shrinkage set $\hat{\mathcal{C}}_{s,1-\alpha}$ and the alternative shrinkage set $\hat{\mathcal{C}}_{s,1-\alpha,1-\delta}$.

Table 3.1 offers the following additional lessons on the influence of the DGP parameters:

- The larger the number of parameters n, the better the relative performance of shrinkage procedures. This does not just apply to joint procedures, but even to marginal procedures, as higher n means more scope for smoothing between response horizons.

- The performance of shrinkage procedures worsens with higher κ, i.e., as the correlation between estimators at different horizons increases. However, even for $\kappa = 0.9$, the joint shrinkage procedures outperform the usual procedures, and the marginal shrinkage procedures outperform the usual procedures at the middle response horizon.

- Proportionally scaling the initial estimator variance Σ does not affect the relative performance of the various procedures.

---

[22]The table does not list coverage rates for the various joint or marginal confidence sets, as these are guaranteed to be exactly $1 - \alpha = 0.9$ up to simulation error.

[23]Unlike for the joint Wald set, the volume of the shrinkage set is not a function of only $q_{1-\alpha,P,I_n,I_n}(\theta^\dagger, \Sigma)$. However, the critical value evaluated at the true IRF is a good guide to the volume if the quantile function is not very sensitive to $\theta^\dagger$.

- Increasing the variance of the long-horizon initial impulse response estimators relative to the short-horizon estimators worsens the performance of shrinkage procedures, but mostly with regard to inference on the impact response.

In unreported simulations, I investigated the performance of a naive confidence set that is centered at the projection shrinkage estimator but uses the critical value for the corresponding Wald set. Casella & Hwang (1987) show that the naive set has valid coverage in the case $\kappa = 0$ and $\varphi = 1$, but its properties are otherwise unknown. My simulations show that the naive set does not control coverage for general choices of Σ, such as the above DGPs.

3.5.2 Time series regression

Now I consider a realistic setting in which we seek to estimate an IRF from time series data, without knowledge of underlying model parameters. The DGP is calibrated to the empirical example in Section 3.2.

DGP AND ESTIMATORS. I generate data from a VAR calibrated to the Gertler & Karadi (2015) data described in Appendix A.3.2. The simulation DGP is given by

$$\begin{pmatrix} y_t \\ w_t \end{pmatrix} = \sum_{k=1}^{2} A_k \begin{pmatrix} y_{t-k} \\ w_{t-k} \end{pmatrix} + bx_t + C\epsilon_t,$$

where $\dim(y_t) = 1$, $\dim(w_t) = 3$, $\epsilon_t \overset{\text{i.i.d.}}{\sim} N(0, I_4)$, and x_t is white noise independent of ϵ_t (its distribution is specified below). The parameter of interest is the IRF of y_t to x_t.

For calibration purposes, I use quadratically detrended monthly data for 1992–2012 on $y_t =$ the excess bond premium and $w_t =$ (log industrial production, log consumer price index, 1-year Treasury rate)$'$.[24] Let $x_t =$ monetary policy shock. The coefficients $A_1, A_2 \in \mathbb{R}^{4\times 4}$

[24]The Bayesian Information Criterion selects a VAR lag length of 2 for the $(y_t, w_t')'$ data.

and $b \in \mathbb{R}^4$ are obtained by least-squares regression on the calibration data, with $C \in \mathbb{R}^{4\times 4}$ given by the Cholesky factor of the covariance matrix of the regression residuals. The true IRF of y_t to x_t implied by the calibrated VAR is plotted in Appendix A.3.5.

I consider several parametrizations based on the above VAR. First, I let n, the number of impulse responses of interest, be either 25 (two years) or 49 (four years). Second, I consider different sample sizes T of simulated data. Third, I let the shock x_t be either i.i.d. standard normal or i.i.d. t-distributed with 5 degrees of freedom (i.e., finite fourth, but not fifth, moment), normalized to have variance 1. Fourth, I consider different HAC lag lengths ℓ when estimating Σ, the asymptotic variance of the local projection estimator.

I investigate the relative performance of the Jordà (2005) local projection estimator (3.1) and the quadratic projection shrinkage transformation of this estimator (see the definition in the previous subsection). The local projection estimator regresses current and future values of y_t on x_t, controlling for w_{t-1}, w_{t-2}. I do not compare with VAR estimators or associated confidence sets as this has already been done by Jordà (2005).

Results. Table 3.2 shows that the main results from the idealized normal model carry over to the realistic time series setting. Despite the need to estimate the asymptotic variance Σ of the local projection estimator by HAC methods, the shrinkage procedures outperform the local projection procedures in the case of joint inference and marginal inference on the middle response horizon. Only in the case of inference on the impact impulse response do the local projection procedures deliver smaller MSE and shorter confidence intervals.

The coverage rates of the joint and marginal shrinkage confidence sets are satisfactory for moderate and large sample sizes. The shrinkage sets have coverage closer to the nominal 90% level than the local projection based confidence sets for all DGPs, with substantial improvements in the case of joint inference. The marginal shrinkage sets have near-correct coverage in all cases except for sample size $T = 100$. The coverage rates do not deteriorate

Simulation results: Projection shrinkage, time series regression

Parameters				Joint				Marginal							
				MSE	CV	Cov	Cov	MSE		Lng $\hat{\mathcal{C}}_{s,1-\delta}$		Cov W		Cov $\hat{\mathcal{C}}_{s,1-\delta}$	
n	T	SD	ℓ			W	$\hat{\mathcal{C}}$	Imp	Mid	Imp	Mid	Imp	Mid	Imp	Mid
25	200	N	12	0.74	0.85	0.79	0.83	1.51	0.71	1.08	0.91	0.86	0.84	0.87	0.85
49	200	N	12	0.59	0.85	0.72	0.83	2.68	0.51	1.12	0.89	0.86	0.82	0.87	0.85
25	100	N	12	0.70	0.85	0.62	0.74	1.32	0.69	1.05	0.93	0.81	0.77	0.83	0.79
25	300	N	12	0.75	0.85	0.84	0.86	1.64	0.72	1.10	0.91	0.88	0.87	0.89	0.87
25	500	N	12	0.77	0.86	0.87	0.88	1.92	0.74	1.14	0.90	0.88	0.87	0.89	0.88
25	200	t	12	0.74	0.85	0.77	0.82	1.48	0.70	1.08	0.92	0.85	0.84	0.87	0.85
25	200	N	6	0.74	0.83	0.84	0.87	1.57	0.72	1.08	0.90	0.87	0.87	0.88	0.87
25	200	N	24	0.75	0.88	0.71	0.80	1.50	0.73	1.09	0.93	0.83	0.81	0.85	0.84

Table 3.2: Simulation results for quadratic projection shrinkage, time series regression on VAR data. Columns 5–6, 9–12: See caption for Table 3.1. Columns 1–4: DGP parameters ("SD" = shock distribution, either Normal or t). Columns 7–8: Coverate rates for 90% joint Wald and shrinkage confidence sets. Columns 13–16: Coverage rates for 90% marginal Wald and shrinkage confidence sets at horizons $i = 0$ ("Imp") and $i = 1 + [n/2]$ ("Mid"). 5000 simulations per DGP, 1000 simulations to compute quantiles, 100 period burn-in, $\alpha = 0.1$, $\delta = 0.01$. HAC: Newey-West.

markedly when the shock x_t is t-distributed. The last rows of Table 3.2 indicate that coverage rates for joint sets are sensitive to the choice of HAC bandwidth parameter ℓ, but optimal HAC estimation is outside the scope of this paper.

3.6 Topics for future research

I finish by discussing several possible topics for future research.

The results on MSE of shrinkage estimators with URE-minimizing shrinkage parameter and on uniform coverage only apply to projection shrinkage. I conjecture that uniform coverage can be proved in a similar manner for general shrinkage confidence sets. Moreover, I conjecture that an MSE-dominance result for general shrinkage estimators can be proved under asymptotics in which the number of response horizons n tends to infinity, following results for i.i.d. regression in Li (1986), Andrews (1991), and Xie, Kou & Brown (2012).

I have not provided analytic conditions for the test-inversion shrinkage confidence sets

to outperform the usual Wald confidence sets in terms of expected volume or other loss. Simulation evidence suggests that dominance is not uniform over the parameter space for general dependence structures, but the shrinkage sets often offer large gains when the true IRF is not too jagged. It would aid our understanding of the relative performance if the shrinkage sets can be shown to have a precise Bayes or Empirical Bayes interpretation.

In unreported simulation experiments I have found that shrinkage confidence sets based on the "Simple Bonferroni" procedure of McCloskey (2015) have high expected volume relative to usual Wald sets precisely when the true IRF is smooth. It would be interesting to investigate whether McCloskey's more computationally demanding "Adjusted Bonferroni" procedure substantially reduces expected volume.

I have not been able to prove that the test inversion shrinkage confidence sets are convex. Simulations suggest that the marginal confidence sets are convex (i.e., intervals). More evidence is needed on the geometry of the joint confidence sets.

The performance of the shrinkage estimators and confidence sets depend on the quality of the HAC or clustered standard errors. While simulations suggest that the need to perform HAC estimation does not compromise the performance of shrinkage procedures *relative* to the initial IRF estimator, methods for improving the quality of HAC or clustered standard errors would presumably help also in this context (Müller, 2014; Imbens & Kolesár, 2016).

Appendix A

Supplemental Material

A.1 Material for Chapter 1

A.1.1 Notation

I_n is the $n \times n$ identity matrix. $\imath$ is the imaginary unit that satisfies $\imath^2 = -1$. If a is a vector, $\mathrm{diag}(a)$ denotes the diagonal matrix with the elements of a along the diagonal in order. If A is a square matrix, $\mathrm{tr}(A)$ is its trace, $\det(A)$ is its determinant, and $\mathrm{diag}(A)$ is the vector consisting of the diagonal elements in order. For an arbitrary matrix B, B' denotes the matrix transpose, $\bar{B}$ denotes the elementwise complex conjugate, $B^* = \bar{B}'$ is the complex conjugate transpose, $\mathrm{Re}(B)$ is the real part of B, $\|B\| = \sqrt{\mathrm{tr}(B^*B)}$ is the Frobenius norm, and $\mathrm{vec}(B)$ is the columnwise vectorization. If C is a positive semidefinite matrix, $\lambda_{\min}(C)$ is its smallest eigenvalues. If Q is an $n \times n$ matrix, it is said to be orthogonal if it is real and $QQ' = I_n$, while it is said to be unitary if $QQ^* = I_n$. The statement $X \perp\!\!\!\perp Y \mid Z$ means that the random variables X and Y are independent conditional on Z. If $\mathcal{K}$ is a set, $\overline{\mathcal{K}}$ denotes its closure and $\mathcal{K}^c$ denotes its complement.

A.1.2 Constructive characterization of the identified set

The result below applies the analysis of Lippi & Reichlin (1994) to the SVMA model; see also Hansen & Sargent (1981) and Komunjer & Ng (2011). I identify a set of IRFs $\Theta = (\Theta_0, \ldots, \Theta_q)$ with the matrix polynomial $\Theta(z) = \sum_{\ell=0}^{q} \Theta_\ell z^\ell$, and I use the notation Θ and $\Theta(z)$ interchangeably where appropriate. In words, the theorem says that if we start with some set of IRFs $\Theta(z)$ contained in the identified set, then we can obtain all other sets of IRFs in the identified set by applying orthogonal rotations to $\Theta(z)$ and/or by "flipping the roots" of $\Theta(z)$. Only a finite sequence of such operations is necessary to jump from one element of the identified set to any other element of the identified set.

Theorem A.1. *Let $\{\Gamma(k)\}_{0\leq k\leq q}$ be an arbitrary ACF. Pick an arbitrary $(\Theta, \sigma) \in \mathcal{S}(\Gamma)$ satisfying $\det(\Theta(0)) \neq 0$. Define $\Psi(z) = \Theta(z)\,\mathrm{diag}(\sigma)$.*

Construct a matrix polynomial $\check{\Psi}(z)$ in either of the following two ways:

(i) Set $\check{\Psi}(z) = \Psi(z)Q$, where Q is an arbitrary orthogonal $n \times n$ matrix.

(ii) Let $\gamma_1, \ldots, \gamma_r$ $(r \leq nq)$ denote the roots of the polynomial $\det(\Psi(z))$. Pick an arbitrary positive integer $k \leq r$. Let $\eta \in \mathbb{C}^n$ be a vector such that $\Psi(\gamma_k)\eta = 0$ (such a vector exists because $\det(\Psi(\gamma_k)) = 0$). Let Q be a unitary matrix whose first column is proportional to η (if γ_k is real, choose Q to be a real orthogonal matrix). All elements of the first column of the matrix polynomial $\Psi(z)Q$ then contain the factor $(z - \gamma_k)$. In each element of the first column, replace the factor $(z-\gamma_k)$ with $(1-\overline{\gamma_k}z)$. Call the resulting matrix polynomial $\check{\Psi}(z)$. If γ_k is real, stop.

If γ_k is not real, let $\tilde{\eta} \in \mathbb{C}^n$ be a vector such that $\check{\Psi}(\overline{\gamma_k})\tilde{\eta} = 0$, and let $\tilde{Q}$ be a unitary matrix whose first column is proportional to $\tilde{\eta}$. All elements of the first column of $\check{\Psi}(z)\tilde{Q}$ then contain the factor $(z-\overline{\gamma_k})$. In each element of the first column, replace the factor $(z - \overline{\gamma_k})$ with $(1 - \gamma_k z)$. Call the resulting matrix polynomial $\tilde{\Psi}(z)$. The matrix

$\tilde{\Psi}(0)\tilde{\Psi}(0)^*$ *is real, symmetric, and positive definite, so pick a real* $n \times n$ *matrix* J *such that* $JJ' = \tilde{\Psi}(0)\tilde{\Psi}(0)^*$. *In an abuse of notation, set* $\check{\Psi}(z) = \tilde{\Psi}(z)\tilde{\Psi}(0)^{-1}J$, *which is guaranteed to be a real matrix polynomial.*

Now obtain a set of IRFs $\check{\Theta}$ *and shock standard deviations* $\check{\sigma}$ *from* $\check{\Psi}(z)$*:*

(a) For each $j = 1, \ldots, n$, *if the* (i_j, j) *element of* $\check{\Psi}(0)$ *is negative, flip the signs of all elements in the* j*-th column of* $\check{\Psi}(z)$, *and call the resulting matrix polynomial* $\check{\check{\Psi}}(z)$. *For each* $j = 1, \ldots, n$, *let* $\check{\sigma}_j$ *denote the* (i_j, j) *element of* $\check{\check{\Psi}}(0)$. *Define* $\check{\sigma} = (\check{\sigma}_1, \ldots, \check{\sigma}_n)$ *and* $\check{\Theta}(z) = \check{\check{\Psi}}(z)\,\mathrm{diag}(\check{\sigma})^{-1}$ *(if the inverse exists).*

Then $(\check{\Theta}, \check{\sigma}) \in \mathcal{S}(\Gamma)$, *provided that all elements of* $\check{\sigma}$ *are strictly positive.*

On the other hand, if $(\check{\Theta}, \check{\sigma}) \in \mathcal{S}(\Gamma)$ *is an arbitrary point in the identified set satisfying* $\det(\check{\Theta}(0)) \neq 0$, *then* $(\check{\Theta}, \check{\sigma})$ *can be obtained from* (Θ, σ) *as follows:*

1. *Start with the initial point* (Θ, σ) *and the associated polynomial* $\Psi(z)$ *defined above.*

2. *Apply an appropriate* finite *sequence of the above-mentioned transformations (i) or (ii), in an appropriate order, to* $\Psi(z)$, *resulting ultimately in a polynomial* $\check{\Psi}(z)$.

3. *Apply the above-mentioned operation (a) to* $\check{\Psi}(z)$. *The result is* $(\check{\Theta}, \check{\sigma})$.

Remarks:

1. An initial point in the identified set can be obtained by following the procedure in Hannan (1970, pp. 64–66) and then applying transformation (a). This essentially corresponds to computing the Wold decomposition of $\{y_t\}$ and applying appropriate normalizations (Hannan, 1970, Thm. $2''$, p. 158). Hence, Theorem A.1 states that any set of structural IRFs that are consistent with a given ACF $\Gamma(\cdot)$ are obtained by applying transformations (i) and (ii) to the Wold IRFs corresponding to $\Gamma(\cdot)$.

2. Transformation (ii) corresponds to "flipping the root" γ_k of $\det(\Psi(z))$. If γ_k is not real, transformation (ii) requires that we also flip the complex conjugate root $\overline{\gamma_k}$, since this ensures that the resulting matrix polynomial will be real after a rotation. The rule used to compute the matrix J in transformation (ii) is not important for the theorem; in particular, J can be the Cholesky factor of $\tilde{\Psi}(0)\tilde{\Psi}(0)^*$.

3. The only purpose of transformation (a) is to enforce the normalizations $\Theta_{i_jj,0} = 1$.

4. To simplify the math, the theorem restricts attention to IRFs satisfying $\det(\Theta(0)) = \det(\Theta_0) \neq 0$. If $\det(\Theta_0) = 0$, there exists a linear combination of $y_{1,t}, \ldots, y_{n,t}$ that is perfectly predictable based on knowledge of shocks $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ occurring before time t. Hence, in most applications, a reasonable prior for Θ ought to assign zero probability to the event $\det(\Theta_0) = 0$.

5. If the IRF parameter space Ξ_Θ were restricted to those IRFs that are invertible (cf. Section 1.2.3), then transformation (ii) would be unnecessary. In this case, the identified set for $\Psi(z) = \Theta(z)\,\mathrm{diag}(\sigma)$ can be obtained by taking any element in the set (e.g., the Wold IRFs) and applying all possible orthogonal rotations, i.e., transformation (i). This is akin to identification in SVARs, cf. Section 1.2.1 and Uhlig (2005, Prop. A.1).

A.1.3 Likelihood evaluation

This subsection provides formulas for computing the exact Gaussian likelihood as well as the Whittle likelihood and score for the SVMA model.

A.1.3.1 Exact likelihood via the Kalman filter

Let $\Psi = \Theta\,\mathrm{diag}(\sigma)$. The state space representation of the SVMA model is

$$y_{i,t} = \Psi_i \alpha_t, \quad i = 1, \ldots, n,\ t = 1, \ldots, T,$$

$$\alpha_t = \begin{pmatrix} 0 & 0 \\ I_{nq} & 0 \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} \tilde{\varepsilon}_t \\ 0 \end{pmatrix}, \quad \tilde{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(0, I_n), \quad t = 2, 3, \ldots, T,$$

$$\alpha_1 \sim N(0, I_{n(q+1)}),$$

where Ψ_i is the $n(q+1)$-dimensional i-th row vector of $\Psi = (\Psi_0, \Psi_1, \ldots, \Psi_q)$, $\tilde{\varepsilon}_t$ is the n-dimensional standardized structural shock vector (each element has variance 1), and $\alpha_t = (\tilde{\varepsilon}_t', \tilde{\varepsilon}_{t-1}', \ldots, \tilde{\varepsilon}_{t-q}')'$ is the $n(q+1)$-dimensional state vector.

I use the "univariate treatment of multivariate series" Kalman filter in Durbin & Koopman (2012, Ch. 6.4), since that algorithm avoids inverting large matrices. For my purposes, the algorithm is as follows.

1. Initialize the state forecast mean $a_{1,1} = 0$ and state forecast variance $Z_{1,1} = I_{n(q+1)}$. Set $t = 1$.

2. For each $i = 1, \ldots, n$:

   (a) Compute the forecast error $v_{i,t} = y_{i,t} - \Psi_i a_{i,t}$, forecast variance $\lambda_{i,t} = \Psi_i Z_{i,t} \Psi_i'$, and Kalman gain $g_{i,t} = (1/\lambda_{i,t}) Z_{i,t} \Psi_i'$.

   (b) Compute the log likelihood contribution: $L_{i,t} = -\frac{1}{2}(\log \lambda_{i,t} + v_{i,t}^2/\lambda_{i,t})$.

   (c) Update the state forecast mean: $a_{i+1,t} = a_{i,t} + g_{i,t} v_{i,t}$.

   (d) Update the state forecast variance: $Z_{i+1,t} = Z_{i,t} - \lambda_{i,t} g_{i,t} g_{i,t}'$.

3. Let $\tilde{a}_{n+1,t}$ denote the first nq elements of $a_{n+1,t}$, and let $\tilde{Z}_{n+1,t}$ denote the upper left $nq \times nq$ block of $Z_{n+1,t}$. Set

$$a_{1,t+1} = \begin{pmatrix} 0 \\ \tilde{a}_{n+1,t} \end{pmatrix}, \quad Z_{1,t+1} = \begin{pmatrix} I_n & 0 \\ 0 & \tilde{Z}_{n+1,t} \end{pmatrix}.$$

4. If $t = T$, stop. Otherwise, increment t by 1 and go to step 2.

The log likelihood $\log p_{Y|\Psi}(Y_T \mid \Psi)$ is given by $\sum_{t=1}^{T}\sum_{i=1}^{n} L_{i,t}$, up to a constant.

A.1.3.2 Whittle likelihood

Let $Y_T = (y_1', y_2', \ldots, y_T')'$ be the stacked data vector. Let $V(\Psi)$ be an $nT \times nT$ symmetric block Toeplitz matrix consisting of $T \times T$ blocks of $n \times n$ matrices, where the (s,t) block is given by $\sum_{\ell=0}^{q-(t-s)} \Psi_{\ell+(t-s)}\Psi_\ell'$ for $t \geq s$ and the sum is taken to equal 0 when $t > s+q$. Then the exact log likelihood function can be written

$$\log p_{Y|\Psi}(Y_T \mid \Psi) = -\tfrac{1}{2}nT\log(2\pi) - \tfrac{1}{2}\log\det(V(\Psi)) - \tfrac{1}{2}Y_T'V(\Psi)^{-1}Y_T. \tag{A.1}$$

This is what the Kalman filter in Appendix A.1.3.1 computes.

For all $k = 0, 1, 2 \ldots, T-1$, define the Fourier frequencies $\omega_k = 2\pi k/T$, the discrete Fourier transform (DFT) of the data $\tilde{y}_k = (2\pi T)^{-1/2}\sum_{t=1}^{T} e^{-\imath\omega_k(t-1)}y_t$, the DFT of the MA parameters $\tilde{\Psi}_k(\Psi) = \sum_{\ell=1}^{q+1} e^{-\imath\omega_k(\ell-1)}\Psi_{\ell-1}$, and the SVMA spectral density matrix $f_k(\Psi) = (2\pi)^{-1}\tilde{\Psi}_k(\Psi)\tilde{\Psi}_k(\Psi)^*$ at frequency ω_k. Let $\|B\|_{\max} = \max_{ij}|B_{ij}|$ denote the maximum norm of any matrix $B = (B_{ij})$. Due to the block Toeplitz structure of $V(\Psi)$,

$$\|V(\Psi) - 2\pi\Delta F(\Psi)\Delta^*\|_{\max} = O(T^{-1}) \tag{A.2}$$

as $T \to \infty$. Δ is an $nT \times nT$ matrix with (s,t) block equal to $T^{-1/2}e^{\imath\omega_{s-1}(t-1)}I_n$, so that $\Delta\Delta^* = I_{nT}$. $F(\Psi)$ is a block diagonal $nT \times nT$ matrix with (s,s) block equal to $f_s(\Psi)$.[1]

The Whittle (1953) approximation to the log likelihood (A.1) is obtained by substituting $V(\Psi) \approx 2\pi\Delta F(\Psi)\Delta^*$. This yields the Whittle log likelihood

$$\log p^W_{Y|\Psi}(Y_T \mid \Psi) = -nT\log(2\pi) - \frac{1}{2}\sum_{k=0}^{T-1}\left\{\log\det(f_k(\Psi)) + \tilde{y}_k^*[f_k(\Psi)]^{-1}\tilde{y}_k\right\}.$$

[1]The result (A.2) is a straight-forward vector generalization of Brockwell & Davis (1991, Prop. 4.5.2).

The Whittle log likelihood is computationally cheap because $\{\tilde{y}_k, \tilde{\Psi}_k(\Psi)\}_{0\leq k\leq T-1}$ can be computed efficiently using the Fast Fourier Transform (Hansen & Sargent, 1981, Sec. 2b; Brockwell & Davis, 1991, Ch. 10.3).[2]

Now I derive the gradient of the Whittle log likelihood. For all $k = 0, 1, \ldots, T-1$, define $C_k(\Psi) = [f_k(\Psi)]^{-1} - [f_k(\Psi)]^{-1}\tilde{y}_k\tilde{y}_k^*[f_k(\Psi)]^{-1}$ and $\tilde{C}_k(\Psi) = \sum_{\ell=1}^{T} e^{-\imath\omega_k(\ell-1)}C_{\ell-1}(\Psi)$. Once $\{C_k(\Psi)\}_{0\leq k\leq T-1}$ have been computed, $\{\tilde{C}_k(\Psi)\}_{0\leq k\leq T-1}$ can be computed using the Fast Fourier Transform.[3] Finally, let $\tilde{C}_k(\Psi) = \tilde{C}_{T+k}(\Psi)$ for $k = -1, -2, \ldots, 1-T$.

Lemma A.1.

$$\frac{\log p^W_{Y|\Psi}(Y_T \mid \Psi)}{\partial \Psi_\ell} = -\sum_{\tilde{\ell}=0}^{q} \mathrm{Re}[\tilde{C}_{\tilde{\ell}-\ell}(\Psi)]\Psi_{\tilde{\ell}}, \quad \ell = 0, 1, \ldots, q. \tag{A.3}$$

The lemma gives the score with respect to Ψ. Since $\Psi_\ell = \Theta_\ell \,\mathrm{diag}(\sigma)$, the chain rule gives the score with respect to Θ and $\log\sigma$:

$$\frac{\partial \log p^W_{Y|\Psi}(Y_T \mid \Psi)}{\partial \Theta_\ell} = \frac{\partial \log p^W_{Y|\Psi}(Y_T \mid \Psi)}{\partial \Psi_\ell}\,\mathrm{diag}(\sigma), \quad \ell = 0, 1, \ldots, q,$$

$$\frac{\partial \log p^W_{Y|\Psi}(Y_T \mid \Psi)}{\partial \log \sigma_j} = \sum_{i=1}^{n}\sum_{\ell=0}^{q} \frac{\partial \log p^W_{Y|\Psi}(Y_T \mid \Psi)}{\partial \Psi_{ij,\ell}}\Psi_{ij,\ell}, \quad j = 1, 2, \ldots, n.$$

A.1.4 Bayesian computation: Algorithms

This subsection details my posterior simulation algorithm and the optional reweighting step that translates Whittle draws into draws from the exact likelihood.

---

[2] As noted by Hansen & Sargent (1981, p. 32), the computation time can be halved by exploiting $\tilde{y}_{T-k} = \overline{\tilde{y}_k}$ and $f_{T-k}(\Psi) = \overline{f_k(\Psi)}$ for $k = 1, 2, \ldots, T$.

[3] Again, computation time can be saved by exploiting $C_{T-k}(\Psi) = \overline{C_k(\Psi)}$ for $k = 1, 2, \ldots, T$.

A.1.4.1 Implementation of Hamiltonian Monte Carlo algorithm

I here describe my implementation of the posterior simulation algorithm. First I outline my method for obtaining an initial value. Then I discuss the modifications I make to the Hoffman & Gelman (2014) algorithm. The calculations below require evaluation of the log prior density, its gradient, the log likelihood, and the score. Evaluation of the multivariate Gaussian log prior and its gradient is straight-forward; this is also the case for many other choices of priors. Evaluation of the Whittle likelihood and its score is described in Appendix A.1.3.2.

Initial value. The HMC algorithm produces draws from a Markov Chain whose long-run distribution is the Whittle posterior of the SVMA parameters, regardless of the initial value used for the chain. However, using an initial value near the mode of the posterior distribution can significantly speed up the convergence to the long-run distribution. I approximate the posterior mode using the following computationally cheap procedure:

1. Compute the empirical ACF of the data.

2. Run q steps of the Innovations Algorithm to obtain an invertible SVMA representation that approximately fits the empirical ACF (Brockwell & Davis, 1991, Prop. 11.4.2).[4] Denote these invertible parameters by $(\hat{\Theta}, \hat{\sigma})$.

3. Let $\mathcal{C}$ denote the (finite) set of complex roots of the SVMA polynomial corresponding to $(\hat{\Theta}, \hat{\sigma})$, cf. Theorem A.1.

4. For each root γ_j in $\mathcal{C}$ (each complex conjugate pair of roots is treated as one root):

[4] In principle, the Innovations Algorithm could be run for more than q steps, but this tends to lead to numerical instability in my trials. The output of the first q steps is sufficiently accurate in my experience.

(a) Let $(\check{\Theta}^{(j)}, \check{\sigma}^{(j)})$ denote the result of flipping root γ_j, i.e., of applying transformation (ii) in Theorem A.1 to $(\hat{\Theta}, \hat{\sigma})$ with this root.

(b) Determine the orthogonal matrix $Q^{(j)}$ such that $\check{\Theta}^{(j)} \operatorname{diag}(\check{\sigma})^{(j)} Q^{(j)}$ is closest to the prior mean $E(\Theta \operatorname{diag}(\sigma))$ in Frobenius norm, cf. Footnote 47.

(c) Obtain parameters $(\tilde{\Theta}^{(j)}, \tilde{\sigma}^{(j)})$ such that $\check{\Theta}^{(j)} \operatorname{diag}(\check{\sigma}^{(j)}) Q^{(j)} = \tilde{\Theta}^{(j)} \operatorname{diag}(\tilde{\sigma}^{(j)})$, i.e., apply transformation (a) in Theorem A.1. Calculate the corresponding value of the prior density $\pi(\tilde{\Theta}^{(j)}, \tilde{\sigma}^{(j)})$.

5. Let $\tilde{j} = \arg\max_j \pi(\tilde{\Theta}^{(j)}, \tilde{\sigma}^{(j)})$.

6. If $\pi(\tilde{\Theta}^{(\tilde{j})}, \tilde{\sigma}^{(\tilde{j})}) \leq \pi(\hat{\Theta}, \hat{\sigma})$, go to Step 7. Otherwise, set $(\hat{\Theta}, \hat{\sigma}) = (\tilde{\Theta}^{(\tilde{j})}, \tilde{\sigma}^{(\tilde{j})})$, remove γ_j (and its complex conjugate) from $\mathcal{C}$, and go back to Step 4.

7. Let the initial value for the HMC algorithm be the parameter vector of the form $((1-x)\check{\Theta} + xE(\Theta), (1-x)\check{\sigma} + xE(\sigma))$ that maximizes the posterior density, where x ranges over the grid $\{0, 0.01, \ldots, 0.99, 1\}$, and $(E(\Theta), E(\sigma))$ is the prior mean of (Θ, σ).

Step 2 computes a set of invertible parameters that yields a high value of the likelihood. Steps 3–6 find a set of possibly noninvertible parameters that yields a high value of the prior density while being observationally equivalent with the parameters from Step 2 (I use a "greedy" search algorithm since it is computationally prohibitive to consider all combinations of root flips). Because Steps 2–6 lexicographically prioritize maximizing the likelihood over maximizing the prior, Step 7 allows the parameters to shrink toward the prior means.

HMC IMPLEMENTATION. I use the HMC variant NUTS from Hoffman & Gelman (2014), which automatically tunes the step size and trajectory length of HMC. See their paper for details on the NUTS algorithm. I downloaded the code from Hoffman's website. I make two modifications to the basic NUTS algorithm, neither of which are essential, although they do

tend to improve the mixing speed of the Markov chain in my trials. These modifications are also used in the NUTS-based statistics software Stan (Stan Development Team, 2015).

First, I allow for step size jittering, i.e., continually drawing a new HMC step size from a uniform distribution over some interval (Neal, 2011, Sec. 5.4.2.2). The jittering is started after the stepsize has been tuned as described in Hoffman & Gelman (2014, Sec. 3.2). For the applications in this paper, the step size is chosen uniformly at random from the interval $[0.5\hat{\epsilon}, 1.5\hat{\epsilon}]$, where $\hat{\epsilon}$ is the tuned step size.

Second, I allow for a diagonal "mass matrix", where the entries along the diagonal are estimates of the posterior standard deviations of the SVMA parameters (Neal, 2011, Sec. 5.4.2.4). I first run the NUTS algorithm for a number of steps with an identity mass matrix. Then I calculate the sample standard deviations of the parameter draws over a window of subsequent steps, after which I update the mass matrix accordingly.[5] I update the mass matrix twice more using windows of increasing length. Finally, I freeze the mass matrix for the remainder of the NUTS algorithm. In this paper, the mass matrix is estimated over steps 300–400, steps 401–600, and steps 601–1000, and it is fixed after step 1000.

A.1.4.2 Reweighting

An optional reweighting step may be used to translate draws obtained from the Whittle-based HMC algorithm into draws from the exact Gaussian posterior density $p_{\Theta,\sigma|Y}(\Theta, \sigma \mid Y_T)$. The Whittle HMC algorithm yields draws $(\Theta^{(1)}, \sigma^{(1)}), \ldots, (\Theta^{(N)}, \sigma^{(N)})$ (after discarding a burn-in sample) from the Whittle posterior density $p^W_{\Theta,\sigma|Y}(\Theta, \sigma \mid Y_T)$. If desired, apply the following reweighting procedure to the Whittle draws:

[5]The sample standard deviations are partially shrunk toward 1 before updating the mass matrix.

1. For each Whittle draw $k = 1, 2, \ldots, N$, compute the relative likelihood weight

$$w_k = \frac{p_{\Theta,\sigma|Y}(\Theta^{(k)}, \sigma^{(k)} \mid Y_T)}{p^W_{\Theta,\sigma|Y}(\Theta^{(k)}, \sigma^{(k)} \mid Y_T)} = \frac{p_{Y|\Psi}(Y_T \mid \Psi(\Theta^{(k)}, \sigma^{(k)}))}{p^W_{Y|\Psi}(Y_T \mid \Psi(\Theta^{(k)}, \sigma^{(k)}))}.$$

2. Compute normalized weights $\tilde{w}_k = w_k / \sum_{b=1}^N w_b$, $k = 1, \ldots, N$.

3. Draw N samples $(\tilde{\Theta}^{(1)}, \tilde{\sigma}^{(1)}), \ldots, (\tilde{\Theta}^{(N)}, \tilde{\sigma}^{(N)})$ from the multinomial distribution with mass points $(\Theta^{(1)}, \sigma^{(1)}), \ldots, (\Theta^{(N)}, \sigma^{(N)})$ and corresponding probabilities $\tilde{w}_1, \ldots, \tilde{w}_N$.

Then $(\tilde{\Theta}^{(1)}, \tilde{\sigma}^{(1)}), \ldots, (\tilde{\Theta}^{(N)}, \tilde{\sigma}^{(N)})$ constitute N draws from the exact posterior distribution. This reweighting procedure is a Sampling-Importance-Resampling procedure (Rubin, 1988) that uses the Whittle posterior as a proposal distribution. The reweighting step is fast, as it only needs to compute the exact likelihood – not the score – for N different parameter values, where N is typically orders of magnitude smaller than the required number of likelihood/score evaluations during the HMC algorithm.

A.1.5 Simulation study: Additional results

Here I provide diagnostics and additional results relating to the simulations in Section 1.4.

A.1.5.1 Diagnostics for simulations

I report diagnostics for the baseline $\rho_{ij} = 0.9$ bivariate simulation, but diagnostics for other specifications in this paper are similar. The average HMC acceptance rate is slightly higher than 0.60, which is the rate targeted by the NUTS algorithm when tuning the HMC step size. The score of the posterior was evaluated about 382,000 times. Figures A.1 and A.2 show the MCMC chains for the IRF and log shock standard deviation draws. Figures A.3 and A.4 show the autocorrelation functions of the draws.

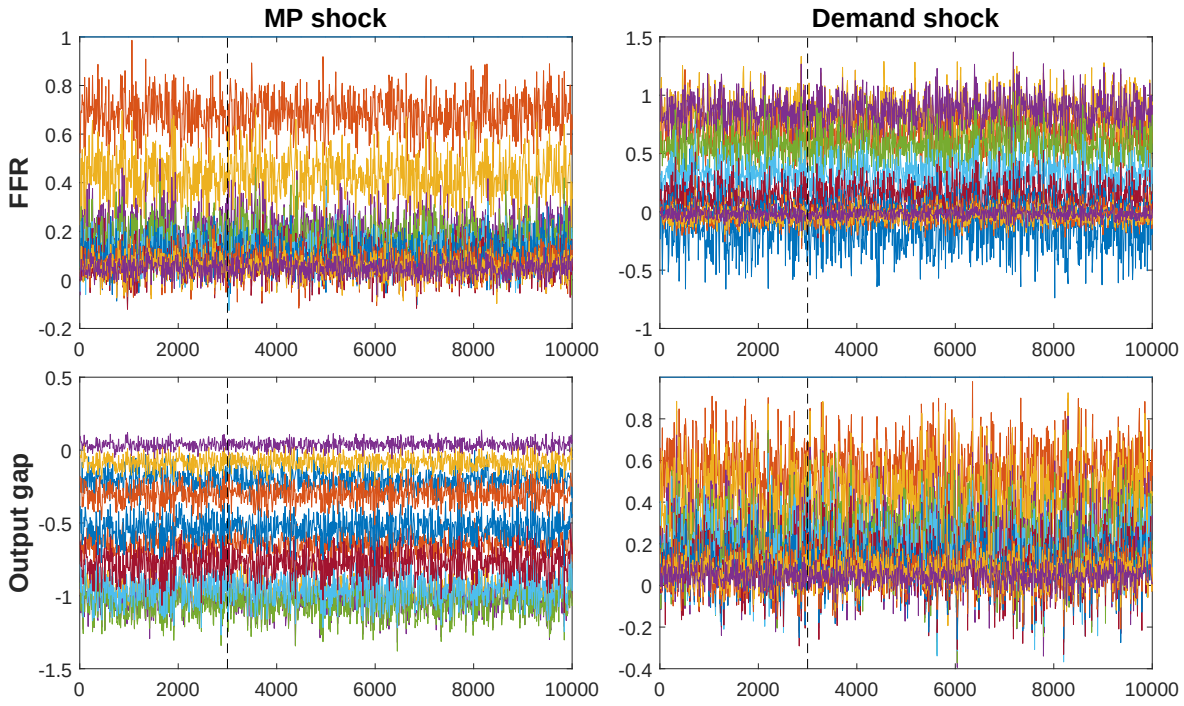

Figure A.1: MCMC chains for each IRF parameter (Θ) in the $\rho_{ij} = 0.9$ simulations in Section 1.4. Each jagged line represents a different impulse response parameter (two of them are normalized at 1). The vertical dashed line marks the burn-in time, before which all draws are discarded. The horizontal axes are in units of MCMC steps, not stored draws (every 10th step is stored).

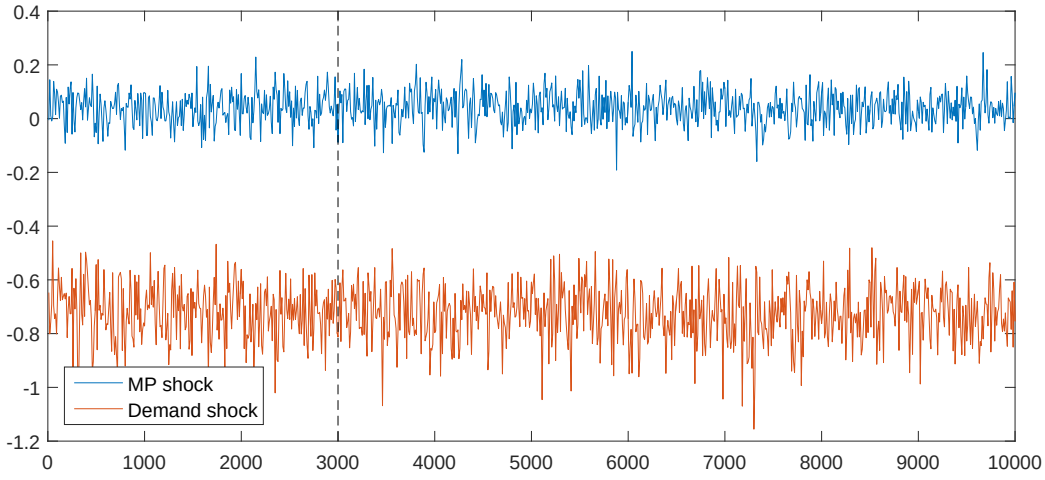

Figure A.2: MCMC chains for each log shock standard deviation parameter ($\log \sigma$) in the $\rho_{ij} = 0.9$ simulations in Section 1.4. See caption for Figure A.1.

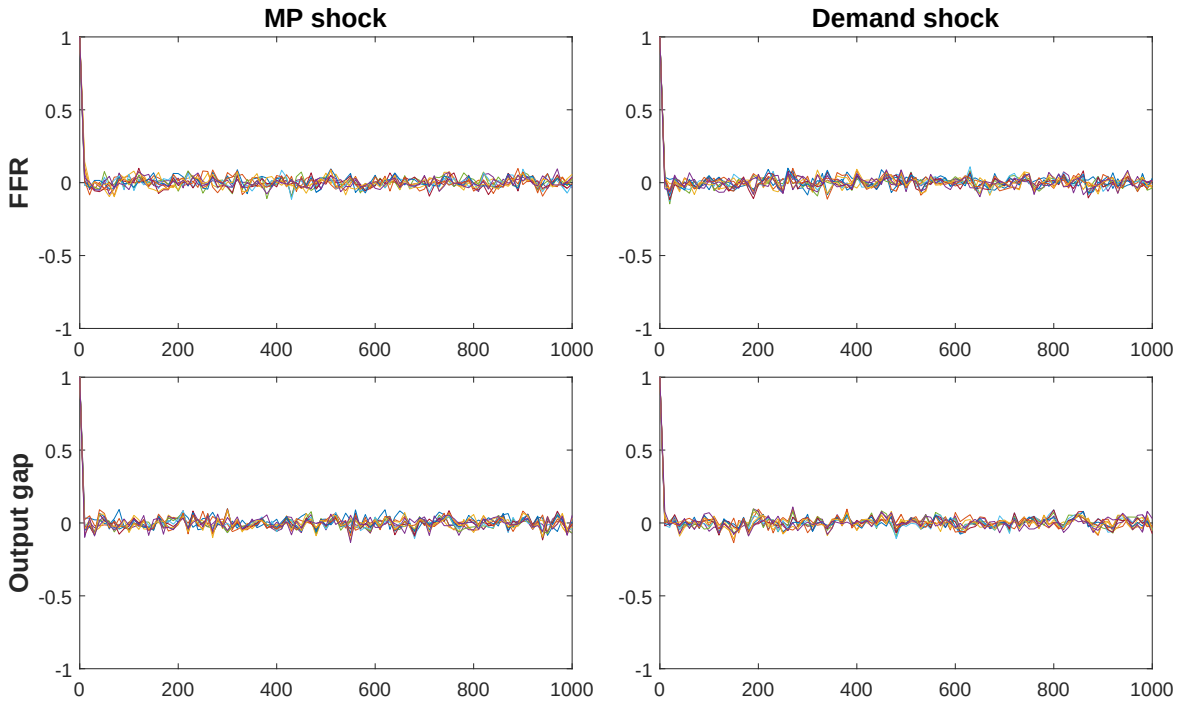

Figure A.3: Autocorrelation functions for HMC draws of each IRF parameter (Θ) in the $\rho_{ij} = 0.9$ simulations in Section 1.4. Each jagged line represents a different impulse response parameter. Only draws after burn-in were used to computed these figures. The autocorrelation lag is shown on the horizontal axes in units of MCMC steps.

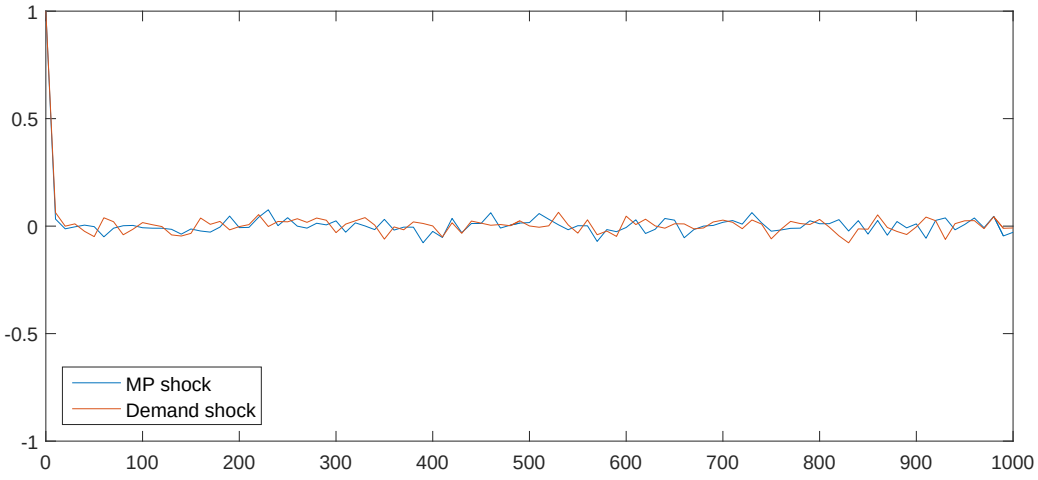

Figure A.4: Autocorrelation functions for HMC draws of each log shock standard deviation parameter ($\log \sigma$) in the $\rho_{ij} = 0.9$ simulations in Section 1.4. See caption for Figure A.3.

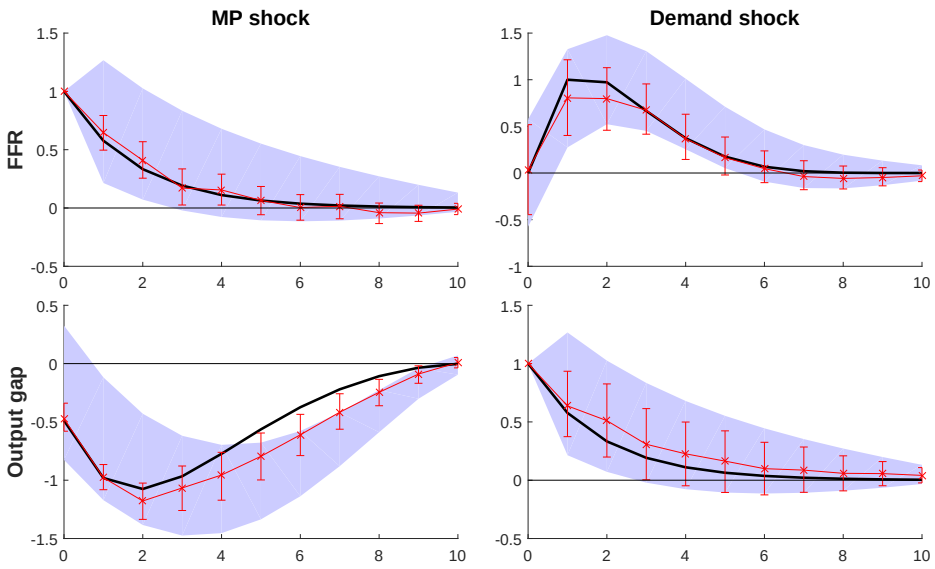

Figure A.5: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with a prior that is too persistent relative to the true parameter values. The plots show true values (thick lines), prior 90% confidence bands (shaded), posterior means (crosses), and posterior 5–95 percentile intervals (vertical bars). The prior means (not shown) are the midpoints of the prior confidence bands, as in Figure 1.4.

A.1.5.2 Simulations with misspecified priors

I here provide simulation results for two bivariate experiments with substantially misspecified priors. I maintain the same prior on IRFs and shock standard deviations as the $\rho_{ij} = 0.9$ prior in Section 1.4, cf. Figures 1.4 and 1.7. Here, however, I modify the true values of the IRFs so they no longer coincide with the prior means.

Misspecified persistence. I first consider an experiment in which the prior overstates the persistence of the shock effects, i.e., the true IRFs die out quicker than indicated by the prior means $\mu_{ij,\ell}$ in Figure 1.4. The true IRFs are set to $\Theta_{ij,\ell} = c_{ij} e^{-0.25\ell} \mu_{ij,\ell}$ for all (i, j, ℓ), where $c_{ij} > 0$ is chosen so that $\max_\ell |\Theta_{ij,\ell}| = \max_\ell |\mu_{ij,\ell}|$ for each IRF. The true shock standard deviations, the prior ($\rho_{ij} = 0.9$), the sample size, and the HMC settings are exactly as in Section 1.4. Figure A.5 compares these true IRFs to the prior distribution. The figure also summarizes the posterior distribution for the IRFs. The posterior is not perfectly

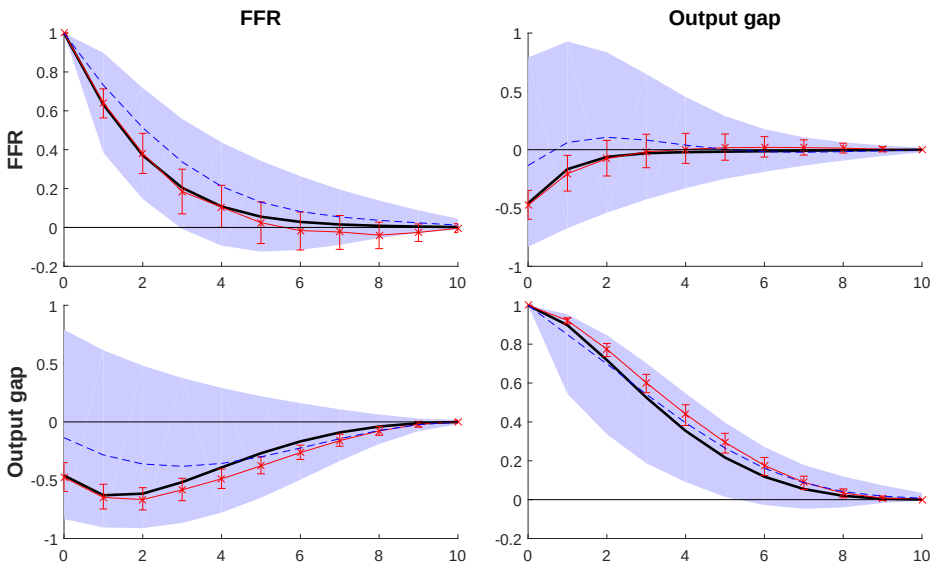

Figure A.6: Posterior auto- and cross-correlation draws for the bivariate SVMA model with a prior that misspecifices the persistence of the IRFs. The displays plot draws of $\text{Corr}(y_{i,t}, y_{j,t-k} \mid \Theta, \sigma)$, where i indexes rows, j indexes columns, and k runs along the horizontal axes. The top right display, say, concerns cross-correlations between the FFR and lags of the output gap. The plots show true values (thick lines), prior means (dashed lines) and 5–95 percentile confidence bands (shaded), and posterior means (crosses) and 5–95 percentile intervals (vertical bars).

centered but is much closer to the truth than the prior is. Figure A.6 shows why this is the case: The prior distribution on (Θ, σ) implies a distribution for auto- and cross-correlations of observed variables that is at odds with the true ACF. Since the data is informative about the ACF, the posterior distribution for IRFs puts higher weight than the prior on IRFs that are consistent with the true auto- and cross-correlations.

Misspecified cross-correlations. The second experiment considers a prior that misspecifies the cross-correlations between the observed variables. I set the true IRFs equal to the prior means in Figure 1.4, except that the true IRF of the output gap to a monetary policy shock equals zero, i.e., $\Theta_{21,\ell} = 0$ for $0 \leq \ell \leq q$. The true shock standard deviations, the prior ($\rho_{ij} = 0.9$), the sample size, and the HMC settings are as above. Figure A.7 shows that posterior inference is accurate despite the misspecified prior. Again, Figure A.8 demonstrates how the data corrects the prior distribution on auto- and cross-correlations,

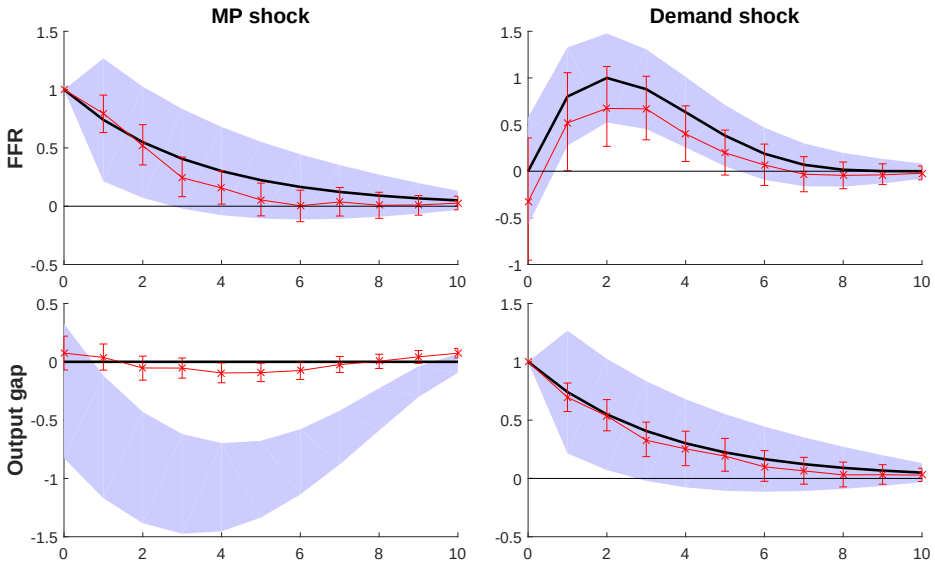

Figure A.7: Summary of posterior IRF (Θ) draws for the bivariate SVMA model with a prior that misspecifies the cross-correlations between variables. See caption for Figure A.5.

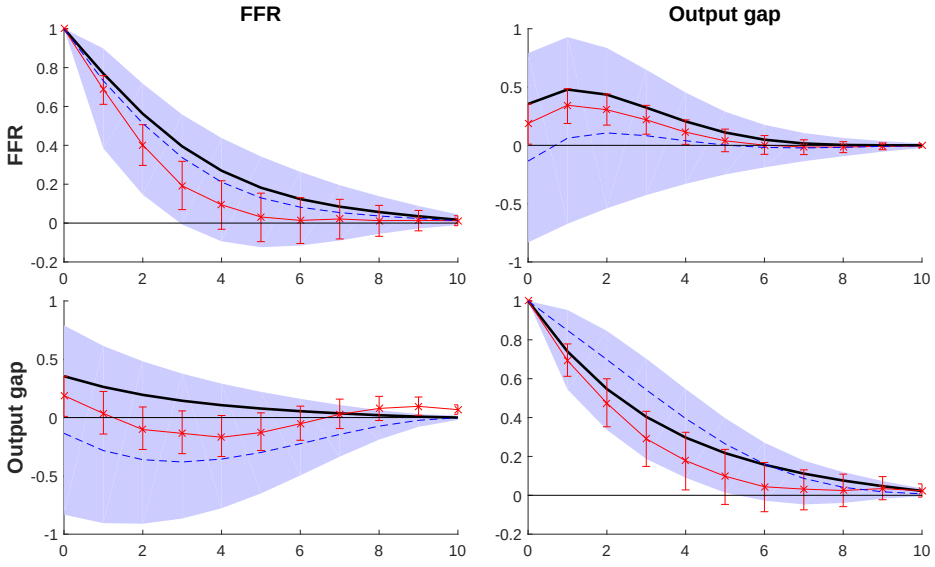

Figure A.8: Posterior autocorrelation draws for the bivariate SVMA model with a prior that misspecifies the cross-correlations between variables. See caption for Figure A.6.

thus pulling the posterior on IRFs toward the true values (although here the true ACF is not estimated as accurately as in Figure A.6).

A.1.6 Application: Additional results

This subsection presents additional results related to the empirical application in Section 1.5. First, I show that the SVMA procedure accurately estimates IRFs on simulated data. Second, I demonstrate how the Kalman smoother can be used to draw inference about the shocks. Third, I examine the sensitivity of posterior inference with respect to the choice of prior. Fourth, I assess the model's fit and suggest ways to improve it.

A.1.6.1 Consistency check with simulated data

I show that the SVMA approach, with the same prior and HMC settings as in Section 1.5, can recover the true IRFs when applied to data generated by the log-linearized Sims (2012) DSGE model. I simulate data for the three observed variables from an SVMA model with i.i.d. Gaussian shocks. The true IRFs are those implied by the log-linearized Sims (2012) model (baseline calibration) out to horizon $q = 16$, yielding a noninvertible representation. The true shock standard deviations are set to $\sigma = (0.5, 0.5, 0.5)'$. Note that the prior for the IRF of TFP growth to the news shock is not centered at the true IRF, as explained in Section 1.5. The sample size is the same as for the actual data ($T = 213$).

Figures A.9 and A.10 summarize the posterior draws produced by the HMC algorithm when applied to the simulated data set. The posterior means accurately locate the true parameter values. The equal-tailed 90% posterior credible intervals are tightly concentrated around the truth in most cases. In particular, inference about the shock standard deviation parameters is precise despite the very diffuse prior.

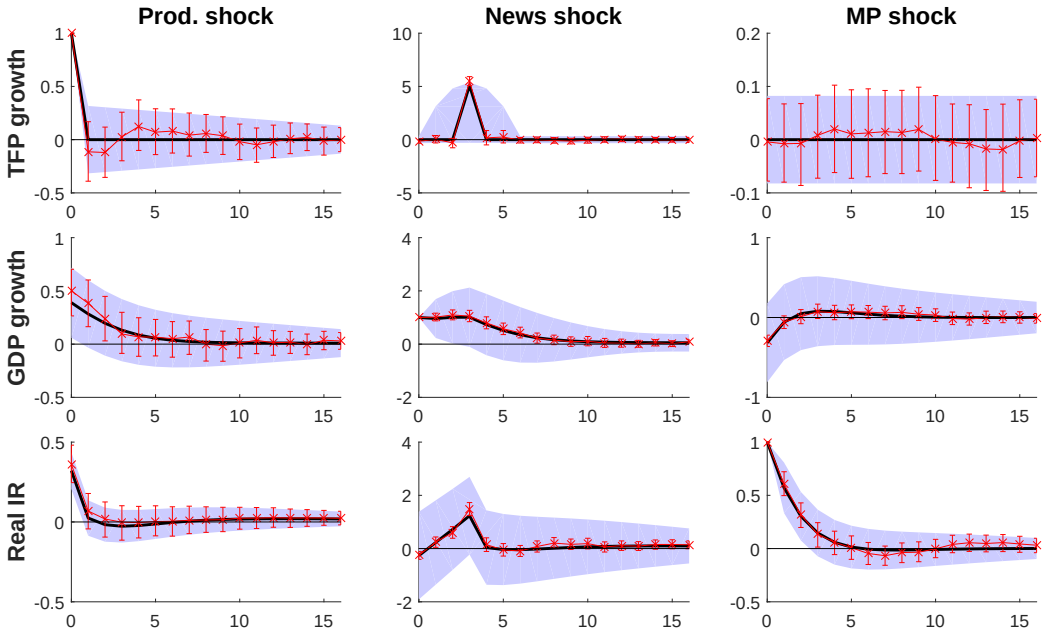

Figure A.9: Summary of posterior IRF (Θ) draws, simulated news shock data. See caption for Figure 1.6.

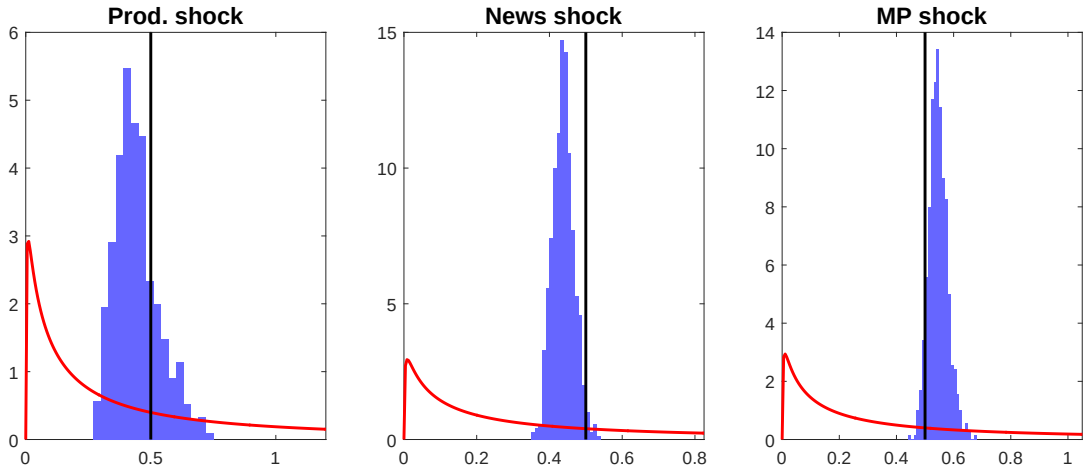

Figure A.10: Summary of posterior shock standard deviation (σ) draws, simulated news shock data. See caption for Figure 1.7.

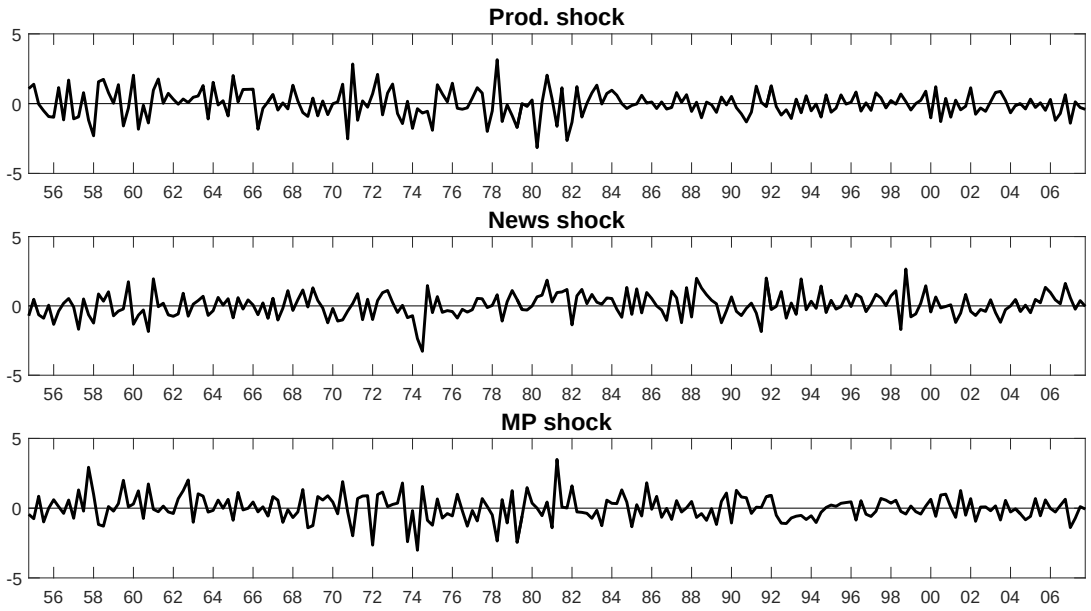

Figure A.11: Posterior means of *standardized* structural shocks ($\varepsilon_{jt}/\sigma_j$) at each point in time, news shock application.

A.1.6.2 Inference about shocks

Figure A.11 shows the time series of posterior means for the structural shocks given the real dataset. For each posterior draw of the structural parameters (Θ, σ), I compute $E(\varepsilon_t \mid \Theta, \sigma, Y_T)$ using the smoothing recursions corresponding to the Gaussian state-space representation in Appendix A.1.3.1 (Durbin & Koopman, 2012, p. 157), and then I average over draws. If the structural shocks are in fact non-Gaussian, the smoother still delivers mean-square-error-optimal linear estimates of the shocks. If desired, draws from the full joint posterior distribution of the shocks can be obtained from a simulation smoother (Durbin & Koopman, 2012, Ch. 4.9). It is also straight-forward to draw from the predictive distribution of future values of the data using standard methods for state-space models.

A.1.6.3 Prior sensitivity

To gauge the robustness of posterior inference with respect to the choice of prior, I compute the sensitivity measure "PS" of Müller (2012). This measure captures the first-order approximate effect on the posterior means of changing the prior mean hyperparameters. Let θ denote the vector containing all impulse responses and log shock standard deviations of the SVMA model, and let e_k denote the k-th unit vector. Because my prior for θ is a member of an exponential family, the Müller (2012) PS measure for parameter θ_k equals

$$PS_k = \max_{\nu:\, \sqrt{\nu' \mathrm{Var}(\theta)^{-1} \nu} = 1} \frac{\partial E(\theta_k \mid Y_T)}{\partial E(\theta)'} \nu = \sqrt{e_k' \mathrm{Var}(\theta \mid Y_T) \mathrm{Var}(\theta)^{-1} \mathrm{Var}(\theta \mid Y_T) e_k}. \tag{A.4}$$

This is the largest (local) change that can be induced in the posterior mean of θ_k from changing the prior means of the components of θ by the multivariate equivalent of 1 prior standard deviation.[6] PS_k depends only on the prior and posterior variance matrices $\mathrm{Var}(\theta)$ and $\mathrm{Var}(\theta \mid Y_T)$, which are easily obtained from the HMC output.

Figure A.12 plots the posterior means of the impulse responses along with $\pm PS_k$ intervals (where the index k corresponds to the (i, j, ℓ) combination for each impulse response). The wider the band around an impulse response, the more sensitive is the posterior mean of that impulse response to (local) changes in the prior. In economic terms, most of the posterior means are seen to be insensitive to changes in the prior means of magnitudes smaller than 1 prior standard deviation. The most prior-sensitive posterior inferences, economically speaking, concern the IRF of GDP growth to a news shock, but large changes in the prior means are necessary to alter the qualitative features of the posterior mean IRF.

---

[6] In particular, $PS_k \geq \max_b |\partial E(\theta_k \mid Y_T)/\partial E(\theta_b)| \sqrt{\mathrm{Var}(\theta_b)}$. Whereas PS_k is a local measure, the effects of large changes in the prior can be evaluated using reweighting (Lopes & Tobias, 2011, Sec. 2.4).

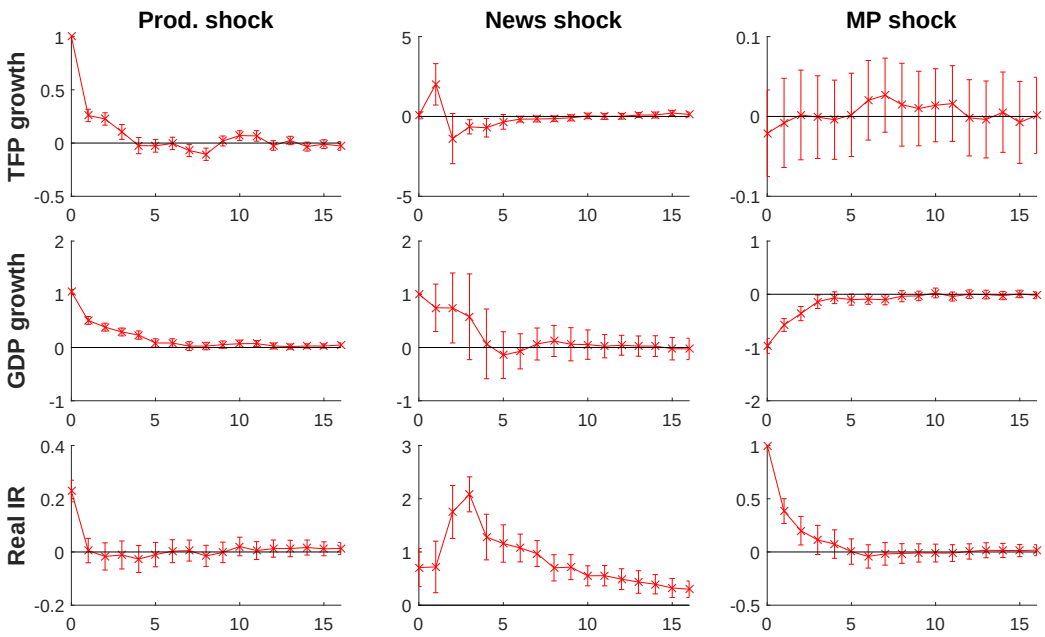

Figure A.12: PS_k measure of the sensitivity of the posterior IRF means with respect to changes in the prior means of all parameters, cf. (A.4), in the news shock application. The symmetric vertical bars have length $2PS_k$ and are centered around the corresponding posterior means (crosses).

A.1.6.4 Posterior predictive analysis

I conduct a posterior predictive analysis to identify ways to improve the fit of the Gaussian SVMA model (Geweke, 2010, Ch. 2.4.2). For each posterior parameter draw produced by HMC, I simulate an artificial dataset of sample size $T = 213$ from a Gaussian SVMA model with the given parameters. On each artificial dataset I compute four checking functions. First and second, the skewness and excess kurtosis of each series. Third, the long-run autocorrelation of each series, defined as the Newey-West long-run variance estimator (20 lags) divided by the sample variance. Fourth, I run a reduced-form VAR regression of the three-dimensional data vector y_t on its 8 first lags and a constant; then I compute the first autocorrelation of the squared VAR residuals for each of the three series. The third measure captures persistence, while the fourth measure captures volatility clustering in forecast errors.

Figure A.13 shows the distribution of checking function values across simulated datasets,

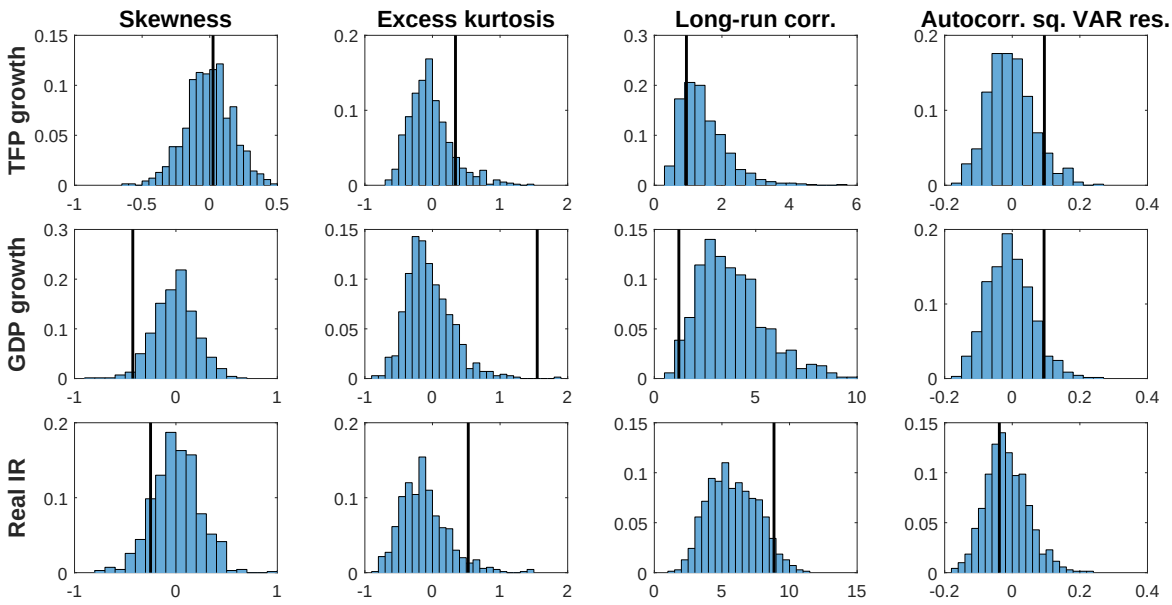

Figure A.13: Posterior predictive checks, news shock application. Observed variables along rows, checking functions along columns. Histograms show the distribution of checking function values on simulated datasets based on the posterior parameter draws; thick vertical lines mark checking function values on actual data. Checking functions from left to right: skewness; excess kurtosis; Newey-West long-run variance estimate (20 lags) divided by sample variance; first autocorrelation of squared residuals from a VAR regression of y_t on a constant and 8 lags.

as well as the corresponding checking function values for the actual data. The Gaussian SVMA model does not capture the skewness and kurtosis of GDP growth; essentially, the model does not generate recessions that are sufficiently severe relative to the size of booms. The model somewhat undershoots the persistence and kurtosis of the real interest rate. The fourth column suggests that forecast errors for TFP and GDP growth exhibit volatility clustering in the data, which is not captured by the Gaussian SVMA model.

The results point to three fruitful model extensions. First, introducing stochastic volatility in the SVMA model would allow for better fit along the dimensions of kurtosis and forecast error volatility clustering. Second, nonlinearities or skewed shocks could capture the negative skewness of GDP growth. Finally, increasing the MA lag length q would allow the model to better capture the persistence of the real interest rate, although this is not a

major concern, as I am primarily interested in shorter-run impulse responses.

A.1.7 Asymptotic theory: Mathematical details

I here give additional details concerning the frequentist asymptotics of Bayes procedures. First I provide high-level sufficient conditions for posterior consistency. Then I state the functional form for the Whittle ACF likelihood mentioned in Section 1.6.2. Finally, I prove posterior consistency for the parameters in the Wold decomposition of a q-dependent time series, which is useful for proving posterior consistency for the ACF (Theorem 1.1). I allow for misspecification of the likelihood functions used to compute the various posterior measures.

All stochastic limits below are taken as $T \to \infty$, and all stochastic limits and expectations are understood to be taken under the true probability measure of the data. The abbreviation "w.p.a. 1" means "with probability approaching 1 as $T \to \infty$".

A.1.7.1 General conditions for posterior consistency

Following Ghosh & Ramamoorthi (2003, Thm. 1.3.4), I give general sufficient conditions for assumption (ii) of Lemma 1.1. Let $\Pi_\Gamma(\cdot)$ denote the marginal prior measure for parameter Γ, with parameter space Ξ_Γ. Let $p_{Y|\Gamma}(Y_T \mid \Gamma)$ denote the (possibly misspecified) likelihood function. The posterior measure is given by

$$P_{\Gamma|Y}(\mathcal{A} \mid Y_T) = \frac{\int_{\mathcal{A}} p_{Y|\Gamma}(Y_T \mid \Gamma)\Pi_\Gamma(d\Gamma)}{\int_{\Xi_\Gamma} p_{Y|\Gamma}(Y_T \mid \Gamma)\Pi_\Gamma(d\Gamma)}$$

for measurable sets $\mathcal{A} \subset \Xi_\Gamma$.[7]

Lemma A.2. *Define the normalized log likelihood ratio* $\hat{\phi}(\Gamma) = T^{-1} \log \frac{p_{Y|\Gamma}(Y_T|\Gamma)}{p_{Y|\Gamma}(Y_T|\Gamma_0)}$ *for all* $\Gamma \in \Xi_\Gamma$. *Assume there exist a function* $\phi\colon \Xi_\Gamma \to \mathbb{R}$, *a neighborhood* $\mathcal{K}$ *of* Γ_0 *in* Ξ_Γ, *and a*

[7] I assume throughout the paper that integrals in the definitions of posterior measures are well-defined.

scalar $\zeta < 0$ *such that the following conditions hold.*

(i) $\sup_{\Gamma \in \mathcal{K}} |\hat{\phi}(\Gamma) - \phi(\Gamma)| \xrightarrow{p} 0$.

(ii) $\phi(\Gamma)$ *is continuous at* $\Gamma = \Gamma_0$.

(iii) $\phi(\Gamma) < 0$ *for all* $\Gamma \neq \Gamma_0$.

(iv) $\sup_{\Gamma \in \mathcal{K}^c} \hat{\phi}(\Gamma) < \zeta$ *w.p.a. 1.*

(v) Γ_0 *is in the support of* $\Pi_\Gamma(\cdot)$.

Then for any neighborhood $\mathcal{U}$ *of* Γ_0 *in* Ξ_Γ, $P_{\Gamma|Y}(\mathcal{U} \mid Y_T) \xrightarrow{p} 1$.

Remarks:

1. The uniform convergence assumption (i) on the log likelihood ratio can often be obtained from pointwise convergence using stochastic equicontinuity (Andrews, 1992).

2. If the likelihood $p_{Y|\Gamma}(Y_T \mid \Gamma)$ is correctly specified (i.e., $p_{Y|\Gamma}(Y_T \mid \Gamma_0)$ is the true density of the data) and assumption (i) holds, $\phi(\Gamma)$ equals the negative Kullback-Leibler divergence and assumptions (ii)–(iii) will typically be satisfied automatically if Γ is identified (Ghosh & Ramamoorthi, 2003, Ch. 1.2–1.3). Even if the likelihood is misspecified, such as with the use of a Whittle likelihood in time series models, it may still be possible to prove the uniform convergence in assumption (i) to some $\phi(\cdot)$ function that uniquely identifies the true parameter Γ_0 through assumptions (ii)–(iii), as in Theorem 1.1. This phenomenon is analogous to the well-known consistency property of quasi maximum likelihood estimators under certain types of misspecification.

A.1.7.2 Whittle likelihood for a q-dependent process

I now state the functional form for the Whittle ACF likelihoods for q-dependent processes mentioned in Section 1.6.2. Define the spectral density for a q-dependent process

parametrized in terms of its ACF:

$$f(\omega;\Gamma) = \frac{1}{2\pi}\left(\Gamma(0) + \sum_{k=1}^{q}\{e^{-\imath k\omega}\Gamma(k) + e^{\imath k\omega}\Gamma(k)'\}\right), \quad \Gamma \in \mathbb{T}_{n,q}.$$

Let $\hat{\Gamma}(k) = T^{-1}\sum_{t=1}^{T-k} y_{t+k}y_t'$, $k = 0, 1, \ldots, T-1$, be the k-th sample autocovariance, and set $\hat{\Gamma}(k) = \hat{\Gamma}(-k)'$ for $k = -1, -2, \ldots, 1-T$. Define the periodogram

$$\hat{I}(\omega) = \frac{1}{2\pi T}\left(\sum_{t=1}^{T} e^{-\imath t\omega}y_t\right)\left(\sum_{t=1}^{T} e^{\imath t\omega}y_t'\right) = \frac{1}{2\pi}\sum_{k=-(T-1)}^{T-1} e^{-\imath k\omega}\hat{\Gamma}(k), \quad \omega \in [-\pi, \pi].$$

The Whittle ACF log likelihood is given by

$$\log p^{W}_{Y|\Gamma}(Y_T \mid \Gamma) = -nT\log(2\pi) - \frac{T}{4\pi}\int_{-\pi}^{\pi}\log\det(f(\omega;\Gamma))\,d\omega - \frac{T}{4\pi}\int_{-\pi}^{\pi}\mathrm{tr}\{f(\omega;\Gamma)^{-1}\hat{I}(\omega)\}\,d\omega.$$

As in Appendix A.1.3.2, it is common to use a discretized Whittle log likelihood that replaces integrals of the form $(2\pi)^{-1}\int_{-\pi}^{\pi} g(\omega)\,d\omega$ (for some 2π-periodic function $g(\cdot)$) with corresponding sums $T^{-1}\sum_{k=0}^{T-1} g(\omega_k)$, where $\omega_k = 2\pi k/T$ for $0 \le k \le T-1$. The discretization makes it possible to compute the periodogram from the DFT $\tilde{y}_k$ of the data (see Appendix A.1.3.2), since $\hat{I}(\omega_k) = \tilde{y}_k\tilde{y}_k^*$ for $0 \le k \le T-1$. The proof of Theorem 1.1 shows that posterior consistency also holds when the discretized Whittle likelihood is used.

A.1.7.3 Posterior consistency for Wold parameters

The proof of Theorem 1.1 relies on a posterior consistency result for the Wold IRFs and prediction covariance matrix in a MA model with q lags. I state this result below. The posterior for the reduced-form parameters is computed using the Whittle likelihood and thus under the working assumption of a MA model with i.i.d. Gaussian innovations. However, consistency only requires Assumption 1.3, so the true data distribution need not be Gaussian or q-dependent. The result in this subsection concerns (invertible) reduced-form IRFs, not

(possibly noninvertible) structural IRFs. While the consistency result may be of general interest, in this paper I use it only as a stepping stone for proving Theorem 1.1.

DEFINITIONS. Fix a finite $q \in \mathbb{N}$, and let $\beta_0(L) = I_n + \sum_{\ell=1}^{q} \beta_{0,\ell} L^\ell$ and Σ_0 denote the MA lag polynomial and prediction covariance matrix, respectively, in the Wold decomposition (Hannan, 1970, Thm. $2''$, p. 158) of a q-dependent stationary n-dimensional process with ACF given by $\{\Gamma_0(k)\}_{0\le k\le q} \in \mathbb{T}_{n,q}$, i.e., the true ACF out to lag q. That is, $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,q}) \in \mathbb{B}_{n,q}$ and $\Sigma_0 \in \mathbb{S}_n$ are the unique parameters such that $\Gamma_0(k) = \sum_{\ell=0}^{q-k} \beta_{0,\ell+k} \Sigma_0 \beta_{0,\ell}'$ for $0 \le k \le q$, where $\beta_{0,0} = I_n$. Here $\mathbb{S}_n$ denotes the space of symmetric positive definite $n \times n$ matrices, while $\mathbb{B}_{n,q}$ is the space of coefficients for which the MA lag polynomial $\beta_0(L)$ has all its roots outside the unit circle, i.e.,

$$\mathbb{B}_{n,q} = \left\{ \beta = (\beta_1, \dots, \beta_q) \in \mathbb{R}^{n \times nq} \colon \det(\Phi(z;\beta)) \neq 0 \ \forall\ |z| \le 1 \right\},$$

$$\Phi(z;\beta) = I_n + \sum_{\ell=1}^{q} \beta_\ell z^\ell, \quad z \in \mathbb{C}.$$

Define the MA spectral density parametrized in terms of (β, Σ):

$$\tilde{f}(\omega;\beta,\Sigma) = \frac{1}{2\pi} \Phi(e^{-\imath\omega};\beta) \Sigma \Phi(e^{\imath\omega};\beta)', \quad \omega \in [-\pi,\pi], \quad (\beta,\Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n.$$

Using the notation introduced in Appendix A.1.7.2 for the periodogram $\hat{I}(\omega)$, the Whittle MA log likelihood is given by

$$\begin{aligned} \log p^W_{Y|\beta,\Sigma}(Y_T \mid \beta, \Sigma) = -nT\log(2\pi) &- \frac{T}{4\pi} \int_{-\pi}^{\pi} \log\det(\tilde{f}(\omega;\beta,\Sigma))\, d\omega \\ &- \frac{T}{4\pi} \int_{-\pi}^{\pi} \operatorname{tr}\{\tilde{f}(\omega;\beta,\Sigma)^{-1} \hat{I}(\omega)\}\, d\omega. \end{aligned}$$

As shown in the proof, the result below goes through if the integrals in the Whittle likelihood are replaced with discretized sums (cf. Appendix A.1.7.2).

RESULT. I now state the posterior consistency result for (β_0, Σ_0). Let $\Pi_{\beta,\Sigma}(\cdot)$ be a prior measure for (β_0, Σ_0) on $\mathbb{B}_{n,q} \times \mathbb{S}_n$. Define the Whittle posterior measure

$$P^W_{\beta,\Sigma|Y}(\mathcal{A} \mid Y_T) = \frac{\int_{\mathcal{A}} p^W_{Y|\beta,\Sigma}(Y_T \mid \beta, \Sigma)\Pi_{\beta,\Sigma}(d\beta, d\Sigma)}{\int_{\mathbb{B}_{n,q}\times\mathbb{S}_n} p^W_{Y|\beta,\Sigma}(Y_T \mid \beta, \Sigma)\Pi_{\beta,\Sigma}(d\beta, d\Sigma)}$$

for any measurable set $\mathcal{A} \subset \mathbb{B}_{n,q} \times \mathbb{S}_n$. Note that the lemma below does not require the true data distribution to be Gaussian or q-dependent.

Lemma A.3. *Let Assumption 1.3 hold. Assume that the pseudo-true parameters $(\beta_0, \Sigma_0) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$ are in the support of the prior $\Pi_{\beta,\Sigma}(\cdot)$. Then the Whittle posterior for (β_0, Σ_0) is consistent, i.e., for any neighborhood $\tilde{\mathcal{U}}$ of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$,*

$$P^W_{\beta,\Sigma|Y}(\tilde{\mathcal{U}} \mid Y_T) \xrightarrow{p} 1.$$

A.2 Material for Chapter 2

A.2.1 Detailed calculations for the case $\Lambda_0 = 0$

Using the definitions of $\hat{F}^k$ and H^k, we get

$$\begin{aligned} T^{-1}\sum_{t=1}^{T} \|\hat{F}^k_t - H^{k\prime}F_t\|^2 &= \operatorname{tr}\left\{(\hat{F}^k - FH^k)(\hat{F}^k - FH^k)'\right\} \\ &= N^{-2}T^{-3}\operatorname{tr}\left\{\tilde{F}^{k\prime}(XX' - F\Lambda_0'\Lambda_0F')(XX' - F\Lambda_0'\Lambda_0F')'\tilde{F}^k\right\}. \end{aligned}$$

Let $\Lambda_0 = 0$. By definition, $\tilde{F}^k$ equals $\sqrt{T}$ times the $T \times k$ matrix whose columns are the eigenvectors of XX' corresponding to its k largest eigenvalues. That is, if we write

$(XX')R = RC$, where R is the orthogonal matrix of eigenvectors and C the diagonal matrix of eigenvalues (in descending order), we have $\sqrt{T}R = (\tilde{F}^k, \breve{F}^k)$ for a $T \times (T-k)$ matrix $\breve{F}^k$ that satisfies $\tilde{F}^{k\prime}\breve{F}^k = 0$. Observe that

$$\tilde{F}^k = \sqrt{T}R(I_k, 0_{k\times(T-k)})',$$

so

$$(XX')\tilde{F}^k = \sqrt{T}(XX')R(I_k, 0_{k\times(T-k)})' = \sqrt{T}RC(I_k, 0_{k\times(T-k)})',$$

and

$$\begin{aligned}\tilde{F}^{k\prime}(XX')(XX')\tilde{F}^k &= T(I_k, 0_{k\times(T-k)})CR'RC(I_k, 0_{k\times(T-k)})' \\ &= TC_k^2,\end{aligned}$$

where $C_k = (I_k, 0_{k\times(T-k)})C(I_k, 0_{k\times(T-k)})'$ denotes the diagonal matrix containing the k largest eigenvalues $\omega_1, \ldots, \omega_k$ of XX'. Hence,

$$\begin{aligned}T^{-1}\sum_{t=1}^{T}\|\hat{F}_t^k - H^{k\prime}F_t\|^2 &= (NT)^{-2}\mathrm{tr}\{C_k^2\} \\ &= (NT)^{-2}\sum_{l=1}^{k}\omega_l^2. \qquad \text{(A.5)}\end{aligned}$$

Example 1 (white noise, continued). Under the assumptions in the main text, the $T \times N$ data matrix X has elements $x_{it} = e_{it} + h_{NT}\xi_{it}$ that are i.i.d. across i and t with mean 0 and variance $\Omega_{NT} = \sigma_e^2 + h_{NT}^2\sigma_\xi^2$. Let Z be a $T \times N$ matrix with elements $z_{it} = x_{it}/\sqrt{\Omega_{NT}}$. Then z_{it} is i.i.d. across i and t with mean zero and unit variance. Let $\tilde{\omega}_1$ denote the largest eigenvalue of the sample covariance matrix $N^{-1}ZZ'$. By Theorem 5.8 of Bai & Silverstein (2009), $\tilde{\omega}_1 \xrightarrow{a.s.} (1+\sqrt{\theta})^2$. Because the largest eigenvalue of $N^{-1}XX'$ satisfies $\omega_1 = \Omega_{NT}\tilde{\omega}_1$, the result (2.7) follows from (A.5).

EXAMPLE 2 (RANDOM WALK, CONTINUED). Let 1_T denote the T-vector of ones. Setting $k = 1$ in equation (A.5), we obtain

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}\|\hat{F}_t^k - H^{k\prime}F_t\|^2 &= (NT)^{-2}\left(\max_{v\in\mathbb{R}^T}\frac{v'XX'v}{v'v}\right)^2 \\ &\geq (NT)^{-2}\left(\frac{1_T'XX'1_T}{T}\right)^2 \\ &= \frac{1}{N^2T^4}\left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T}x_{it}\right)^2\right]^2. \end{aligned}$$

Jensen's inequality and cross-sectional i.i.d.-ness of x_{it} imply

$$E\left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T}x_{it}\right)^2\right]^2 \geq \left[NE\left(\sum_{t=1}^{T}x_{it}\right)^2\right]^2,$$

so inequality (2.8) follows.

EXAMPLE 3 (SINGLE LARGE BREAK, CONTINUED). In the large break model, $w_{it} = \Delta_i F_t \mathbf{1}_{\{t\geq\kappa+1\}}$. Denote the last $(T-\kappa)$ elements of the T-vector F by $F_{\kappa+1:T}$. Then we can write

$$w = (w_1,\ldots,w_T)' = h_{NT}\begin{pmatrix} 0_{\kappa\times N} \\ F_{\kappa+1:T}\otimes\Delta' \end{pmatrix},$$

so that

$$ww' = h_{NT}^2\begin{pmatrix} 0_{\kappa\times\kappa} & 0_{\kappa\times(T-\kappa)} \\ 0_{(T-\kappa)\times\kappa} & (F_{\kappa+1:T}F_{\kappa+1:T}')\|\Delta\|^2 \end{pmatrix}.$$

It follows that the eigenvalues of ww' are 0 (with multiplicity κ) along with $h_{NT}^2\|\Delta\|^2$ times the $(T-\kappa)$ eigenvalues of $F_{\kappa+1:T}F_{\kappa+1:T}'$. But the eigenvalues of $F_{\kappa+1:T}F_{\kappa+1:T}'$ are just $\|F_{\kappa+1:T}\|^2$ (with multiplicity 1) and 0 (with multiplicity $T-\kappa-1$). The k largest eigenvalues $\omega_1,\ldots,\omega_k$

of $XX' = ww'$ are therefore

$$\omega_1 = h_{NT}^2 \|\Delta\|^2 \|F_{\kappa+1:T}\|^2, \quad \omega_2 = \omega_3 = \cdots = \omega_k = 0.$$

Consequently, regardless of the number of estimated factors k,

$$\begin{aligned} T^{-1}\sum_{t=1}^{T} \|\hat{F}_t^k - H^{k\prime} F_t\|^2 &= (NT)^{-2} \sum_{l=1}^{k} \omega_l^2 \\ &= \frac{h_{NT}^4}{(NT)^2} \|\Delta\|^4 \|F_{\kappa+1:T}\|^4 \\ &= \frac{h_{NT}^4}{N^2} \|\Delta\|^4 (1-\bar{\tau})^2 \left(\frac{1}{T-\kappa} \sum_{t=\kappa+1}^{T} F_t^2 + o_p(1) \right)^2 \end{aligned}$$

where the last equality uses $T - \kappa = (1 - \bar{\tau})T(1 + o(1))$. Expression (2.9) follows.

A.2.2 Comparison of our Monte Carlo calibration with Eickmeier, Lemke & Marcellino (2015)

Eickmeier et al. (2015) use a two-step maximum likelihood procedure to estimate a five-factor DFM with time-varying parameters on quarterly U.S. data from 1972 to 2007. As in some of our simulations, the factor loadings in their model evolve as independent random walks. From their smoothed estimates of the factor loading paths (restricting attention to the paths that exhibit non-negligible time variation) one obtains a median standard deviation of the innovations equal to 0.0165 for loadings on the first factor, which has the largest median loading innovation standard deviation of the five factors. Because their sample size is $T = 140$, the random walk specification implies a median standard deviation of $\lambda_{iT1} - \lambda_{i01}$ of about 0.20; the 95th percentile of the implied standard deviation of $\lambda_{iT1} - \lambda_{i01}$ is about

0.75. The 5–95 percentile range of estimated initial factor loadings is $[-0.87, 0.28]$.[8] As explained in the main text, in our random walk design with $a = \beta = \rho = 0$, $c = 2$ and $T = 200$, the standard deviation of $\lambda_{iT1} - \lambda_{i01}$ is 0.53 for all i, while the 5–95 percentile range for initial factor loadings is $[-0.74, 0.74]$. Our $c = 2$ calibration is therefore similar to the Eickmeier et al. (2015) estimated amount of factor loading time variation in U.S. data, while our $c = 3.5$ simulations appear to exhibit substantially more instability.

A.3 Material for Chapter 3

A.3.1 Notation

For $x \in \mathbb{R}$, define $x_+ = \max\{x, 0\}$. I_n is the $n \times n$ identity matrix. Denote the space of symmetric positive semidefinite $n \times n$ matrices by $\mathbb{S}_n$ and the subspace of positive definite matrices by $\mathbb{S}_+^n$. For $A \in \mathbb{S}_n$, let $\rho(A)$ be its largest eigenvalue. For $A \in \mathbb{S}_+^n$ and $B \in \mathbb{R}^{m \times n}$, define the weighted Frobenius norm $\|B\|_A = \sqrt{\text{tr}(B'AB)}$, the usual Frobenius norm $\|B\| = \|B\|_{I_n}$, and the max norm $\|B\|_\infty = \max_{k,\ell} |B_{k\ell}|$. For $B \in \mathbb{R}^{m \times n}$, denote the rank of B by $\text{rk}(B)$, and let $\text{span}(B)$ be the linear space spanned by the columns of B. Denote the trace of a square matrix C by $\text{tr}(C)$. The $1-\alpha$ quantile of the $\chi^2(1)$ distribution is denoted $z_{1,1-\alpha}$.

A.3.2 Data and empirical specification

The data used in Sections 3.1, 3.2 and 3.5 is from the replication files for Gertler & Karadi (2015), which are available on the American Economic Association Website.[9] The specification for the non-smooth Jordà local projection IRF estimate follows Ramey (2016, Sec. 3.5.3). The response variable is the Gilchrist & Zakrajšek (2012) excess bond premium. The

[8] We are grateful to Wolfgang Lemke and Massimiliano Marcellino for helping us obtain these figures.

[9] https://www.aeaweb.org/articles?id=10.1257/mac.20130329. Downloaded April 11, 2016.

shock variable is the Gertler & Karadi (2015) monetary policy shock identified from high-frequency changes in 3-month-ahead Federal Funds Futures prices around Federal Open Market Committee announcements. The regressions control for two lags of the response and shock variables, as well as two lags of the following: log industrial production, log consumer price index, and the interest rate on 1-year Treasury bills. Unlike Ramey, I additionally control for a quadratic time trend. The regression sample is January 1991 through June 2012, but data points from late 1990 are used by lagged series.

In IRF plots, the shock is normalized to have unit standard deviation, corresponding to 4.9 basis points. To interpret the units, a monthly regression without controls of the first difference of the Effective Federal Funds Rate (in basis points) on a 1-standard-deviation monetary policy shock yields a coefficient of 11 on the 1991–2012 sample.

A.3.3 Details for projection shrinkage estimator

Here I provide analytic derivations and quantile simulation strategies for the projection shrinkage estimators. Let P be a symmetric and idempotent matrix. Then $\|\beta - \hat{\beta}\|^2 = \|P(\beta - \hat{\beta})\|^2 + \|(I_n - P)(\beta - \hat{\beta})\|^2$, implying that $\hat{\beta}_P(\lambda)$ defined in (3.6) satisfies

$$P\hat{\beta}_P(\lambda) = \frac{1}{1+\lambda} P\hat{\beta}, \quad (I_n - P)\hat{\beta}_P(\lambda) = (I_n - P)\hat{\beta}.$$

URE. The URE simplifies in the case of projection shrinkage. The matrix $\Theta_P(\lambda) := \Theta_{P,I_n}(\lambda) = (I_n + \lambda P)^{-1}$ satisfies $P\Theta_P(\lambda) = (1+\lambda)^{-1}P$ and $(I_n - P)\Theta_P(\lambda) = I_n - P$. Hence, for $M = P$ and $W = \tilde{W} = I_n$, the URE (3.7) can be written

$$\begin{aligned}\hat{R}_{P,I_n,I_n}(\lambda) = T\|P(\hat{\beta}_P(\lambda) - \hat{\beta})\|^2 + T\|(I_n - P)(\hat{\beta}_P(\lambda) - \hat{\beta})\|^2 \\ + 2\operatorname{tr}\{P\Theta_P(\lambda)\hat{\Sigma}\} + 2\operatorname{tr}\{(I_n - P)\Theta_P(\lambda)\hat{\Sigma}\}\end{aligned}$$

$$= \left(\frac{\lambda}{1+\lambda}\right)^2 T\|P\hat{\beta}\|^2 + \left(1 - \frac{\lambda}{1+\lambda}\right)\mathrm{tr}(\hat{\Sigma}_P) + \text{constant},$$

where $\hat{\Sigma}_P = P\hat{\Sigma}P$. The value of $\lambda \geq 0$ that minimizes the above quadratic form satisfies

$$\frac{\hat{\lambda}_P}{1+\hat{\lambda}_P} = \min\left\{\frac{\mathrm{tr}(\hat{\Sigma}_P)}{T\|P\hat{\beta}\|^2}, 1\right\}.$$

Quantile simulation. For projection shrinkage, the simulation of the quantile functions in Section 3.2.4 simplifies drastically, since $\hat{\theta}_{M,W,\tilde{W}}(\eta,\Sigma)$ defined in (3.9) reduces to

$$\hat{\theta}_{P,I_n,I_n}(\eta,\Sigma) = \eta - \min\left\{\frac{\mathrm{tr}(P\Sigma)}{\|P\eta\|^2}, 1\right\} P\eta.$$

Conditional quantile bound. In the notation of Section 3.2.4, $q_{s,1-\alpha,P,I_n,I_n}(\theta,\Sigma)$ is the $1-\alpha$ quantile of $\{s'\hat{\theta}_{P,I_n,I_n}(\zeta u+\theta,\Sigma) - s'\theta\}^2$, $u \sim N(0, s'\Sigma s)$. By an argument in the proof of Proposition 3.2, $\|\hat{\theta}_{P,I_n,I_n}(\eta,\Sigma) - \eta\| \leq \sqrt{\mathrm{tr}(P\Sigma)}$ for all η, Σ. Using $s'\zeta = 1$, it follows that

$$|s'\hat{\theta}_{P,I_n,I_n}(\zeta u+\theta,\Sigma) - s'\theta| \leq |s'(\zeta u+\theta) - s'\theta| + \|s\|\sqrt{\mathrm{tr}(P\Sigma)} = |u| + \|s\|\sqrt{\mathrm{tr}(P\Sigma)},$$

implying $\sqrt{q_{s,1-\alpha,P,I_n,I_n}(\theta,\Sigma)} \leq \sqrt{(s'\Sigma s)z_{1,1-\alpha}} + \|s\|\sqrt{\mathrm{tr}(P\Sigma)}$ for all θ, Σ.

A.3.4 URE and the bias-variance tradeoff

The URE (3.7) can also be motivated from the bias-variance perspective used by Claeskens & Hjort (2003) to derive their Focused Information Criterion. Informally, suppose that $E(\hat{\beta}) = \beta^\dagger$ and $E[(\hat{\beta}-\beta^\dagger)(\hat{\beta}-\beta^\dagger)'] = T^{-1}\Sigma$, and set $W = \tilde{W} = I_n$ to simplify notation. For given $\lambda \geq 0$, the MSE of $\hat{\beta}_{M,I_n}(\lambda)$ can be decomposed into bias and variance terms:

$$\begin{aligned} R_{M,I_n,I_n}(\lambda) = {} & TE[\hat{\beta}_{M,I_n}(\lambda) - \beta^\dagger]'E[\hat{\beta}_{M,I_n}(\lambda) - \beta^\dagger] \\ & + \mathrm{tr}\left\{TE[(\hat{\beta}_{M,I_n}(\lambda) - E[\hat{\beta}_{M,I_n}(\lambda)])(\hat{\beta}_{M,I_n}(\lambda) - E[\hat{\beta}_{M,I_n}(\lambda)])']\right\} \end{aligned}$$

$$= T \operatorname{tr}\left\{[I_n - \Theta_{M,I_n}(\lambda)]^2 \beta^\dagger \beta^{\dagger\prime}\right\} + \operatorname{tr}\left\{\Theta_{M,I_n}(\lambda)^2 \Sigma\right\}.$$

Since $E(\hat{\beta}\hat{\beta}') = \beta^\dagger\beta^{\dagger\prime} + T^{-1}\Sigma$, consider the estimator of $R_{M,I_n,I_n}(\lambda)$ obtained by substituting in the unbiased estimator $\hat{\beta}\hat{\beta}' - T^{-1}\hat{\Sigma}$ of $\beta^\dagger\beta^{\dagger\prime}$, and substituting $\hat{\Sigma}$ for Σ:

$$\begin{aligned}
\tilde{R}_M(\lambda) &:= \operatorname{tr}\left\{[I_n - \Theta_{M,I_n}(\lambda)]^2 (T\hat{\beta}\hat{\beta}' - \hat{\Sigma})\right\} + \operatorname{tr}\left\{\Theta_{M,I_n}(\lambda)^2 \hat{\Sigma}\right\} \\
&= T\|\hat{\beta}_{M,I_n}(\lambda) - \hat{\beta}\|^2 + \operatorname{tr}\left\{\left(\Theta_{M,I_n}(\lambda)^2 - [I_n - \Theta_{M,I_n}(\lambda)]^2\right)\hat{\Sigma}\right\} \\
&= T\|\hat{\beta}_{M,I_n}(\lambda) - \hat{\beta}\|^2 + \operatorname{tr}\left\{(2\Theta_{M,I_n}(\lambda) - I_n)\hat{\Sigma}\right\} \\
&= \hat{R}_{M,I_n,I_n}(\lambda) - \operatorname{tr}(\hat{\Sigma}).
\end{aligned}$$

Hence, the criterion $\tilde{R}_M(\lambda)$ is equivalent with the URE criterion $\hat{R}_{M,I_n,I_n}(\lambda)$ for the purposes of selecting the shrinkage parameter λ.

A.3.5 Supplemental simulation results

Supplementing the analysis in Section 3.5, I provide simulation results for the SmIRF estimator and plot the true IRF in the VAR DGP.

Table A.1 shows that the MSE performance of the SmIRF estimator is similar to that of the quadratic projection shrinkage estimator. The SmIRF estimator is given by $\hat{\theta}_{M,I_n,I_n}$, as defined in Section 3.4.1, where M is the $(n-2)\times n$ second difference matrix (3.5). Quadratic projection shrinkage tends to do slightly better than SmIRF when $K = 0$ or 0.5, but SmIRF does better for $K = 1$ or 2, i.e., when the true IRF is more jagged. This is unsurprising, as the second difference penalty in the SmIRF objective function (3.2) is more lenient toward a sine curve than is the quadratic projection penalty in (3.6).

Table A.2 illustrates the performance of the SmIRF-based confidence sets for three of the DGPs considered in Table 3.1 in Section 3.5.1. The results in this table are based on a smaller number of simulations than other results in this paper due to the computational cost

Simulation results: SmIRF MSE

Parameters					Joint		Marginal	
					MSE	Var	MSE	
n	K	κ	σ_0	φ			Imp	Mid
10	0.5	0.5	0.2	3	0.72	0.67	1.93	0.66
25	0.5	0.5	0.2	3	0.43	0.41	1.66	0.39
50	0.5	0.5	0.2	3	0.26	0.24	1.02	0.22
25	0	0.5	0.2	3	0.34	0.34	0.82	0.26
25	1	0.5	0.2	3	0.51	0.47	1.81	0.39
25	2	0.5	0.2	3	0.65	0.61	1.70	0.54
25	0.5	0	0.2	3	0.22	0.21	0.71	0.18
25	0.5	0.9	0.2	3	0.87	0.84	1.95	0.88
25	0.5	0.5	0.1	3	0.43	0.41	1.50	0.39
25	0.5	0.5	0.4	3	0.38	0.36	1.38	0.33
25	0.5	0.5	0.2	1	0.41	0.39	0.83	0.35
25	0.5	0.5	0.2	5	0.42	0.40	2.88	0.38

Table A.1: Simulation results for MSE of SmIRF estimator. See caption for Table 3.1. 5000 simulations per DGP. Numerical optimization: Matlab's `fmincon`, algorithm "interior-point".

Simulation results: SmIRF confidence sets

Parameters					Joint			Marginal			
					MSE	Var	CV	MSE		Lng $\hat{\mathcal{C}}_{s,1-\delta}$	
n	K	κ	σ_0	φ				Imp	Mid	Imp	Mid
25	0	0.5	0.2	3	0.32	0.32	0.56	0.84	0.23	0.85	0.81
25	0.5	0.5	0.2	3	0.43	0.41	0.84	1.59	0.37	0.95	0.82
25	1	0.5	0.2	3	0.54	0.49	0.91	2.32	0.39	1.08	0.82

Table A.2: Simulation results for SmIRF confidence sets. See caption for Table 3.1. 1000 simulations per DGP, 500 simulations to compute quantiles, 30 grid points to compute marginal confidence set length, $\alpha = 0.1$, $\delta = 0.01$. Numerical optimization: Matlab's `fmincon`, algorithm "active-set".

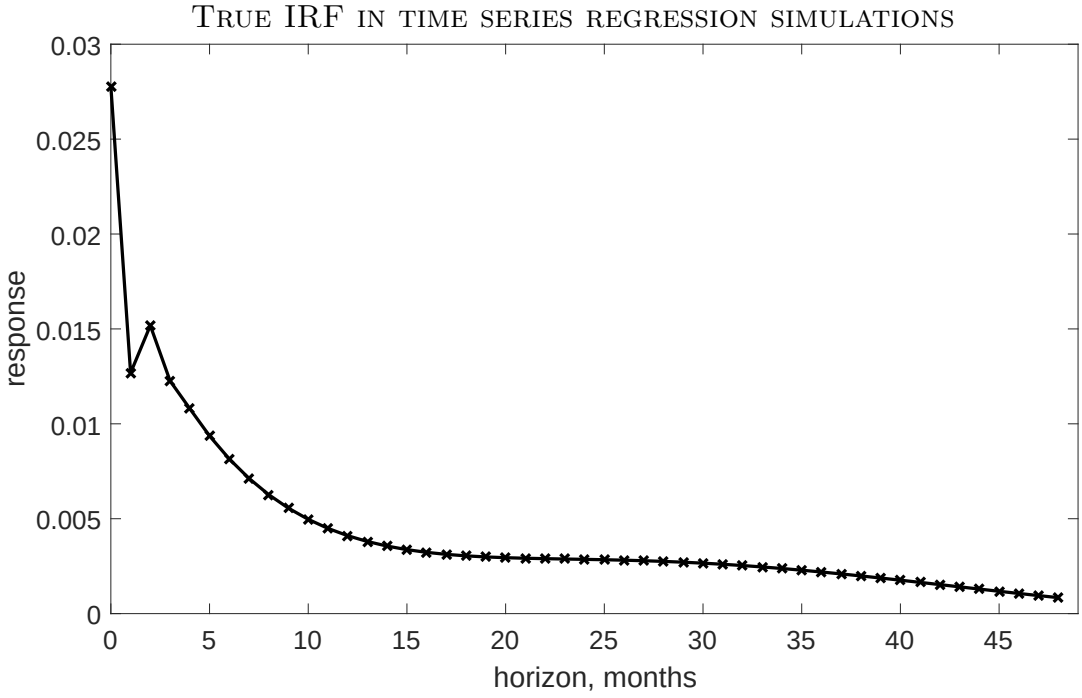

Figure A.14: True VAR-implied IRF in time series regression simulations.

of computing the URE-minimizing shrinkage parameter for the SmIRF estimator. The table shows that the SmIRF confidence sets do as well as or better than the quadratic projection shrinkage confidence sets.

Figure A.14 shows the true IRF implied by the data generating VAR(2) model used in the simulations in Section 3.5.2.

Appendix B

Proofs

B.1 Proofs for Chapter 1

B.1.1 Proof of Theorem A.1

As in Lippi & Reichlin (1994, p. 311), define the rational matrix function

$$R(\gamma, z) = \begin{pmatrix} \frac{z-\gamma}{1-\bar{\gamma}z} & 0 \\ 0 & I_{n-1} \end{pmatrix}, \quad \gamma, z \in \mathbb{C}.$$

Transformation (ii) corresponds to the transformation $\check{\Psi}(z) = \Psi(z)QR(\gamma_k, z)^{-1}$ if γ_k is real. If γ_k is not real, the transformation corresponds to $\check{\Psi}(z) = \tilde{\Psi}(z)\check{Q}$, where $\tilde{\Psi}(z) = \Psi(z)QR(\gamma_k, z)^{-1}\tilde{Q}R(\overline{\gamma_k}, z)^{-1}$ and $\check{Q} = \tilde{\Psi}(0)^{-1}J$ is a unitary matrix. I proceed in three steps.

Step 1. Consider the first claim of the theorem. Let $f(\omega; \Gamma) = (2\pi)^{-1}\sum_{k=-q}^{q}\Gamma(k)e^{-\imath k\omega}$, $\omega \in [-\pi, \pi]$, denote the spectral density matrix function associated with the ACF $\Gamma(\cdot)$. Since $\Psi(z) = \Theta(z)\,\mathrm{diag}(\sigma)$ with $(\Theta, \sigma) \in \mathcal{S}(\Gamma)$, we must have $\Psi(e^{-\imath\omega})\Psi(e^{-\imath\omega})^* = 2\pi f(\omega; \Gamma)$ for all ω by the usual formula for the spectral density of a vector MA process (Brockwell & Davis, 1991, Example 11.8.1). Because $R(\gamma, e^{-\imath\omega})R(\gamma, e^{-\imath\omega})^* = I_n$ for any (γ, ω), it is easy to verify that $\check{\Psi}(z)$ – constructed by applying transformation (i) or transformation (ii) to $\Psi(z)$ – also

satisfies $\check{\Psi}(e^{-\imath\omega})\check{\Psi}(e^{-\imath\omega})^* = 2\pi f(\omega;\Gamma)$. Hence, $\check{\Psi}(z) = \sum_{\ell=0}^{q} \check{\Psi}_\ell z^\ell$ is a matrix MA polynomial satisfying $\sum_{\ell=0}^{q-k} \check{\Psi}_{\ell+k}\check{\Psi}_\ell^{\ *} = \Gamma(k)$ for all $k = 0, 1, \ldots, q$. In Step 2 below I show that $\check{\Psi}(z)$ is a matrix polynomial with real coefficients. By construction of $\check{\Theta}(z) = \sum_{\ell=0}^{q} \check{\Theta}_\ell z^\ell$ and $\check{\sigma}$, we then have $\sum_{\ell=0}^{q-k} \check{\Theta}_{\ell+k}\,\mathrm{diag}(\check{\sigma})^2\check{\Theta}_\ell{}' = \Gamma(k)$ for all $k = 0, 1, \ldots, q$, so $(\check{\Theta}, \check{\sigma}) \in \mathcal{S}(\Gamma)$, as claimed.

Step 2. I now show that transformation (ii) yields a real matrix polynomial $\check{\Psi}(z)$. This fact was asserted by Lippi & Reichlin (1994, pp. 317–318). I am grateful to Professor Marco Lippi for providing me with the proof arguments for Step 2; all errors are my own.

$\check{\Psi}(z)$ is clearly real if the flipped root γ_k is real (since η and Q can be chosen to be real in this case), so consider the case where we flip a pair of complex conjugate roots γ_k and $\overline{\gamma_k}$. Recall that in this case, $\check{\Psi}(z) = \tilde{\Psi}(z)\check{Q}$, where $\tilde{\Psi}(z) = \Psi(z)QR(\gamma_k, z)^{-1}\tilde{Q}R(\overline{\gamma_k}, z)^{-1}$ and $\check{Q}$ is unitary. It follows from the same arguments as in Step 1 that the complex-valued matrix polynomial $\tilde{\Psi}(z) = \sum_{\ell=0}^{q} \tilde{\Psi}_\ell z^\ell$ satisfies $\sum_{\ell=0}^{q-k} \tilde{\Psi}_{\ell+k}\tilde{\Psi}_\ell^{\ *} = \Gamma(k)$ for all $k = 0, 1, \ldots, q$

Let $\bar{\tilde{\Psi}}(z) = \sum_{\ell=0}^{q} \overline{\tilde{\Psi}_\ell} z^\ell$ denote the matrix polynomial obtained by conjugating the coefficients of the polynomial $\tilde{\Psi}(z)$. By construction, the roots of $\det(\tilde{\Psi}(z))$ are real or appear as complex conjugate pairs, so $\det(\bar{\tilde{\Psi}}(z))$ has the same roots as $\det(\tilde{\Psi}(z))$. Furthermore, for $k = 0, 1, \ldots, q$,

$$\sum_{\ell=0}^{q-k} \bar{\tilde{\Psi}}_{\ell+k}\bar{\tilde{\Psi}}_\ell^{\ *} = \overline{\Gamma(k)} = \Gamma(k) = \sum_{\ell=0}^{q-k} \tilde{\Psi}_{\ell+k}\tilde{\Psi}_\ell^*.$$

By Theorem 3(b) of Lippi & Reichlin (1994), there exists a unitary $n \times n$ matrix $\tilde{\tilde{Q}}$ such that $\bar{\tilde{\Psi}}(z) = \tilde{\Psi}(z)\tilde{\tilde{Q}}$ for $z \in \mathbb{R}$. The matrix polynomial $\tilde{\Psi}(z)\tilde{\Psi}(0)^{-1}$ then has real coefficients:[1] For all $z \in \mathbb{R}$,

$$\tilde{\Psi}(z)\tilde{\Psi}(0)^{-1} = \left(\tilde{\Psi}(z)\tilde{\tilde{Q}}\right)\left(\tilde{\Psi}(0)\tilde{\tilde{Q}}\right)^{-1} = \bar{\tilde{\Psi}}(z)\bar{\tilde{\Psi}}(0)^{-1} = \overline{\tilde{\Psi}(z)\tilde{\Psi}(0)^{-1}}.$$

[1] $\tilde{\Psi}(0)$ is nonsingular because $\det(\Psi(0)) \neq 0$.

Consequently, with the real matrix J defined as in the theorem, $\check{\Psi}(z) = \tilde{\Psi}(z)\tilde{\Psi}(0)^{-1}J$ is a matrix polynomial with real coefficients. Finally, since $\tilde{\tilde{Q}}$ is unitary, the matrix

$$\tilde{\Psi}(0)\tilde{\Psi}(0)^* = \left(\bar{\tilde{\Psi}}(0)\tilde{\tilde{Q}}\right)\left(\bar{\tilde{\Psi}}(0)\tilde{\tilde{Q}}\right)^* = \overline{\tilde{\Psi}(0)\tilde{\Psi}(0)^*}$$

is real, symmetric, and positive definite, so J is well-defined.

Step 3. Finally, I prove the second claim of the theorem. Suppose we have a fixed element $(\check{\Theta}, \check{\sigma})$ of the identified set that we want to end up with after transforming the initial element (Θ, σ) appropriately. Define $\check{\Psi}(z) = \check{\Theta}(z)\,\mathrm{diag}(\check{\sigma})$. Since $(\Theta, \sigma), (\check{\Theta}, \check{\sigma}) \in \mathcal{S}(\Gamma)$, the two sets of SVMA parameters correspond to the same spectral density, i.e., $\Psi(e^{-\imath\omega})\Psi(e^{-\imath\omega})^* = \check{\Psi}(e^{-\imath\omega})\check{\Psi}(e^{-\imath\omega})^*$ for all $\omega \in [-\pi, \pi]$. As in the proof of Theorem 2 in Lippi & Reichlin (1994), we can apply transformation (ii) finitely many (say, b) times to $\Psi(z)$, flipping all the roots that are inside the unit circle, thus ending up with a polynomial

$$B(z) = \Psi(z)Q_1R(\gamma_{k_1}, z)^{-1}\cdots Q_bR(\gamma_{k_b}, z)^{-1}Q_{b+1}$$

for which all roots of $\det(B(z))$ lie on or outside the unit circle. Likewise, denote the (finitely many) roots of $\det(\check{\Psi}(z))$ by $\check{\gamma}_k$, $k = 1, 2, \ldots$, and apply to $\check{\Psi}(z)$ a finite sequence of transformation (ii) to arrive at a polynomial

$$\check{B}(z) = \check{\Psi}(z)\check{Q}_1R(\check{\gamma}_{\check{k}_1}, z)^{-1}\cdots \check{Q}_{\check{b}}R(\check{\gamma}_{\check{k}_{\check{b}}}, z)^{-1}\check{Q}_{\check{b}+1}$$

for which all roots of $\det(\check{B}(z))$ lie on or outside the unit circle. Since $\det(B(z))$ and $\det(\check{B}(z))$ have all roots on or outside the unit circle, and we have $B(e^{-\imath\omega})B(e^{-\imath\omega})^* = \check{B}(e^{-\imath\omega})\check{B}(e^{-\imath\omega})^* = 2\pi f(\omega; \Gamma)$ for all ω, there must exist an orthogonal matrix Q such that

$\check{B}(z) = B(z)Q$ (Lippi & Reichlin, 1994, p. 313; Hannan, 1970, p. 69). Thus,

$$\check{\Psi}(z) = \Psi(z)Q_1R(\gamma_{k_1}, z)^{-1} \cdots Q_bR(\gamma_{k_b}, z)^{-1}Q_{b+1}Q\check{Q}^*_{\check{b}+1}R(\check{\gamma}_{\check{k}_{\check{b}}}, z)\check{Q}^*_{\check{b}} \cdots R(\check{\gamma}_{\check{k}_1}, z)\check{Q}^*_1,$$

and

$$\det(\check{\Psi}(z)) = \det(\Psi(z))\frac{(z - \check{\gamma}_{\check{k}_1}) \cdots (z - \check{\gamma}_{\check{k}_{\check{b}}})(1 - \overline{\gamma_{k_1}}z) \cdots (1 - \overline{\gamma_{k_b}}z)}{(z - \gamma_{k_1}) \cdots (z - \gamma_{k_b})(1 - \overline{\check{\gamma}_{\check{k}_1}}z) \cdots (1 - \overline{\check{\gamma}_{\check{k}_{\check{b}}}}z)},$$

so any root $\check{\gamma}_{\check{k}}$ of $\det(\check{\Psi}(z))$ must either equal γ_k or it must equal $1/\overline{\gamma_k}$, where γ_k is some root of $\det(\Psi(z))$. It follows that we can apply a finite sequence of transformation (ii) (i.e., an appropriate sequence of root flips) to $\Psi(z)$ to obtain a real matrix polynomial $\tilde{\check{\Psi}}(z)$ satisfying $\det(\tilde{\check{\Psi}}(z)) = \det(\check{\Psi}(z))$ for all $z \in \mathbb{C}$. Theorem 3(b) in Lippi & Reichlin (1994) then implies that $\check{\Psi}(z)$ can be obtained from $\tilde{\check{\Psi}}(z)$ through transformation (i) (i.e., an orthogonal rotation, which clearly must be real). Finally, obtain $(\check{\Theta}, \check{\sigma})$ from $\check{\Psi}(z)$ by transformation (a). □

B.1.2 Proof of Lemma A.1

Suppressing the arguments (Ψ), let $L_k = \log\det(f_k) + \tilde{y}_k^* f_k^{-1}\tilde{y}_k$. Then

$$\frac{\partial L_k}{\partial(f_k')} = f_k^{-1} - f_k^{-1}\tilde{y}_k\tilde{y}_k^* f_k^{-1} = C_k.$$

Writing $f_k' = \overline{\tilde{\Psi}_k}\tilde{\Psi}_k'$, we have

$$\frac{\partial \operatorname{vec}(f_k')}{\partial \operatorname{vec}(\Psi_\ell)'} = (\tilde{\Psi}_k \otimes I_n)e^{\imath\omega_k\ell} + (I_n \otimes \overline{\tilde{\Psi}_k})K_n' e^{-\imath\omega_k\ell},$$

where K_n is the $n^2 \times n^2$ commutation matrix such that $\operatorname{vec}(B') = K_n \operatorname{vec}(B)$ for any $n \times n$ matrix B (Magnus & Neudecker, 2007, Ch. 3.7). Using $\operatorname{vec}(ABC) = (C' \otimes A)\operatorname{vec}(B)$,

$$\frac{\partial L_k}{\partial \operatorname{vec}(\Psi_\ell)'} = \frac{\partial L_k}{\partial \operatorname{vec}(f_k')'}\frac{\partial \operatorname{vec}(f_k')}{\partial \operatorname{vec}(\Psi_\ell)'} = \operatorname{vec}\left(C_k\tilde{\Psi}_k e^{\imath\omega_k\ell} + \overline{C_k^*\tilde{\Psi}_k e^{\imath\omega_k\ell}}\right)'.$$

Since $C_k^* = C_k$, we get $\partial L_k / \partial \Psi_\ell = 2 \operatorname{Re}\left(C_k \tilde{\Psi}_k e^{\imath \omega_k \ell}\right)$, so

$$\begin{aligned}
\frac{\partial \log p^W_{Y|\Psi}(Y_T \mid \Psi)}{\partial \Psi_\ell} &= -\frac{1}{2} \sum_{k=0}^{T-1} \frac{\partial L_k}{\partial \Psi_\ell} \\
&= -\sum_{k=0}^{T-1} \operatorname{Re}\left(C_k \sum_{\tilde{\ell}=1}^{q+1} e^{-\imath \omega_k (\tilde{\ell}-1)} \Psi_{\tilde{\ell}-1} e^{\imath \omega_k \ell} \right) \\
&= -\sum_{\tilde{\ell}=0}^{q} \operatorname{Re}\left(\sum_{k=0}^{T-1} C_k e^{-\imath \omega_k (\tilde{\ell}-\ell)} \right) \Psi_{\tilde{\ell}}.
\end{aligned}$$

Finally, $\sum_{k=0}^{T-1} C_k e^{-\imath \omega_k (\tilde{\ell}-\ell)} = \sum_{k=0}^{T-1} C_k e^{-\imath \omega_{\tilde{\ell}-\ell} k} = \tilde{C}_{\tilde{\ell}-\ell}$ for $\tilde{\ell} \geq \ell$, and $\sum_{k=0}^{T-1} C_k e^{-\imath \omega_k (\tilde{\ell}-\ell)} = \sum_{k=0}^{T-1} C_k e^{-\imath \omega_k (T+\tilde{\ell}-\ell)} = \sum_{k=0}^{T-1} C_k e^{-\imath \omega_{T+\tilde{\ell}-\ell} k} = \tilde{C}_{\tilde{\ell}-\ell}$ for $\tilde{\ell} < \ell$. □

B.1.3 Proof of Lemma 1.1

By the triangle inequality,

$$\|P_{\theta|Y}(\cdot \mid Y_T) - \Pi_{\theta|\Gamma}(\cdot \mid \hat{\Gamma})\|_{L_1} \leq \|\Pi_{\theta|\Gamma}(\cdot \mid \hat{\Gamma}) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1} + \|P_{\theta|Y}(\cdot \mid Y_T) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1}.$$

If $\hat{\Gamma} \xrightarrow{p} \Gamma_0$, the first term above tends to 0 in probability by assumption (i) and the continuous mapping theorem. Hence, the statement of the lemma follows if I can show that the second term above tends to 0 in probability.

Let $\epsilon > 0$ be arbitrary. By assumption (i), there exists a neighborhood $\mathcal{U}$ of Γ_0 in Ξ_Γ such that $\|\Pi_{\theta|\Gamma}(\cdot \mid \Gamma) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1} < \epsilon/2$ for all $\Gamma \in \mathcal{U}$. By assumption (ii), $P_{\Gamma|Y}(\mathcal{U}^c \mid Y_T) < \epsilon/4$ w.p.a. 1. The decomposition (1.10) then implies

$$\begin{aligned}
\|P_{\theta|Y}(\cdot \mid Y_T) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1} &= \left\| \int [\Pi_{\theta|\Gamma}(\cdot \mid \Gamma) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)] P_{\Gamma|Y}(d\Gamma \mid Y_T) \right\|_{L_1} \\
&\leq \int_{\mathcal{U}} \|\Pi_{\theta|\Gamma}(\cdot \mid \Gamma) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1} P_{\Gamma|Y}(d\Gamma \mid Y_T) \\
&\quad + \int_{\mathcal{U}^c} \|\Pi_{\theta|\Gamma}(\cdot \mid \Gamma) - \Pi_{\theta|\Gamma}(\cdot \mid \Gamma_0)\|_{L_1} P_{\Gamma|Y}(d\Gamma \mid Y_T)
\end{aligned}$$

$$
\begin{aligned}
&\leq \int_{\mathcal{U}} \frac{\epsilon}{2} P_{\Gamma|Y}(d\Gamma \mid Y_T) + 2P_{\Gamma|Y}(\mathcal{U}^c \mid Y_T) \\
&\leq \frac{\epsilon}{2} + 2\frac{\epsilon}{4} \\
&= \epsilon
\end{aligned}
$$

w.p.a. 1. Here I use that the L_1 distance between probability measures is bounded by 2. □

B.1.4 Proof of Lemma A.2

I follow the proof of Theorem 1.3.4 in Ghosh & Ramamoorthi (2003). Set $\kappa_2 = \sup_{\Gamma \in \mathcal{U}^c} \phi(\Gamma)$. Notice that $\hat{\phi}(\Gamma_0) = 0$ and assumption (i) together imply $\phi(\Gamma_0) = 0$. By assumptions (ii)–(iii), we can therefore find a small neighborhood $\mathcal{V}$ of Γ_0 in Ξ_Γ such that $\kappa_1 = \inf_{\Gamma \in \mathcal{V}} \phi(\Gamma)$ satisfies $\max\{\kappa_2, \zeta\} < \kappa_1 < 0$. We may shrink $\mathcal{V}$ to ensure that it also satisfies $\mathcal{V} \subset \mathcal{U} \cap \mathcal{K}$. Choose $\delta > 0$ such that $\kappa_1 - \delta > \max\{\kappa_2 + \delta, \zeta\}$. Write

$$
\begin{aligned}
P_{\Gamma|Y}(\mathcal{U} \mid Y_T) &= \left(1 + \frac{\int_{\mathcal{U}^c} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}{\int_{\mathcal{U}} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}\right)^{-1} \\
&\geq \left(1 + \frac{\int_{\mathcal{K}^c} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma) + \int_{\mathcal{U}^c \cap \mathcal{K}} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}{\int_{\mathcal{V}} e^{T\hat{\phi}(\Gamma)} \Pi_\Gamma(d\Gamma)}\right)^{-1}.
\end{aligned}
$$

Assumptions (i) and (iv) imply that the following three inequalities hold w.p.a. 1:

$$
\sup_{\Gamma \in \mathcal{V}} \hat{\phi}(\Gamma) > \kappa_1 - \delta, \quad \sup_{\Gamma \in \mathcal{U}^c \cap \mathcal{K}} \hat{\phi}(\Gamma) < \kappa_2 + \delta, \quad \sup_{\Gamma \in \mathcal{K}^c} \hat{\phi}(\Gamma) < \zeta.
$$

We then have

$$
\begin{aligned}
P_{\Gamma|Y}(\mathcal{U} \mid Y_T) &\geq \left(1 + \frac{\int_{\mathcal{K}^c} e^{\zeta T} \Pi_\Gamma(d\Gamma) + \int_{\mathcal{U}^c \cap \mathcal{K}} e^{(\kappa_2+\delta)T} \Pi_\Gamma(d\Gamma)}{\int_{\mathcal{V}} e^{(\kappa_1-\delta)T} \Pi_\Gamma(d\Gamma)}\right)^{-1} \\
&\geq \left(1 + \frac{e^{\zeta T} + e^{(\kappa_2+\delta)T}}{\Pi_\Gamma(\mathcal{V}) e^{(\kappa_1-\delta)T}}\right)^{-1}
\end{aligned}
$$

w.p.a. 1. Since $\Pi_\Gamma(\mathcal{V}) > 0$ by assumption (v), and $\kappa_1 - \delta > \max\{\kappa_2 + \delta, \zeta\}$, I conclude that $P_{\Gamma|Y}(\mathcal{U} \mid Y_T) \xrightarrow{p} 1$ as $T \to \infty$. □

B.1.5 Proof of Theorem 1.1

The proof exploits the one-to-one mapping between the ACF Γ_0 and the Wold parameters (β_0, Σ_0) defined in Appendix A.1.7.3, which allows me to use Lemma A.3 to infer posterior consistency for Γ_0 under the Whittle likelihood.

Let $M\colon \mathbb{T}_{n,q} \to \mathbb{B}_{n,q} \times \mathbb{S}_n$ denote the function that maps a q-dependent ACF $\Gamma(\cdot)$ into its Wold representation $(\beta(\Gamma), \Sigma(\Gamma))$ (Hannan, 1970, Thm. $2''$, p. 158). By construction, the map $M(\cdot)$ is continuous (and measurable). The inverse map $M^{-1}(\cdot)$ is given by $\Gamma(k) = \sum_{\ell=0}^{q-k} \beta_{\ell+k} \Sigma \beta_\ell'$ (with $\beta_0 = I_n$) and so also continuous. The prior $\Pi_\Gamma(\cdot)$ for the ACF Γ induces a particular prior measure for the Wold parameters (β, Σ) on $\mathbb{B}_{n,q} \times \mathbb{S}_n$ given by $\Pi_{\beta,\Sigma}(\mathcal{A}) = \Pi_\Gamma(M^{-1}(\mathcal{A}))$ for any measurable set $\mathcal{A}$. Let $P^W_{\beta,\Sigma|Y}(\cdot \mid Y_T)$ be the posterior measure for (β, Σ) computed using the induced prior $\Pi_{\beta,\Sigma}(\cdot)$ and the Whittle MA likelihood $p^W_{Y|\beta,\Sigma}(Y_T \mid \beta, \Sigma)$, cf. Appendix A.1.7.3.

I first show that the induced posterior for (β_0, Σ_0) is consistent. Let $\tilde{\mathcal{U}}$ be any neighborhood of $(\beta_0, \Sigma_0) = M(\{\Gamma_0(k)\}_{0 \le k \le q})$ in $\mathbb{B}_{n,q} \times \mathbb{S}_n$. Since $M(\cdot)$ is continuous, $M^{-1}(\tilde{\mathcal{U}})$ is a neighborhood of $\{\Gamma_0(k)\}_{0 \le k \le q}$ in $\mathbb{T}_{n,q}$. Hence, since $\{\Gamma_0(k)\}_{0 \le k \le q}$ is in the support of $\Pi_\Gamma(\cdot)$, (β_0, Σ_0) is in the support of $\Pi_{\beta,\Sigma}(\cdot)$:

$$\Pi_{\beta,\Sigma}(\tilde{\mathcal{U}}) = \Pi_\Gamma(M^{-1}(\tilde{\mathcal{U}})) > 0.$$

Due to Assumption 1.3 and the fact that (β_0, Σ_0) is in the support of $\Pi_{\beta,\Sigma}(\cdot)$, Lemma A.3 implies that $P_{\beta,\Sigma|Y}(\tilde{\mathcal{U}} \mid Y_T) \xrightarrow{p} 1$ for any neighborhood $\tilde{\mathcal{U}}$ of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$.

I now prove posterior consistency for Γ_0. Since $\tilde{f}(\omega; M(\Gamma)) = f(\omega; \Gamma)$ for all $\omega \in [-\pi, \pi]$ and $\Gamma \in \mathbb{T}_{n,q}$, we have $p^W_{Y|\beta,\Sigma}(Y_T \mid M(\Gamma)) = p^W_{Y|\Gamma}(Y_T \mid \Gamma)$ for all $\Gamma \in \mathbb{T}_{n,q}$. Consequently,

$P^W_{\Gamma|Y}(\mathcal{A} \mid Y_T) = P^W_{\beta,\Sigma|Y}(M(\mathcal{A}) \mid Y_T)$ for all measurable sets $\mathcal{A}$. Let $\mathcal{U}$ be an arbitrary neighborhood of $\{\Gamma_0(k)\}_{0\leq k\leq q}$ in $\mathbb{T}_{n,q}$. Since $M^{-1}(\cdot)$ is continuous at (β_0, Σ_0), the set $\tilde{\mathcal{U}} = M(\mathcal{U})$ is a neighborhood of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$. It follows from Step 1 that

$$P^W_{\Gamma|Y}(\mathcal{U} \mid Y_T) = P^W_{\beta,\Sigma|Y}(\tilde{\mathcal{U}} \mid Y_T) \xrightarrow{p} 1.$$

Moreover, the proof of Lemma A.3 implies that the Whittle posterior is consistent regardless of whether the Whittle likelihood is based on integrals or discretized sums. □

B.1.6 Proof of Theorem 1.2

By the calculation in Eqn. 11 of Moon & Schorfheide (2012), the Whittle posterior $P^W_{\theta|Y}(\cdot \mid Y_T)$ satisfies a decomposition of the form (1.10), where the posterior measure for the ACF Γ is given by $P^W_{\Gamma|Y}(\cdot \mid Y_T)$. By Theorem 1.1, the latter posterior measure is consistent for Γ_0 provided that the induced prior $\Pi_\Gamma(\cdot)$ has Γ_0 in its support.

Γ_0 is indeed in the support of $\Pi_\Gamma(\cdot)$, for the following reason. Let $\Gamma(\Theta, \sigma)$ denote the map (1.7) from structural parameters $(\Theta, \sigma) \in \Xi_\Theta \times \Xi_\sigma$ to ACFs $\Gamma \in \mathbb{T}_{n,q}$. There exists a (non-unique) set of IRFs and shock standard deviations $(\check{\Theta}, \check{\sigma}) \in \Xi_\Theta \times X_\sigma$ such that $\Gamma_0 = \Gamma(\check{\Theta}, \check{\Sigma})$ (Hannan, 1970, pp. 64–66). Let $\mathcal{U}$ be an arbitrary neighborhood of Γ_0 in $\mathbb{T}_{n,q}$. The map $\Gamma(\cdot, \cdot)$ is continuous, so $\Gamma^{-1}(\mathcal{U})$ is a neighborhood of $(\check{\Theta}, \check{\sigma})$ in $\Xi_\Theta \times \Xi_\sigma$. Because $\Pi_{\Theta,\sigma}(\cdot)$ has full support on $\Xi_\Theta \times \Xi_\sigma$, we have $\Pi_\Gamma(\mathcal{U}) = \Pi_{\Theta,\sigma}(\Gamma^{-1}(\mathcal{U})) > 0$. Since the neighborhood $\mathcal{U}$ was arbitrary, Γ_0 lies in the support of the induced prior $\Pi_\Gamma(\cdot)$.

Finally, note that the empirical autocovariances $\hat{\Gamma}$ are consistent for the true ACF Γ_0 under Assumption 1.3. Hence, the assumptions of the general Lemma 1.1 are satisfied for the Whittle SVMA posterior, and Theorem 1.2 follows. □

B.1.7 Proof of Lemma A.3

The proof closely follows the steps in Dunsmuir & Hannan (1976, Sec. 3) for proving consistency of the Whittle maximum likelihood estimator in a reduced-form identified VARMA model. Note that the only properties of the data generating process used in Dunsmuir & Hannan (1976, Sec. 3) are covariance stationarity and ergodicity for second moments, as in Assumption 1.3. Dunsmuir & Hannan also need $T^{-1}y_t y'_{t+T-k} \xrightarrow{p} 0$ for fixed t and k, which follows from Markov's inequality under covariance stationarity. Where Dunsmuir & Hannan (1976) appeal to almost sure convergence, I substitute convergence in probability.

Define the normalized log likelihood ratio

$$\hat{\phi}(\beta, \Sigma) = T^{-1} \log \frac{p^W_{\beta,\Sigma}(Y_T \mid \beta, \Sigma)}{p^W_{\beta,\Sigma}(Y_T \mid \beta_0, \Sigma_0)}.$$

By the Kolmogorov-Szegö formula, for any $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log\det(\tilde{f}(\omega; \beta, \Sigma))\, d\omega = \log\det(\Sigma) - n\log(2\pi). \tag{B.1}$$

Hence,

$$\hat{\phi}(\beta, \Sigma) = \frac{1}{2}\log\det(\Sigma_0\Sigma^{-1}) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \operatorname{tr}\left\{[\tilde{f}(\omega; \beta_0, \Sigma_0)^{-1} - \tilde{f}(\omega; \beta, \Sigma)^{-1}]\hat{I}(\omega)\right\} d\omega. \tag{B.2}$$

Define also the function

$$\phi(\beta, \Sigma) = \frac{1}{2}\log\det(\Sigma_0\Sigma^{-1}) + \frac{1}{2}\int_{-\pi}^{\pi} \operatorname{tr}\{I_n - \tilde{f}(\omega; \beta, \Sigma)^{-1}\tilde{f}(\omega; \beta_0, \Sigma_0)\}\, d\omega.$$

$\phi(\beta, \Sigma)$ is continuous. By the argument in Dunsmuir & Hannan (1976, p. 5) (see also Brockwell & Davis, 1991, Prop. 10.8.1, for the univariate case), we have $\phi(\beta, \Sigma) \leq \phi(\beta_0, \Sigma_0) = 0$ for all $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$, with equality if and only if $(\beta, \Sigma) = (\beta_0, \Sigma_0)$.

The remainder of the proof verifies the conditions of Lemma A.2 in five steps.

STEP 1. I first show that there exists a neighborhood $\mathcal{K}$ of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$ such that

$$\sup_{(\beta,\Sigma)\in\mathcal{K}} |\hat{\phi}(\beta, \Sigma) - \phi(\beta, \Sigma)| = o_p(1). \tag{B.3}$$

By definition of the Wold decomposition of a time series with a non-singular spectral density, all the roots of $z \mapsto \det(\Phi(\beta_0; z))$ lie strictly outside the unit circle. $\tilde{f}(\omega; \beta, \Sigma)^{-1} = \Phi(\beta; e^{\imath\omega})^{-1\prime}\Sigma^{-1}\Phi(\beta; e^{-\imath\omega})^{-1}$ is therefore uniformly continuous in (ω, β, Σ) for all $\omega \in [-\pi, \pi]$ and (β, Σ) in a small neighborhood of (β_0, Σ_0). Denoting this neighborhood by $\mathcal{K}$, the discussion around Lemma 1 in Dunsmuir & Hannan (1976, p. 350) implies (B.3).

STEP 2. For any $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$ and $z \in \mathbb{C}$, define the adjoint of $\Phi(\beta; z)$ as

$$\Phi_{\text{adj}}(\beta; z) = \Phi(\beta; z)^{-1} \det(\Phi(\beta; z)),$$

so $\tilde{f}(\omega; \beta, \Sigma) = |\det(\Phi(\beta; e^{-\imath\omega}))|^2\Phi_{\text{adj}}(\beta; e^{-\imath\omega})^{-1}\Sigma\Phi_{\text{adj}}(\beta; e^{-\imath\omega})^{-1*}$. The elements of $\Phi_{\text{adj}}(\beta; z)$ are polynomials in z, each polynomial of order $\kappa \leq q(n-1)$ (Dunsmuir & Hannan, 1976, p. 354). Write the matrix polynomial as $\Phi_{\text{adj}}(\beta; z) = I_n + \sum_{\ell=1}^{\kappa} \beta_{\text{adj},\ell} z^\ell$, and define $\tilde{\Phi}_{\text{adj}}(\beta) = (\sum_{\ell=1}^{\kappa} \|\beta_{\text{adj},\ell}\|^2)^{1/2}$.

Now define, for $\delta \geq 0$,

$$\tilde{f}_\delta(\omega; \beta, \Sigma) = (|\det(\Phi(\beta; e^{-\imath\omega}))|^2 + \delta)\Phi_{\text{adj}}(\beta; e^{-\imath\omega})^{-1}\Sigma\Phi_{\text{adj}}(\beta; e^{-\imath\omega})^{-1*},$$

$$\hat{\phi}_\delta(\beta, \Sigma) = \frac{1}{2}\log\det(\Sigma_0\Sigma^{-1}) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \text{tr}\left\{[\tilde{f}(\omega; \beta_0, \Sigma_0)^{-1} - \tilde{f}_\delta(\omega; \beta, \Sigma)^{-1}]\hat{I}(\omega)\right\} d\omega,$$

and

$$\phi_\delta(\beta, \Sigma) = \frac{1}{2}\log\det(\Sigma_0\Sigma^{-1}) + \frac{1}{2}\int_{-\pi}^{\pi} \text{tr}\{I_n - \tilde{f}_\delta(\omega; \beta, \Sigma)^{-1}\tilde{f}(\omega; \beta_0, \Sigma_0)\} d\omega.$$

Because $\hat{I}(\omega)$ is positive semidefinite for each $\omega \in [-\pi, \pi]$, we have $\hat{\phi}(\beta, \Sigma) \leq \hat{\phi}_\delta(\beta, \Sigma)$ for all $(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n$ and $\delta > 0$.

Finally, for any $c_1, c_2, c_3 > 0$, define the set

$$\tilde{\mathcal{K}}(c_1, c_2, c_3) = \{(\beta, \Sigma) \in \mathbb{B}_{n,q} \times \mathbb{S}_n \colon \lambda_{\min}(\Sigma) \geq c_1, \|\Sigma\| \leq c_2, \tilde{\Phi}_{\text{adj}}(\beta) \leq c_3\},$$

where $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ.

The discussion surrounding Lemma 3 in Dunsmuir & Hannan (1976, p. 351) then gives

$$\sup_{(\beta,\Sigma)\in\tilde{\mathcal{K}}(c_1,c_2,c_3)} |\hat{\phi}_\delta(\beta, \Sigma) - \phi_\delta(\beta, \Sigma)| = o_p(1),$$

for any $c_1, c_2, c_3 > 0$. Because $\phi_\delta(\beta, \Sigma)$ is continuous in (β, Σ, δ) at $(\beta = \beta_0, \Sigma = \Sigma_0, \delta = 0)$, and $\phi(\beta, \Sigma) = \phi_{\delta=0}(\beta, \Sigma)$ is uniquely maximized at (β_0, Σ_0),

$$\inf_{\delta>0} \sup_{(\beta,\Sigma)\notin\mathcal{K}} \phi_\delta(\beta, \Sigma) < \phi(\beta_0, \Sigma_0) = 0.$$

I conclude that for all $c_1, c_2, c_3 > 0$ there exist $\delta > 0$ and $\zeta > 0$ such that

$$\sup_{(\beta,\Sigma)\in\tilde{\mathcal{K}}(c_1,c_2,c_3)\cap\mathcal{K}^c} \hat{\phi}(\beta, \Sigma) \leq \sup_{(\beta,\Sigma)\in\tilde{\mathcal{K}}(c_1,c_2,c_3)\cap\mathcal{K}^c} \hat{\phi}_\delta(\beta, \Sigma) \leq -\zeta + o_p(1).$$

Step 3. Let $\zeta > 0$ be the scalar found in the previous step. The proof of Theorem 4(i) in Dunsmuir & Hannan (1976, pp. 354–355) (see also the beginning of the proof of their Theorem 3, pp. 352–353) shows that there exist $c_1, c_2, c_3 > 0$ such that

$$\sup_{(\beta,\Sigma)\in(\mathbb{B}_{n,q}\times\mathbb{S}_n)\cap\tilde{K}(c_1,c_2,c_3)^c} \hat{\phi}(\beta, \Sigma) \leq -\zeta.$$

Step 4. Steps 1–3 imply that the sufficient conditions in Lemma A.2 hold. I conclude that $P^W_{\beta,\Sigma|Y}(\tilde{\mathcal{U}} \mid Y_T) \xrightarrow{p} 1$ for any neighborhood $\tilde{\mathcal{U}}$ of (β_0, Σ_0) in $\mathbb{B}_{n,q} \times \mathbb{S}_n$.

STEP 5. Finally, I prove an assertion in Appendix A.1.7.3: Lemma A.3 holds for the *discretized* Whittle likelihood that replaces integrals $(2\pi)^{-1}\int_{-\pi}^{\pi} g(\omega)\,d\omega$ (for a 2π-periodic function $g(\cdot)$) in the definition of $\log p^{W}_{Y|\beta,\Sigma}(Y_T \mid \beta,\Sigma)$ with sums $T^{-1}\sum_{k=0}^{T-1} g(\omega_k)$, $\omega_k = 2\pi k/T$.

The proof of Theorem 4(ii) of Dunsmuir & Hannan (1976, p. 356) shows that steps 1–3 above carry through if the integral in expression (B.2) is replaced with a discretized sum. The only other effect of discretizing the integrals in the Whittle likelihood is that the Kolmogorov-Szegö formula (B.1) does not hold exactly. Instead,

$$T^{-1}\sum_{j=0}^{T-1}\log\det(\tilde{f}(\omega_j;\beta,\Sigma)) = \log\det(\Sigma) - n\log(2\pi) + T^{-1}\sum_{j=0}^{T-1}\log|\det(\Phi(\beta;e^{-\imath\omega_j}))|^2.$$

The posterior consistency result for the discretized Whittle posterior follows from steps 1–4 above if I show

$$\sum_{j=0}^{T-1}\log|\det(\Phi(\beta;e^{-\imath\omega_j}))|^2 \leq 2nq\log 2 \tag{B.4}$$

for all $(\beta,\Sigma)\in\mathbb{B}_{n,q}\times\mathbb{S}_n$, and furthermore,

$$\sum_{j=0}^{T-1}\log|\det(\Phi(\beta;e^{-\imath\omega_j}))|^2 = o_p(1) \tag{B.5}$$

uniformly in a small neighborhood of (β_0,Σ_0) in $\mathbb{B}_{n,q}\times\mathbb{S}_n$.

For any $\beta\in\mathbb{B}_{n,q}$ and $z\in\mathbb{C}$, write $\det(\Phi(z;\beta)) = \det(I_n + \sum_{\ell=1}^{q}\beta_\ell z^\ell) = \prod_{b=1}^{nq}(1-a_b(\beta)z)$ for some complex scalars $\{a_b(\beta)\}_{1\leq b\leq nq}$ that depend on β and satisfy $|a_b(\beta)|<1$ (Brockwell & Davis, 1991, p. 191). From the Taylor series $\log(1-z) = -\sum_{s=1}^{\infty} z^s/s$ (valid for $z\in\mathbb{C}$ inside the unit circle) we get, for all $\beta\in\mathbb{B}_{n,q}$,

$$\sum_{k=0}^{T-1}\log\det(\Phi(e^{-\imath\omega_k};\beta)) = -\sum_{k=0}^{T-1}\sum_{b=1}^{nq}\sum_{s=1}^{\infty}\frac{(a_b(\beta)e^{-\imath\omega_k})^s}{s} = -\sum_{b=1}^{nq}\sum_{s=1}^{\infty}\frac{(a_b(\beta))^s}{s}\sum_{k=0}^{T-1}e^{-\imath\omega_k s}.$$

Since $\sum_{k=0}^{T-1} e^{-\imath\omega_k s}$ equals T when s is an integer multiple of T, and equals 0 otherwise,

$$\sum_{k=0}^{T-1} \log\det(\Phi(e^{-\imath\omega_k};\beta)) = -\sum_{b=1}^{nq}\sum_{s=1}^{\infty}\frac{(a_b(\beta))^{sT}}{s} = \sum_{b=1}^{nq}\log\left(1-(a_b(\beta))^T\right).$$

Hence,

$$\begin{aligned}\sum_{k=0}^{T-1}\log|\det(\Phi(e^{-\imath\omega_k};\beta))|^2 &= \sum_{k=0}^{T-1}\log\det(\Phi(e^{-\imath\omega_k};\beta)) + \sum_{k=0}^{T-1}\log\det(\Phi(e^{-\imath\omega_k};\beta)^*)\\ &= \sum_{b=1}^{nq}\log\left|1-(a_b(\beta))^T\right|^2\\ &\le nq\log 4,\end{aligned}$$

where the inequality uses $|1-(a_b(\beta))^T|<2$. Claim (B.4) follows. For β in a small neighborhood of β_0, $\max_b|a_b(\beta)|$ is uniformly bounded away from 1. This implies claim (B.5). □

B.2 Proofs for Chapter 2

B.2.1 Proof of Theorem 2.1

To lighten the notation, we denote $\sum_i = \sum_{i=1}^N$ (the same for j) and $\sum_s = \sum_{s=1}^T$ (the same for t). A double sum $\sum_{i=1}^N\sum_{j=1}^N$ is denoted $\sum_{i,j}$.

We extend the proof of Theorem 1 in Bai & Ng (2002). By the definition of the estimator $\hat{F}^k$, we have $\hat{F}^k = (NT)^{-1}XX'\tilde{F}^k$, where $\tilde{F}^{k\prime}\tilde{F}^k/T = I_k$ (Bai & Ng, 2008). Define $e = (e_1,\ldots,e_T)'$ and $w = (w_1,\ldots,w_T)'$. Since

$$XX' = F\Lambda_0'\Lambda_0F' + F\Lambda_0'(e+w)' + (e+w)\Lambda_0F' + (e+w)(e+w)',$$

we can write

$$\hat{F}_t^k - H^{k'}F_t = (NT)^{-1}\Big\{\tilde{F}^{k'}F\Lambda_0'e_t + \tilde{F}^{k'}e\Lambda_0F_t + \tilde{F}^{k'}ee_t + \tilde{F}^{k'}F\Lambda_0'w_t + \tilde{F}^{k'}w\Lambda_0F_t + \tilde{F}^{k'}ww_t + \tilde{F}^{k'}ew_t + \tilde{F}^{k'}we_t\Big\}.$$

Label the eight terms on the right-hand side $A_{1t}, \ldots, A_{8t}$, respectively. By Loève's inequality,

$$T^{-1}\sum_t \|\hat{F}_t^k - H^{k'}F_t\|^2 \leq 8\sum_{n=1}^{8}\left(T^{-1}\sum_t \|A_{nt}\|^2\right). \tag{B.6}$$

Bai & Ng (2002) have shown that the terms corresponding to $n = 1, 2, 3$ are $O_p(C_{NT}^{-2})$ under Assumptions 2.1 to 2.3. We proceed to bound the remaining terms in probability.

We have

$$\|A_{4t}\|^2 \leq \left(T^{-1}\sum_s \|\tilde{F}_s^k\|^2\right)\left(T^{-1}\sum_s \|F_s\|^2\right)\left\|N^{-1}\Lambda_0'w_t\right\|^2.$$

The first factor equals $\mathrm{tr}(\tilde{F}^{k'}\tilde{F}^k/T) = \mathrm{tr}(I_k) = k$. The second factor is $O_p(1)$ by Assumption 2.1. Also,

$$\begin{aligned}
E\left\|\frac{\Lambda_0'w_t}{N}\right\|^2 &\leq N^{-2}\sum_{i,j}|E(w_{it}w_{jt})\lambda_{i0}'\lambda_{j0}| \\
&\leq \bar{\lambda}^2h_{NT}^2N^{-2}\sum_{i,j}|E(\xi_{it}F_t\xi_{it}F_t)| \\
&\leq r^2\bar{\lambda}^2\sup_{p,q}h_{NT}^2N^{-2}\sum_{i,j}|E(\xi_{itp}F_{tp}\xi_{itq}F_{tq})| \\
&= O(h_{NT}^2N^{-2}Q_1(N,T)),
\end{aligned}$$

uniformly in t, by Assumption 2.4.1. Hence,

$$T^{-1}\sum_t \|A_{4t}\|^2 = O_p(h_{NT}^2N^{-2}Q_1(N,T)).$$

Similarly,

$$\|A_{5t}\|^2 \leq \left(T^{-1}\sum_s \|\tilde{F}_s^k\|^2\right)\left((N^2T)^{-1}\sum_s (w_s'\Lambda_0 F_t)^2\right),$$

where the first term is $O(1)$ and

$$\begin{aligned}(N^2T)^{-1}E\sum_s (w_s'\Lambda_0 F_t)^2 &\leq (N^2T)^{-1}\sum_s\sum_{i,j}|E(w_{is}w_{js}\lambda_{i0}'F_t\lambda_{j0}'F_t)| \\ &\leq r^4\bar{\lambda}^2 \sup_{p_1,q_1,p_2,q_2} h_{NT}^2(N^2T)^{-1}\sum_s\sum_{i,j}|E(\xi_{isp_1}\xi_{jsq_1}F_{sp_1}F_{sq_1}F_{tp_2}F_{tq_2})|.\end{aligned}$$

By summing over t, dividing by T and using Assumption 2.4.2 we obtain

$$T^{-1}\sum_t \|A_{5t}\|^2 = O_p(h_{NT}^2N^{-2}T^{-2}Q_2(N,T)).$$

For the sixth term,

$$\begin{aligned}E\|A_{6t}\|^2 &\leq E\left\{\left(T^{-1}\sum_s \|\tilde{F}_s^k\|^2\right)\left((N^2T)^{-1}\sum_s (w_s'w_t)^2\right)\right\} \\ &= k(N^2T)^{-1}\sum_s\sum_{i,j}E(w_{is}w_{it}w_{js}w_{jt}) \\ &\leq kr^4 \sup_{p_1,q_1,p_2,q_2}\frac{h_{NT}^4}{N^2T}\sum_s\sum_{i,j}|E(\xi_{isp_1}\xi_{jsq_1}\xi_{itp_2}\xi_{jtq_2}F_{sp_1}F_{sq_1}F_{tp_2}F_{tq_2})|.\end{aligned}$$

By Assumption 2.4.3, it follows that

$$T^{-1}\sum_t \|A_{6t}\|^2 = O_p(h_{NT}^4N^{-2}T^{-2}Q_3(N,T)).$$

Regarding the seventh term, using Assumption 2.5,

$$\begin{aligned}E\|A_{7t}\|^2 &\leq E\left\{\left(T^{-1}\sum_s \|\tilde{F}_s^k\|^2\right)\left((N^2T)^{-1}\sum_s (e_s'w_t)^2\right)\right\} \\ &= k(N^2T)^{-1}\sum_s\sum_{i,j}E(e_{is}e_{js})E(w_{it}w_{jt})\end{aligned}$$

$$
\begin{aligned}
&\leq k(N^2T)^{-1}\sum_s\sum_{i,j}(E(e_{is}^2)E(e_{js}^2))^{1/2}|E(w_{it}w_{jt})| \\
&\leq kr^2M\sup_{p,q}h_{NT}^2(N^2T)^{-1}\sum_s\sum_{i,j}|E(\xi_{itp}\xi_{jtq}F_{tp}F_{tq})| \\
&= O(h_{NT}^2N^{-2}Q_1(N,T)),
\end{aligned}
$$

uniformly in t. The second-to-last line uses $E(e_{it}^2) \leq M$, whereas the last follows from Assumption 2.4.1. We conclude that

$$
T^{-1}\sum_t \|A_{7t}\|^2 = O_p(h_{NT}^2N^{-2}Q_1(N,T)).
$$

A similar argument gives

$$
T^{-1}\sum_t \|A_{8t}\|^2 = O_p(h_{NT}^2N^{-2}Q_1(N,T)).
$$

Hence, the right-hand side of inequality (B.6) is the sum of variables of four stochastic orders: $O_p(C_{NT}^{-2})$, $O_p(h_{NT}^2N^{-2}Q_1(N,T))$, $O_p(h_{NT}^2N^{-2}T^{-2}Q_2(N,T))$ and $O_p(h_{NT}^4N^{-2}T^{-2}Q_3(N,T))$. The statement of the theorem follows. □

B.3 Proofs for Chapter 3

B.3.1 Proof of Proposition 3.1

To simplify notation, I write $\beta_T^\dagger$ without the T subscript. Expand

$$
\begin{aligned}
\|\hat\beta_{M,W}(\lambda)-\beta^\dagger\|_{\tilde W}^2 &= \|\hat\beta_{M,W}(\lambda)-\hat\beta\|_{\tilde W}^2 + 2\hat\beta_{M,W}(\lambda)'\tilde W(\hat\beta-\beta^\dagger) + \|\beta^\dagger\|_{\tilde W}^2 - \|\hat\beta\|_{\tilde W}^2 \\
&= \|\hat\beta_{M,W}(\lambda)-\hat\beta\|_{\tilde W}^2 + 2\,\mathrm{tr}\{\tilde W\Theta_{M,W}(\lambda)\hat\beta(\hat\beta-\beta^\dagger)'\} + \|\beta^\dagger\|_{\tilde W}^2 - \|\hat\beta\|_{\tilde W}^2 \\
&= T^{-1}\hat R_{M,W,\tilde W}(\lambda) + 2T^{-1}\,\mathrm{tr}\left\{\tilde W\Theta_{M,W}(\lambda)[T\hat\beta(\hat\beta-\beta^\dagger)'-\hat\Sigma]\right\} + \|\beta^\dagger\|_{\tilde W}^2 - \|\hat\beta\|_{\tilde W}^2.
\end{aligned}
$$

Since $\Theta_{M,W}(\lambda)(I_n - P_M) = I_n - P_M$, the following random variable does not depend on λ:

$$\hat{C} = 2\,\mathrm{tr}\left\{\tilde{W}\Theta_{M,W}(\lambda)(I_n - P_M)[T\hat{\beta}(\hat{\beta} - \beta^\dagger)' - \hat{\Sigma}]\right\} + T\|\beta^\dagger\|^2_{\tilde{W}} - T\|\hat{\beta}\|^2_{\tilde{W}}.$$

We have

$$\begin{aligned}\left|R_{M,W,\tilde{W}}(\lambda) - E\left(\hat{R}_{M,W,\tilde{W}}(\lambda) + \hat{C}\right)\right| &= \left|\mathrm{tr}\left\{\tilde{W}\Theta_{M,W}(\lambda)P_M E[T\hat{\beta}(\hat{\beta} - \beta^\dagger)' - \hat{\Sigma}]\right\}\right| \\ &\leq \|\tilde{W}\|\|\Theta_{M,W}(\lambda)\|\left\|E[TP_M\hat{\beta}(\hat{\beta} - \beta^\dagger)' - P_M\hat{\Sigma}]\right\|.\end{aligned}$$

Note that $\|\Theta_{M,W}(\lambda)\| \leq \sqrt{n}\|W\|\rho((W + \lambda M'M)^{-1}) \leq \sqrt{n}\|W\|\rho(W^{-1})$ for all $\lambda \geq 0$. It remains to show that $E[TP_M\hat{\beta}(\hat{\beta} - \beta^\dagger)' - P_M\hat{\Sigma}] \to 0$. Write

$$TP_M\hat{\beta}(\hat{\beta} - \beta^\dagger)' - P_M\hat{\Sigma} = TP_M(\hat{\beta} - \beta^\dagger)(\hat{\beta} - \beta^\dagger)' - P_M\hat{\Sigma} + \sqrt{T}P_M\beta^\dagger\sqrt{T}(\hat{\beta} - \beta^\dagger)'.$$

By Assumption 3.1, the first term above converges in distribution to P_MUU', where $U \sim N(0, \Sigma)$, with uniformly integrable norm; hence, its expectation converges to $E(P_MUU') = P_M\Sigma$. Similarly, the expectation of the second term above converges to $-P_M\Sigma$. The last term above converges in distribution to hU', and uniform integrability of its norm follows easily from Assumption 3.1; hence, its expectation converges to $E(hU') = 0$. □

B.3.2 Proof of Proposition 3.2

I proceed in three steps.

Step 1. By the continuous mapping theorem,

$$\begin{aligned}\sqrt{T}(\hat{\beta}_P(\hat{\tau}) - \beta^\dagger) = \sqrt{T}(\hat{\beta} - \beta^\dagger) \\ - \min\left\{\frac{\hat{\tau}}{\|P\sqrt{T}(\hat{\beta} - \beta^\dagger) + \sqrt{T}P\beta^\dagger\|^2}, 1\right\}\{P\sqrt{T}(\hat{\beta} - \beta^\dagger) + \sqrt{T}P\beta^\dagger\}\end{aligned}$$

$$\xrightarrow{d} V,$$

where

$$V = U - \min\left\{\frac{\tau}{\|PU+h\|^2}, 1\right\}(PU+h), \quad U \sim N(0,\Sigma).$$

Note that $\min\{\tau/x^2, 1\}x \leq \sqrt{\tau}$ for all $\tau, x \geq 0$, implying that

$$\|\sqrt{T}(\hat{\beta}_P(\hat{\tau}) - \hat{\beta})\| \leq \sqrt{\hat{\tau}}.$$

Hence, $T\|\hat{\beta}_P(\hat{\tau}) - \beta^\dagger\|^2 \leq 2(T\|\hat{\beta} - \beta^\dagger\|^2 + \hat{\tau})$, so that the left-hand side is uniformly integrable by assumption. It follows that

$$\lim_{T\to\infty} E\left(T\|\hat{\beta}_P(\hat{\tau}) - \beta^\dagger\|^2\right) = E\|V\|^2.$$

The rest of the proof calculates an upper bound for the right-hand side above.

Step 2. Define the random variable

$$\tilde{V} = U - \frac{\tau}{\|PU+h\|^2}(PU+h).$$

I now show that $E\|V\|^2 \leq E\|\tilde{V}\|^2$ using the proof of Theorem 5.4, p. 356, in Lehmann & Casella (1998). Define the scalar random variable $B = \tau/\|PU+h\|^2$. Then $PV + h = (1-B)_+(PU+h)$ and $P\tilde{V} + h = (1-B)(PU+h)$, so that $\|P\tilde{V}+h\|^2 \geq \|PV+h\|^2$. Since $(I_n - P)V = (I_n - P)\tilde{V}$, we have

$$\begin{aligned} E\|\tilde{V}\|^2 - E\|V\|^2 &= E\|P\tilde{V}\|^2 + E\|PV\|^2 \\ &= E\|P\tilde{V}+h\|^2 - E\|PV+h\|^2 - 2E[h'P(\tilde{V}-V)] \\ &\geq -2E[h'P(\tilde{V}-V)] \end{aligned}$$

$$= 2E[(B-1)h'(PU+h) \mid B > 1] \Pr(B > 1),$$

where the last equality uses that $\tilde{V} = V$ on the event $\{B \leq 1\}$, while $PV + h = 0$ on the complementary event $\{B > 1\}$. We have $E\|\tilde{V}\|^2 \geq E\|V\|^2$ if I show that $E[h'(PU+h) \mid B = b] \geq 0$ for all $b > 1$, which in turn would be implied by $E[h'(PU+h) \mid \|PU+h\|^2 = c] \geq 0$ for all $c > 0$.

Note that $Ph = \lim_{T\to\infty} \sqrt{T} P^2 \beta^\dagger = h$. Let $\tilde{m} = \mathrm{rk}(P)$, and write $P = AA'$ for some $A \in \mathbb{R}^{n\times\tilde{m}}$ with full column rank and satisfying $A'A = I_{\tilde{m}}$. Diagonalize $A'\Sigma A = QDQ'$, where $QQ' = I_{\tilde{m}}$ and $D \in \mathbb{R}^{\tilde{m}\times\tilde{m}}$ is diagonal. Let $\tilde{h} = Q'A'h$. Then $(h'(PU+h), \|PU+h\|)$ has the same distribution as $(\tilde{h}'(\tilde{U}+\tilde{h}), \|\tilde{U}+\tilde{h}\|)$, where $\tilde{U} \sim N(0, D)$. Denote the i-th element of $\tilde{U}$ by $\tilde{U}_i$. To show $E\|\tilde{V}\|^2 \geq E\|V\|^2$, it suffices to show $\tilde{h}_i E[\tilde{U}_i + \tilde{h}_i \mid \sum_{j=1}^{\tilde{m}} (\tilde{U}_j + \tilde{h}_j)^2 = c] \geq 0$ for all $i = 1, \ldots, \tilde{m}$. The latter follows from essentially the same arguments as Lehmann & Casella (1998, bottom of p. 356) use for the case $D = I_{\tilde{m}}$.

Step 3. It remains to bound $E\|\tilde{V}\|^2$. From here on I follow the proof of Theorem 2 in Hansen (2016b). Using $E\|U\|^2 = \mathrm{tr}(\Sigma)$, we have

$$E\|\tilde{V}\|^2 = \mathrm{tr}(\Sigma) + \tau^2 E\left(\frac{1}{\|PU+h\|^2}\right) - 2\tau E[\eta(U+h)'PU], \tag{B.7}$$

where $\eta(x) = x/\|Px\|^2$ for $x \in \mathbb{R}^n$. By Stein's Lemma (Hansen, 2016b, Lem. 2),

$$\begin{aligned}
E[\eta(U+h)'PU] &= E\left[\mathrm{tr}\left(\frac{\partial}{\partial x}\eta(U+h)'P\Sigma\right)\right] \\
&= E\left[\mathrm{tr}\left\{\left(\frac{1}{\|PU+h\|^2} I_n - \frac{2}{\|PU+h\|^4} P(U+h)(U+h)'\right) P\Sigma\right\}\right] \\
&= E\left[\frac{\mathrm{tr}(\Sigma_P)}{\|PU+h\|^2} - \frac{2\,\mathrm{tr}\{(PU+h)'\Sigma_P(PU+h)\}}{\|PU+h\|^4}\right] \\
&\geq E\left[\frac{\mathrm{tr}(\Sigma_P)}{\|PU+h\|^2} - \frac{2\rho(\Sigma_P)\|PU+h\|^2}{\|PU+h\|^4}\right]
\end{aligned}$$

$$= E\left[\frac{\text{tr}(\Sigma_P) - 2\rho(\Sigma_P)}{\|PU + h\|^2}\right].$$

Inserting this into equation (B.7), we obtain

$$\begin{aligned} E\|\tilde{V}\|^2 &\leq \text{tr}(\Sigma) - \tau E\left(\frac{2(\text{tr}(\Sigma_P) - 2\rho(\Sigma_P)) - \tau}{\|PU + h\|^2}\right) \\ &\leq \text{tr}(\Sigma) - \tau\frac{2(\text{tr}(\Sigma_P) - 2\rho(\Sigma_P)) - \tau}{E\|PU + h\|^2}, \end{aligned}$$

where the last line uses Jensen's inequality and the assumption $0 \leq \tau \leq 2(\text{tr}(\Sigma_P) - 2\rho(\Sigma_P))$. Finally, observe that $E\|PU + h\|^2 = \text{tr}(\Sigma_P) + \|h\|^2$. □

B.3.3 Proof of Corollary 3.1

I verify the conditions of Proposition 3.2. With $\hat{\tau} = \text{tr}(\hat{\Sigma}_P)$, Assumption 3.1 implies $\tau = \text{plim}_{T\to\infty}\hat{\tau} = \text{tr}(\Sigma_P)$. It remains to show that $\{T\|\hat{\beta} - \beta_T^\dagger\|^2 + \hat{\tau}\}_{T\geq 1}$ is uniformly integrable, which follows from Assumption 3.1 and $\hat{\tau} = \text{tr}(\hat{\Sigma}P) \leq \|\hat{\Sigma}P\| \leq \|\hat{\Sigma}\|\rho(P) = \|\hat{\Sigma}\|$. □

B.3.4 Proof of Proposition 3.3

I follow the proofs of Theorem 1 in Andrews & Guggenberger (2010) and of "Theorem Bonf" in McCloskey (2015). There exist a sequence $\{\beta_T, \Sigma_T, \gamma_T\}_{T\geq 1} \in \mathbb{R}^n \times \mathcal{S} \times \Gamma$ and a subsequence $\{k_T\}_{T\geq 1}$ of $\{T\}_{T\geq 1}$ such that the left-hand side of (3.14) equals

$$\lim_{T\to\infty} \text{Prob}_{F_{k_T}(\beta_{k_T}, \Sigma_{k_T}, \gamma_{k_T})}\left(\hat{S}(\beta_{k_T}) \leq q_{1-\alpha}(\sqrt{k_T}\beta_{k_T}, \hat{\Sigma})\right). \tag{B.8}$$

Define $\tilde{\beta}_{k_T} = P\beta_{k_T}$, and let $\tilde{\beta}_{i,k_T}$ denote its i-th element, $i = 1, \ldots, n$. For an index i, either (a) $\limsup_{T\to\infty} |\sqrt{k_T}\tilde{\beta}_{i,k_T}| < \infty$ or (b) $\limsup_{T\to\infty} |\sqrt{k_T}\tilde{\beta}_{i,k_T}| = \infty$. In case (a), there exist

an $h_i \in \mathbb{R}$ and a further subsequence $\{\tilde{k}_T\}_{T\geq 1}$ of $\{k_T\}_{T\geq 1}$ such that

$$\lim_{T\to\infty} \sqrt{\tilde{k}_T}\tilde{\beta}_{i,\tilde{k}_T} = h_i. \tag{B.9}$$

In case (b), there exists a further subsequence $\{\tilde{k}_T\}_{T\geq 1}$ of $\{k_T\}_{T\geq 1}$ such that

$$\lim_{T\to\infty} \sqrt{\tilde{k}_T}\tilde{\beta}_{i,\tilde{k}_T} \in \{-\infty, \infty\}. \tag{B.10}$$

Moreover, since $\mathcal{S}$ is compact, there exist $\tilde{\Sigma} \in \mathcal{S}$ and a further subsequence $\{\tilde{k}_T\}_{T\geq 1}$ of $\{k_T\}_{T\geq 1}$ such that

$$\lim_{T\to\infty} \Sigma_{\tilde{k}_T} = \tilde{\Sigma}. \tag{B.11}$$

By sequentially choosing further subsequences for each $i = 1, \ldots, n$, we can find a subsequence $\{\tilde{k}_T\}_{T\geq 1}$ of $\{k_T\}_{T\geq 1}$ such that – for every $i = 1, \ldots, n$ – either (B.9) or (B.10) holds, and such that (B.11) holds. Since any subsequence of a convergent sequence converges to the same limit, expression (B.8) equals

$$\lim_{T\to\infty} \text{Prob}_{F_{\tilde{k}_T}(\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T})} \left(\hat{S}(\beta_{\tilde{k}_T}) \leq q_{1-\alpha}(\sqrt{\tilde{k}_T}\beta_{\tilde{k}_T}, \hat{\Sigma}) \right). \tag{B.12}$$

Write

$$\begin{aligned} \hat{S}(\beta_{\tilde{k}_T}) = g\Big(& \sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) \\ & - f(\|P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2, \hat{\Sigma})\Big\{P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\Big\}, \hat{\Sigma}\Big). \end{aligned} \tag{B.13}$$

There are now two cases to consider.

Case I. Suppose first that (B.10) holds for some i. Then $\lim_{T\to\infty} \|\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\| = \infty$, and by Assumption 3.2 and equation (B.11),

$$\|P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\| \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T})]{p} \infty.$$

Hence, by Assumptions 3.2 and 3.3,

$$\begin{aligned}
&\left\| f(\|P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2, \hat{\Sigma})(P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}) \right\| \\
&\qquad \leq \left| f(\|P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2, \hat{\Sigma}) \right| \left\| P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}) \right\| \\
&\qquad \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T})]{p} 0.
\end{aligned}$$

Consequently, using Assumption 3.2 on expression (B.13),

$$\hat{S}(\beta_{\tilde{k}_T}) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T})]{d} g(U, \tilde{\Sigma}), \quad U \sim N(0, \tilde{\Sigma}).$$

A similar argument shows that $q_{1-\alpha}(\sqrt{\tilde{k}_T}\beta_{\tilde{k}_T}, \hat{\Sigma})$ converges in probability under the sequence $F_{\tilde{k}_T}(\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T})$ to the $1-\alpha$ quantile of $g(U, \tilde{\Sigma})$, where $U \sim N(0, \tilde{\Sigma})$. It follows that (B.12) – and thus the left-hand side of (3.14) – equals $1 - \alpha$.

Case II. Suppose instead that (B.9) holds for all $i = 1, \ldots, n$. Let $h = (h_1, \ldots, h_n)'$, and note that $Ph = h$. By expression (B.13) and Assumptions 3.2 and 3.3

$$\hat{S}(\beta_{\tilde{k}_T}) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T})]{d} g\left(U - f(\|PU + h\|^2, \tilde{\Sigma})(PU + h), \tilde{\Sigma}\right), \quad U \sim N(0, \tilde{\Sigma}).$$

Moreover, using the continuous mapping theorem and $q_{1-\alpha}(\theta, \Sigma) = q_{1-\alpha}(P\theta, \Sigma)$ for all θ, Σ,

$$q_{1-\alpha}(\sqrt{\tilde{k}_T}\beta_{\tilde{k}_T}, \hat{\Sigma}) = q_{1-\alpha}(\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}, \hat{\Sigma}) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T})]{p} q_{1-\alpha}(h, \tilde{\Sigma}).$$

The definition of $q_{1-\alpha}(\cdot,\cdot)$ implies that (B.12) – and thus the left-hand side of (3.14) – equals $1-\alpha$. □

B.3.5 Proof of Proposition 3.4

I focus on the proof of (3.15), although I remark at the end how (3.16) can be obtained from similar arguments. Using the same steps as in the proof of Proposition 3.3, find a subsequence $\{\tilde{k}_T\}_{T\geq 1}$ of $\{T\}_{T\geq 1}$ and a sequence $\{\beta_{\tilde{k}_T}, \Sigma_{\tilde{k}_T}, \gamma_{\tilde{k}_T}\}_{T\geq 1}$ such that the left-hand side of (3.15) equals

$$\lim_{T\to\infty} \text{Prob}_{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}\left(\hat{S}_s(s'\beta_{\tilde{k}_T}) \leq q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T})+\hat{\nu}),\hat{\Sigma}\Big)\right), \tag{B.14}$$

and – for all $i=1,\ldots,n$ – either (B.9) or (B.10) holds, and moreover (B.11) holds, where $\tilde{\beta}_{i,\tilde{k}_T}$ is the i-th element of $P\tilde{\beta}_{\tilde{k}_T}$. Write

$$\begin{aligned}\hat{S}_s(s'\beta_{\tilde{k}_T}) = \Bigg(&\sqrt{\tilde{k}_T}s'(\hat{\beta}-\beta_{\tilde{k}_T}) \\ &- f(\|P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2,\hat{\Sigma})\Big\{s'P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}s'\tilde{\beta}_{\tilde{k}_T}\Big\}\Bigg)^2.\end{aligned} \tag{B.15}$$

There are two cases to consider.

Case I. Suppose that (B.10) holds for some $i=1,\ldots,n$. As in the proof of Proposition 3.3,

$$\begin{aligned}&\Big|f(\|P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2,\hat{\Sigma})\Big\{s'P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}s'\tilde{\beta}_{\tilde{k}_T}\Big\}\Big| \\ &\quad\leq \Big|f(\|P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2,\hat{\Sigma})\Big|\Big\|P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}s'\tilde{\beta}_{\tilde{k}_T}\Big\|\|s\| \\ &\quad\xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{p} 0.\end{aligned}$$

Thus, from (B.15),

$$\hat{S}_s(s'\beta_{\tilde{k}_T}) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{d} (s'\tilde{\Sigma}s)\chi^2(1).$$

Moreover,

$$\begin{aligned} q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T})+\hat{\nu}),\hat{\Sigma}\Big) &= q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{P}\beta_{\tilde{k}_T}+(I_n-\hat{P})\hat{\beta}),\hat{\Sigma}\Big) \\ &= q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}\beta_{\tilde{k}_T}+(I_n-\hat{P})\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T}),\hat{\Sigma}\Big) \\ &= q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}+P(I_n-\hat{P})\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T}),\hat{\Sigma}\Big) \qquad \text{(B.16)} \\ &\xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{p} (s'\tilde{\Sigma}s)z_{1,1-\alpha}, \end{aligned}$$

since $q_{s,1-\alpha}(\theta,\Sigma)=q_{s,1-\alpha}(P\theta,\Sigma)$ for all θ,Σ, and

$$\|\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}+P(I_n-\hat{P})\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})\| \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{p} \infty.$$

The above displays imply that (B.14) – and thus the left-hand side of (3.15) – equals $1-\alpha$.

Case II. Suppose now that (B.9) holds for all $i=1,\dots,n$. Let $h=(h_1,\dots,h_n)'$, and note that $Ph=h$. Then

$$\hat{S}_s(s'\beta_{\tilde{k}_T}) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{d} V,$$

where

$$V=\Big\{s'U-f(\|PU+h\|^2,\tilde{\Sigma})(s'PU+s'h)\Big\}^2, \quad U\sim N(0,\tilde{\Sigma}). \qquad \text{(B.17)}$$

Jointly with the above convergence, expression (B.16) yields

$$q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T})+\hat{\nu}),\hat{\Sigma}\Big) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{d} q_{s,1-\alpha}\Big(h+(I_n-\tilde{P})U,\tilde{\Sigma}\Big),$$

where $\tilde{P} = \tilde{\zeta}s'$, $\tilde{\zeta} = (s'\tilde{\Sigma}s)^{-1}\tilde{\Sigma}s$. I will have shown that (B.14) – and thus the left-hand side of (3.15) – equals $1-\alpha$ if I show that

$$\text{Prob}\left(V \leq q_{s,1-\alpha}\Big(h + (I_n - \tilde{P})U, \tilde{\Sigma}\Big)\right) = 1 - \alpha. \tag{B.18}$$

Using $U = \tilde{P}U + (I_n - \tilde{P})U = \tilde{\zeta}(s'U) + (I_n - \tilde{P})U$, write

$$V = \Big\{(s'U) - f(\|P\{\tilde{\zeta}(s'U) + h + (I_n - \tilde{P})U\}\|^2, \tilde{\Sigma})s'P\{\tilde{\zeta}(s'U) + h + (I_n - \tilde{P})U\}\Big\}^2.$$

The two jointly Gaussian variables $s'U$ and $(I_n - \tilde{P})U$ are uncorrelated and thus independent:

$$E[s'UU'(I_n - \tilde{P})'] = s'\Sigma(I_n - \tilde{P})' = s'(I_n - \tilde{P})\Sigma = 0,$$

using $\Sigma\tilde{P}' = \tilde{P}\Sigma$ and $s'\tilde{P} = s'$. By the independence,

$$\begin{aligned}
&\text{Prob}\left(V \leq q_{s,1-\alpha}\Big(h + (I_n - \tilde{P})U, \tilde{\Sigma}\Big) \,\Big|\, (I_n - \tilde{P})U = \nu\right) \\
&\quad = \text{Prob}\Bigg(\Big\{(s'U) - f(\|P\{\tilde{\zeta}(s'U) + h + \nu\}\|^2, \tilde{\Sigma})s'P\{\tilde{\zeta}(s'U) + h + \nu\}\Big\}^2 \\
&\qquad\qquad \leq q_{s,1-\alpha}\Big(h + \nu, \tilde{\Sigma}\Big)\Bigg) \\
&\quad = 1 - \alpha,
\end{aligned}$$

where the last equality holds by definition of $q_{s,1-\alpha}(\cdot,\cdot)$. The above conditional result implies the unconditional statement (B.18).

PROOF OF (3.16). As in the above proof of (3.15), pick a subsequence of parameters such that the left-hand side of (3.16) equals

$$\lim_{T} \text{Prob}_{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}\Big(\hat{S}_{s,W}(s'\beta_{\tilde{k}_T}) \le \hat{c}^2_{1-\delta}, \hat{S}_s(s'\beta_{\tilde{k}_T}) \le \tilde{q}_{s,\frac{1-\alpha}{1-\delta}}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T})+\hat{\nu}), \hat{\Sigma}, \hat{c}_{1-\delta}\Big)\Big). \tag{B.19}$$

As above, there are two cases. In "Case I", the limit (B.19) equals

$$\text{Prob}\left(u^2 \le c^2_{1-\delta}, u^2 \le \tilde{z}_{\frac{1-\alpha}{1-\delta}}(s'\tilde{\Sigma}s, c_{1-\delta})\right),$$

where $u \sim N(0, s'\tilde{\Sigma}s)$, $c_{1-\delta} = \sqrt{(s'\tilde{\Sigma}s)z_{1,1-\delta}}$, and $\tilde{z}_{\frac{1-\alpha}{1-\delta}}(s'\tilde{\Sigma}s, c_{1-\delta})$ equals the $\frac{1-\alpha}{1-\delta}$ quantile of the distribution of the square of a truncated normal variable with mean 0, variance parameter $s'\tilde{\Sigma}s$ and truncation interval $|u| \le c_{1-\delta}$. The above display equals, by definition of $\tilde{z}_{\frac{1-\alpha}{1-\delta}}(\cdot,\cdot)$,

$$\text{Prob}\left(u^2 \le c^2_{1-\delta}\right)\text{Prob}\left(u^2 \le \tilde{z}_{\frac{1-\alpha}{1-\delta}}(s'\tilde{\Sigma}s, c_{1-\delta}) \,\middle|\, u^2 \le c^2_{1-\delta}\right) = (1-\delta)\frac{1-\alpha}{1-\delta} = 1-\alpha.$$

In "Case II", the limit (B.19) equals

$$\text{Prob}\left((s'U)^2 \le c^2_{1-\delta},\ V \le \tilde{q}_{s,\frac{1-\alpha}{1-\delta}}\Big(h + (I_n - \tilde{P})U, \tilde{\Sigma}, c_{1-\delta}\Big)\right), \tag{B.20}$$

where $U \sim N(0, \tilde{\Sigma})$ and V is given by (B.17). The variables $s'U$ and $(I_n - \tilde{P})U$ are independent. Hence, conditional on $(I_n - \tilde{P})U = \nu$, the event in (B.20) has probability

$$\text{Prob}\left((s'U)^2 \le c^2_{1-\delta}\right)\text{Prob}\Big(\Big\{(s'U) - f(\|P\{\tilde{\zeta}(s'U)+h+\nu\}\|^2, \tilde{\Sigma})s'P\{\tilde{\zeta}(s'U)+h+\nu\}\Big\}^2 \le \tilde{q}_{s,\frac{1-\alpha}{1-\delta}}\Big(h+\nu, \tilde{\Sigma}, c_{1-\delta}\Big) \,\Big|\, (s'U)^2 \le c^2_{1-\delta}\Big),$$

which equals $1-\alpha$ by definition of $\tilde{q}_{s,\frac{1-\alpha}{1-\delta}}(\cdot,\cdot,\cdot)$. Thus, the unconditional probability (B.20) also equals $1-\alpha$. □

Bibliography

Adjemian, S., Bastani, H., Juillard, M., Karamé, F., Mihoubi, F., Perendia, G., Pfeifer, J., Ratto, M. & Villemot, S. (2011). Dynare: Reference Manual, Version 4. Dynare Working Papers, 1, CEPREMAP.

Alessi, L., Barigozzi, M. & Capasso, M. (2011). Non-Fundamentalness in Structural Econometric Models: A Review. *International Statistical Review* 79(1), 16–47.

Amengual, D. & Watson, M. W. (2007). Consistent Estimation of the Number of Dynamic Factors in a Large N and T Panel. *Journal of Business and Economic Statistics* 25(1), 91–96.

Andrews, D. W. (1991). Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* 47(2), 359–377.

Andrews, D. W. (1992). Generic Uniform Convergence. *Econometric Theory* 8(02), 241–257.

Andrews, D. W. & Guggenberger, P. (2010). Asymptotic Size and a Problem With Subsampling and With the m Out of n Bootstrap. *Econometric Theory* 26(02), 426–468.

Andrews, D. W. K., Cheng, X. & Guggenberger, P. (2011). Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests. Cowles Foundation Discussion Paper No. 1813.

Angrist, J. D., Jordà, Ò. & Kuersteiner, G. (2013). Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited. NBER Working Paper No. 19355.

Arias, J. E., Rubio-Ramírez, J. F. & Waggoner, D. F. (2014). Inference Based on SVARs Identified with Sign and Zero Restrictions: Theory and Applications. Federal Reserve

Bank of Atlanta Working Paper No. 2014-1.

Bai, J. & Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70(1), 191–221.

Bai, J. & Ng, S. (2006a). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica* 74(4), 1133–1150.

Bai, J. & Ng, S. (2006b). Determining the number of factors in approximate factor models, Errata. Manuscript, Columbia University.

Bai, J. & Ng, S. (2008). Large Dimensional Factor Analysis. *Foundations and Trends in Econometrics* 3(2), 89–163.

Bai, Z. & Silverstein, J. (2009). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer.

Banerjee, A., Marcellino, M. & Masten, I. (2008). Forecasting Macroeconomic Variables Using Diffusion Indexes in Short Samples with Structural Change. In D. E. Rapach & M. E. Wohar (Eds.), *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Vol. 3 of *Frontiers of Economics and Globalization*, Ch. 4, 149–194. Emerald Group Publishing.

Barnichon, R. & Matthes, C. (2015). Gaussian Mixture Approximations of Impulse Responses and The Non-Linear Effects of Monetary Shocks. Working paper.

Baumeister, C. & Hamilton, J. D. (2015a). Optimal Inference about Impulse-Response Functions and Historical Decompositions in Incompletely Identified Structural Vector Autoregressions. Working paper.

Baumeister, C. & Hamilton, J. D. (2015b). Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information. *Econometrica* 83(5), 1963–1999.

Baumeister, C. & Hamilton, J. D. (2015c). Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Demand Shocks. Working paper.

Beaudry, P., Fève, P., Guay, A. & Portier, F. (2015). When is Nonfundamentalness in VARs

a Real Problem? An Application to News Shocks. NBER Working Paper No. 21466.

Beaudry, P. & Portier, F. (2014). News-Driven Business Cycles: Insights and Challenges. *Journal of Economic Literature* 52(4), 993–1074.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives* 28(2), 29–50.

Beran, R. (2010). The unbearable transparency of Stein estimation. In J. Antoch, M. Hušková, & P. K. Sen (Eds.), *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, 25–34. Institute of Mathematical Statistics.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer.

Blanchard, O. J., L'Huillier, J. P. & Lorenzoni, G. (2013). News, Noise, and Fluctuations: An Empirical Exploration. *American Economic Review* 103(7), 3045–3070.

Bock, M. E. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Annals of Statistics* 3(1), 209–218.

Breitung, J. & Eickmeier, S. (2011). Testing for structural breaks in dynamic factor models. *Journal of Econometrics* 163(1), 71–84.

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods* (2nd ed.). Springer Series in Statistics. Springer.

Broda, C. & Parker, J. A. (2014). The Economic Stimulus Payments of 2008 and the aggregate demand for consumption. *Journal of Monetary Economics* 68, S20–S36. Supplement issue: October 19–20, 2012 Research Conference on "Financial Markets, Financial Policy, and Macroeconomic Activity".

Brown, L. D., Casella, G. & Hwang, J. T. G. (1995). Optimal Confidence Sets, Bioequivalence, and the Lamaçon of Pascal. *Journal of the American Statistical Association* 90(431), 880–889.

Casella, G. & Hwang, J. T. (1987). Employing vague prior information in the construction of confidence sets. *Journal of Multivariate Analysis* 21(1), 79–104.

Casella, G. & Hwang, J. T. G. (2012). Shrinkage Confidence Procedures. *Statistical Science* 27(1), 51–60.

Chan, J. C. C., Eisenstat, E. & Koop, G. (2015). Large Bayesian VARMAs. Working paper.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9), 2633–2679.

Chib, S. (2001). Markov Chain Monte Carlo Methods: Computation and Inference. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of Econometrics*, Vol. 5, Ch. 57, 3569–3649. North-Holland.

Chib, S. & Greenberg, E. (1994). Bayes inference in regression models with ARMA (p, q) errors. *Journal of Econometrics* 64, 183–206.

Choudhuri, N., Ghosal, S. & Roy, A. (2005). Bayesian Methods for Function Estimation. In D. Dey & C. Rao (Eds.), *Handbook of Statistics*, Vol. 25, Ch. 13, 373–414. Elsevier.

Christiano, L. J. & Vigfusson, R. J. (2003). Maximum likelihood in the frequency domain: the importance of time-to-plan. *Journal of Monetary Economics* 50(4), 789–815.

Chudik, A. & Pesaran, M. H. (2011). Infinite-dimensional VARs and factor models. *Journal of Econometrics* 163(1), 4–22.

Claeskens, G. & Hjort, N. L. (2003). The Focused Information Criterion. *Journal of the American Statistical Association* 98(464), 900–916.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Clements, M. & Hendry, D. (1998). *Forecasting Economic Time Series*. Cambridge University Press.

Cochrane, J. H. & Piazzesi, M. (2002). The Fed and Interest Rates—A High-Frequency

Identification. *American Economic Review* 92(2), 90–95.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Advanced Texts in Econometrics. Oxford University Press.

Del Negro, M. & Schorfheide, F. (2004). Priors from General Equilibrium Models for VARs. *International Economic Review* 45(2), 643–673.

Drèze, J. H. & Richard, J.-F. (1983). Bayesian Analysis of Simultaneous Equation Models. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of Econometrics*, Vol. I, Ch. 9, 517–598. North-Holland.

Dunsmuir, W. & Hannan, E. J. (1976). Vector Linear Time Series Models. *Advances in Applied Probability* 8(2), 339–364.

Durbin, J. & Koopman, S. (2012). *Time Series Analysis by State Space Methods* (2nd ed.). Oxford Statistical Science Series. Oxford University Press.

Dwyer, M. (1998). Impulse Response Priors for Discriminating Structural Vector Autoregressions. Working paper.

Eickmeier, S., Lemke, W. & Marcellino, M. (2015). Classical time varying factor-augmented vector auto-regressive models—estimation, forecasting and structural analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3), 493–533.

Eickmeier, S. & Ziegler, C. (2008). How Successful are Dynamic Factor Models at Forecasting Output and Inflation? A Meta-Analytic Approach. *Journal of Forecasting* 27(3), 237–265.

Engle, R. & Watson, M. W. (1981). A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates. *Journal of the American Statistical Association* 76(376), 774–781.

Fernald, J. (2014). A Quarterly, Utilization-Adjusted Series on Total Factor Productivity. Federal Reserve Bank of San Francisco Working Paper No. 2012-19.

Fernández-Villaverde, J., Rubio-Ramírez, J. F., Sargent, T. J. & Watson, M. W. (2007). ABCs (and Ds) of Understanding VARs. *American Economic Review* 97(3), 1021–1026.

Fessler, P. & Kasy, M. (2016). How to use economic theory to improve estimators, with an application to labor demand and wage inequality. Working paper.

Fève, P. & Jidoud, A. (2012). Identifying News Shocks from SVARs. *Journal of Macroeconomics* 34(4), 919–932.

Forni, M. & Gambetti, L. (2014). Sufficient information in structural VARs. *Journal of Monetary Economics* 66, 124–136.

Forni, M., Gambetti, L., Lippi, M. & Sala, L. (2013). Noisy News in Business Cycles. Working paper.

Forni, M., Gambetti, L. & Sala, L. (2014). No News in Business Cycles. *Economic Journal* 124(581), 1168–1191.

Forni, M., Giannone, D., Lippi, M. & Reichlin, L. (2009). Opening the Black Box: Structural Factor Models With Large Cross Sections. *Econometric Theory* 25(05), 1319–1347.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2000). The Generalized Dynamic-Factor Model: Identification and Estimation. *Review of Economics and Statistics* 82(4), 540–554.

Frisch, R. (1933). Propagation Problems and Impulse Problems in Dynamic Economics. In *Economic Essays in Honor of Gustav Cassel*. George Allen & Unwin.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Gertler, M. & Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics* 7(1), 44–76.

Geweke, J. (1977). The Dynamic Factor Analysis of Economic Time Series. In D. J. Aigner & A. S. Goldberger (Eds.), *Latent Variables in Socio-Economic Models*. North-Holland.

Geweke, J. (1988). The Secular and Cyclical Behavior of Real GDP in 19 OECD Countries, 1957–1983. *Journal of Business & Economic Statistics* 6(4), 479–486.

Geweke, J. (2010). *Complete and Incomplete Econometric Models*. The Econometric and Tinbergen Institutes Lectures. Princeton University Press.

Ghosh, J. K. & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer.

Giacomini, R. & Kitagawa, T. (2015). Robust inference about partially identified SVARs. Working paper.

Giannone, D., Lenza, M. & Primiceri, G. E. (2015). Prior Selection for Vector Autoregressions. *Review of Economics and Statistics* 97(2), 436–451.

Giannone, D. & Reichlin, L. (2006). Does Information Help Recovering Structural Shocks from Past Observations? *Journal of the European Economic Association* 4(2–3), 455–465.

Gilchrist, S. & Zakrajšek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *American Economic Review* 102(4), 1692–1720.

Gordon, S. & Boccanfuso, D. (2001). Learning from Structural Vector Autoregression Models. Working paper.

Gospodinov, N. & Ng, S. (2015). Minimum Distance Estimation of Possibly Noninvertible Moving Average Models. *Journal of Business & Economic Statistics* 33(3), 403–417.

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* 57(2), 357–384.

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

Hannan, E. (1970). *Multiple Time Series*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Hansen, B. E. (2010). Multi-Step Forecast Model Selection. Working paper.

Hansen, B. E. (2016a). A Stein-Like 2SLS Estimator. *Econometric Reviews*. Forthcoming.

Hansen, B. E. (2016b). Efficient shrinkage in parametric models. *Journal of Econometrics* 190(1), 115–132.

Hansen, B. E. (2016c). Stein Shrinkage for Vector Autoregressions. Working paper.

Hansen, L. P. & Sargent, T. (1981). Exact linear rational expectations models: Specification and estimation. Federal Reserve Bank of Minneapolis Staff Report No. 71.

Hansen, L. P. & Sargent, T. J. (1991). Two Difficulties in Interpreting Vector Autoregressions. In L. P. Hansen & T. J. Sargent (Eds.), *Rational Expectations Econometrics*, Underground Classics in Economics, Ch. 4, 77–119. Westview Press.

Hendry, D., Pagan, A. & Sargan, J. (1984). Dynamic Specification. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of Econometrics*, Vol. II, Ch. 18, 1023–1100. Elsevier.

Hodrick, R. J. & Prescott, E. C. (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking* 29(1), 1–16.

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15, 1593–1623.

Imbens, G. W. & Kolesár, M. (2016). Robust Standard Errors in Small Samples: Some Practical Advice. *Review of Economics and Statistics*. Forthcoming.

Ingram, B. F. & Whiteman, C. H. (1994). Supplanting the 'Minnesota' prior: Forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics* 34(3), 497–510.

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

Ito, T. & Quah, D. (1989). Hypothesis Testing with Restricted Spectral Density Matrices, with an Application to Uncovered Interest Parity. *International Economic Review* 30(1), 203–215.

Jacobson, L. S., LaLonde, R. J. & Sullivan, D. G. (1993). Earnings Losses of Displaced Workers. *American Economic Review* 83(4), 685–709.

James, W. & Stein, C. M. (1961). Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: Contribu-

tions to the Theory of Statistics, 361–379. University of California Press.

Jordà, O. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review* 95(1), 161–182.

Joshi, V. M. (1969). Admissibility of the Usual Confidence Sets for the Mean of a Univariate or Bivariate Normal Population. *Annals of Mathematical Statistics* 40(3), 1042–1067.

Kilian, L. & Murphy, D. P. (2012). Why Agnostic Sign Restrictions Are Not Enough: Understanding The Dynamics of Oil Market VAR Models. *Journal of the European Economic Association* 10(5), 1166–1188.

Klaeffing, M. (2003). Monetary Policy Shocks – A Nonfundamental Look at the Data. European Central Bank Working Paper No. 228.

Kline, B. & Tamer, E. (2015). Bayesian Inference in a Class of Partially Identified Models. Working paper.

Kocięcki, A. (2010). A Prior for Impulse Responses in Bayesian Structural VAR Models. *Journal of Business & Economic Statistics* 28(1), 115–127.

Komunjer, I. & Ng, S. (2011). Dynamic Identification of Dynamic Stochastic General Equilibrium Models. *Econometrica* 79(6), 1995–2032.

Koopman, S. J. & Shephard, N. (1992). Exact Score for Time Series Models in State Space Form. *Biometrika* 79(4), 823–826.

Korobilis, D. (2013). Assessing the Transmission of Monetary Policy Using Time-varying Parameter Dynamic Factor Models. *Oxford Bulletin of Economics and Statistics* 75(2), 157–179.

Lanne, M. & Saikkonen, P. (2013). Noncausal Vector Autoregression. *Econometric Theory* 29(03), 447–481.

Leeb, H. & Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory* 21(01), 21–59.

Leeper, E. M., Sims, C. A. & Zha, T. (1996). What Does Monetary Policy Do? *Brookings*

Papers on Economic Activity (2), 1–63.

Leeper, E. M., Walker, T. B. & Yang, S.-C. S. (2013). Fiscal Foresight and Information Flows. *Econometrica* 81(3), 1115–1145.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer Texts in Statistics. Springer.

Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd ed.). Springer Texts in Statistics. Springer.

Li, K.-C. (1986). Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing. *Annals of Statistics* 14(3), 1101–1112.

Lippi, M. & Reichlin, L. (1994). VAR analysis, nonfundamental representations, Blaschke matrices. *Journal of Econometrics* 63(1), 307–325.

Lopes, H. F. & Tobias, J. L. (2011). Confronting Prior Convictions: On Issues of Prior Sensitivity and Likelihood Robustness in Bayesian Analysis. *Annual Review of Economics* 3(1), 107–131.

Lütkepohl, H. (2014). Fundamental Problems with Nonfundamental Shocks. In N. Haldrup, M. Meitz, & P. Saikkonen (Eds.), *Essays in Nonlinear Time Series Econometrics*, Ch. 8, 198–214. Oxford University Press.

Magnus, J. & Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3rd ed.). Wiley Series in Probability and Statistics. John Wiley & Sons.

Mallows, C. L. (1973). Some Comments on C_P. *Technometrics* 15(4), 661–675.

McCloskey, A. (2015). Bonferroni-Based Size-Correction for Nonstandard Testing Problems. Working paper.

Mertens, K. & Ravn, M. O. (2010). Measuring the Impact of Fiscal Policy in the Face of Anticipation: A Structural VAR Approach. *Economic Journal* 120(544), 393–413.

Mertens, K. & Ravn, M. O. (2013). The Dynamic Effects of Personal and Corporate Income

Tax Changes in the United States. *American Economic Review* 103(4), 1212–1247.

Moon, H. R. & Schorfheide, F. (2012). Bayesian and Frequentist Inference in Partially Identified Models. *Econometrica* 80(2), 755–782.

Müller, U. K. (2012). Measuring prior sensitivity and prior informativeness in large Bayesian models. *Journal of Monetary Economics* 59(6), 581–597.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica* 81(5), 1805–1849.

Müller, U. K. (2014). HAC Corrections for Strongly Autocorrelated Time Series. *Journal of Business & Economic Statistics* 32(3), 311–322.

Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, Ch. 5, 113–162. CRC Press.

Oman, S. D. (1982). Contracting towards subspaces when estimating the mean of a multivariate normal distribution. *Journal of Multivariate Analysis* 12(2), 270–290.

Pagan, A. (1984). Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review* 25(1), 221–247.

Pesaran, H. M., Pettenuzzo, D. & Timmermann, A. (2006). Forecasting Time Series Subject to Multiple Structural Breaks. *Review of Economic Studies* 73(4), 1057–1084.

Pesaran, M. H. & Timmermann, A. (2005). Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics* 129(1–2), 183–217.

Pesaran, M. H. & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137(1), 134–161.

Phillips, P. (1989). Partially Identified Econometric Models. *Econometric Theory* 5(02), 181–240.

Phillips, P. C. B. & Solo, V. (1992). Asymptotics for Linear Processes. *Annals of Statistics* 20(2), 971–1001.

Poirier, D. J. (1998). Revising Beliefs in Nonidentified Models. *Econometric Theory* 14(04), 483–509.

Qu, Z. & Tkachenko, D. (2012a). Frequency Domain Analysis of Medium Scale DSGE Models with Application to Smets and Wouters (2007). In N. Balke, F. Canova, F. Milani, & M. A. Wynne (Eds.), *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments*, Vol. 28 of *Advances in Econometrics*, 319–385. Emerald Group Publishing.

Qu, Z. & Tkachenko, D. (2012b). Identification and frequency domain quasi-maximum likelihood estimation of linearized dynamic stochastic general equilibrium models. *Quantitative Economics* 3(1), 95–132.

Ramamoorthi, R. V., Sriram, K. & Martin, R. (2015). On Posterior Concentration in Misspecified Models. *Bayesian Analysis* 10(4), 759–789.

Ramey, V. A. (2016). Macroeconomic Shocks and Their Propagation. NBER Working Paper No. 21978. Draft of chapter to appear in *Handbook of Macroeconomics*, Vol. 2.

Rigobon, R. (2003). Identification Through Heteroskedasticity. *Review of Economics and Statistics* 85(4), 777–792.

Rotemberg, J. & Woodford, M. (1997). An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy. In B. S. Bernanke & J. Rotemberg (Eds.), *NBER Macroeconomics Annual*, Vol. 12, 297–346. MIT Press.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics 3*, 395–402. Oxford University Press.

Sala, L. (2015). DSGE Models in the Frequency Domain. *Journal of Applied Econometrics* 30(2), 219–240.

Sargent, T. J. (1989). Two Models of Measurements and the Investment Accelerator. *Journal of Political Economy* 97(2), 251–287.

Sargent, T. J. & Sims, C. A. (1977). Business Cycle Modeling Without Pretending to Have Too Much A Priori Economic Theory. In C. A. Sims (Ed.), *New Methods in Business Research*. Federal Reserve Bank of Minneapolis.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics* 3, 1039–1074.

Shiller, R. J. (1973). A Distributed Lag Estimator Derived from Smoothness Priors. *Econometrica* 41(4), 775–788.

Sims, C. (1980). Macroeconomics and Reality. *Econometrica* 48(1), 1–48.

Sims, C. A. & Zha, T. (1998). Bayesian Methods for Dynamic Multivariate Models. *International Economic Review* 39(4), 949–968.

Sims, C. A. & Zha, T. (1999). Error Bands for Impulse Responses. *Econometrica* 67(5), 1113–1155.

Sims, C. A. & Zha, T. (2006). Does Monetary Policy Generate Recessions? *Macroeconomic Dynamics* 10(02), 231–272.

Sims, E. R. (2012). News, Non-Invertibility, and Structural VARs. In N. Balke, F. Canova, F. Milani, & M. A. Wynne (Eds.), *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments*, Vol. 28 of *Advances in Econometrics*, 81–135. Emerald Group Publishing.

Stan Development Team (2015). *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*.

Stein, C. M. (1956). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: Contributions to the Theory of Statistics, 197–206. University of California Press.

Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics* 9(6), 1135–1151.

Stock, J. H. & Watson, M. W. (1989). New Indexes of Coincident and Leading Economic Indicators. In O. J. Blanchard & S. Fischer (Eds.), *NBER Macroeconomics Annual*, Vol. 4, 351–409. MIT Press.

Stock, J. H. & Watson, M. W. (1998a). Diffusion Indexes. NBER Working Paper No. 6702.

Stock, J. H. & Watson, M. W. (1998b). Median Unbiased Estimation of Coefficient Variance in a Time-Varying Parameter Model. *Journal of the American Statistical Association* 93(441), 349–358.

Stock, J. H. & Watson, M. W. (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association* 97(460), 1167–1179.

Stock, J. H. & Watson, M. W. (2008). Recent Developments in Structural VAR Modeling. NBER Summer Institute Minicourse 2008, "What's New in Econometrics – Time Series", Lecture 7.

Stock, J. H. & Watson, M. W. (2009). Forecasting in Dynamic Factor Models Subject to Structural Instability. In D. F. Hendry, J. Castle, & N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, 173–205. Oxford University Press.

Stock, J. H. & Watson, M. W. (2011). Dynamic Factor Models. In M. P. Clement & D. F. Hendry (Eds.), *The Oxford Handbook of Economic Forecasting*, 35–59. Oxford University Press.

Stock, J. H. & Watson, M. W. (2012a). Disentangling the Channels of the 2007–09 Recession. *Brookings Papers on Economic Activity* (Spring), 81–135.

Stock, J. H. & Watson, M. W. (2012b). Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business & Economic Statistics* 30(4), 481–493.

Stock, J. H. & Watson, M. W. (2016). Factor Models and Structural Vector Autoregressions in Macroeconomics. Draft of chapter to appear in *Handbook of Macroeconomics*, Vol. 2.

Tamaki, K. (2008). The Bernstein-von Mises Theorem for Stationary Processes. *Journal of the Japan Statistical Society* 38(2), 311–323.

Tseng, Y.-L. & Brown, L. D. (1997). Good exact confidence sets for a multivariate normal mean. *Annals of Statistics* 25(5), 2228–2258.

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

Uhlig, H. (2015). Shocks, Sign Restrictions and Identification. Presentation slides, Econometric Society World Congress August 2015.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vinod, H. D. (1978). A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares. *Review of Economics and Statistics* 60(1), 121–131.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Watson, M. W. (1994). Vector autoregressions and cointegration. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of Econometrics*, Vol. IV, Ch. 47, 2843–2915. North-Holland.

Watson, M. W. & Engle, R. F. (1983). Alternative Algorithms for the Estimation of Dynamic Factor, MIMIC and Varying Coefficient Regression Models. *Journal of Econometrics* 23(3), 385–400.

Whittle, P. (1953). The Analysis of Multiple Stationary Time Series. *Journal of the Royal Statistical Society, Series B (Methodological)* 15(1), 125–139.

Xie, X., Kou, S. C. & Brown, L. D. (2012). SURE Estimates for a Heteroscedastic Hierarchical Model. *Journal of the American Statistical Association* 107(500), 1465–1479.