Estimation of Smooth Impulse Response Functions

MAJOR REVISION IN PROGRESS – PLEASE DO NOT CITE

Mikkel Plagborg-Møller*

May 19, 2016

Abstract: 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 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.

Keywords: impulse response function, James-Stein estimator, local projection, shrinkage, test inversion, unbiased risk estimate, uniform inference.

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

*Harvard University, email: plagborg@fas.harvard.edu. This paper is based on Chapter 3 of my PhD dissertation. I am grateful for comments from Isaiah Andrews, Gary Chamberlain, Gita Gopinath, Max Kasy, Michal Kolesar, Eben Lazarus, Daniel Lewis, Adam McCloskey, Anna Mikusheva, Ulrich Müller, Emi Nakamura, Neil Shephard, Jim Stock, and seminar participants at Harvard University. I thank Mark Gertler and Peter Karadi for making their data available online.

[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).

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 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 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 smoothness hypothesis. Consequently, I prove that a version of the SmIRF estimator with

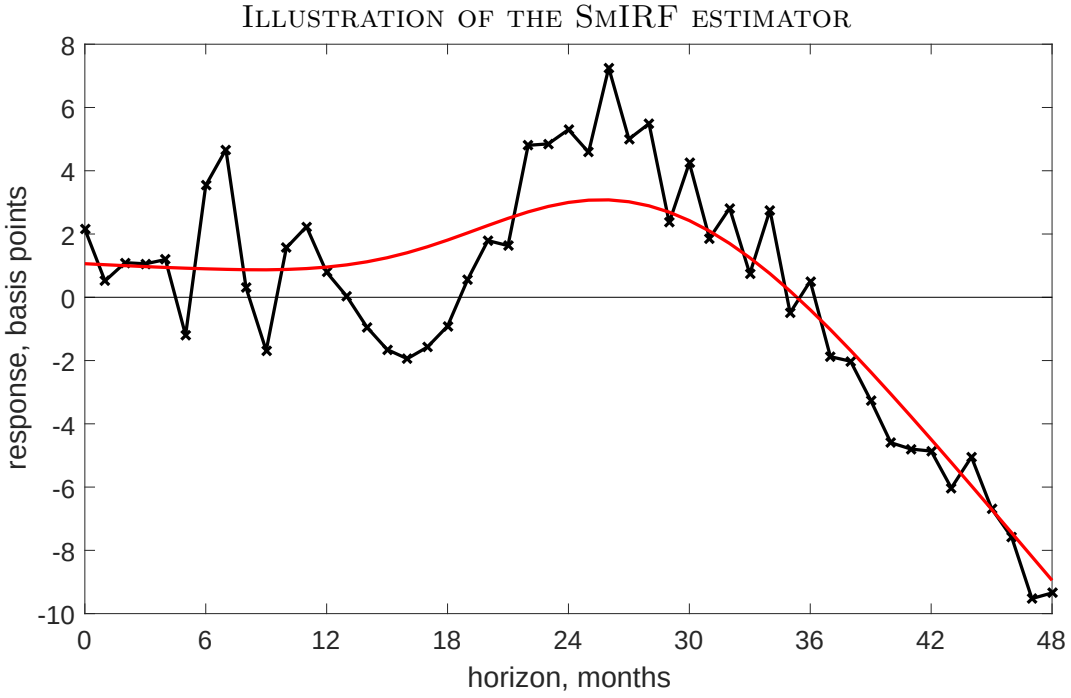

Figure 1: Local projection (jagged line, crosses) and SmIRF (smooth curve) estimators of the IRF of the excess bond premium to a 1-standard-deviation monetary policy shock, monthly U.S. data. Response: Gilchrist & Zakrajšek (2012) excess bond premium. Shock: Gertler & Karadi (2015) monetary policy shock. 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.2.

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

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

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, 2015). 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 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, 2012; 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 developed 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

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

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 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 2 is a user's guide to SmIRF and other shrinkage estimators, the URE, and confidence set construction. Section 3 presents theoretical results on the MSE of shrinkage estimators. Section 4 derives valid shrinkage confidence sets. Section 5 contains a simulation study. Section 6 lists topics for future research. Appendix A defines notation and data, gives technical details, and provides additional simulation results, while proofs are relegated to Appendix B.

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 1 in the Introduction.

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

$$y_{t+i} = \beta_i x_t + \text{controls} + \varepsilon_{t+i|t}, \tag{1}$$

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 (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 1 in the Introduction. If the IRF is estimated from horizon-by-horizon regressions, there is nothing 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]

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

[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.

[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.

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 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.

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

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

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

[6] For local projection (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.

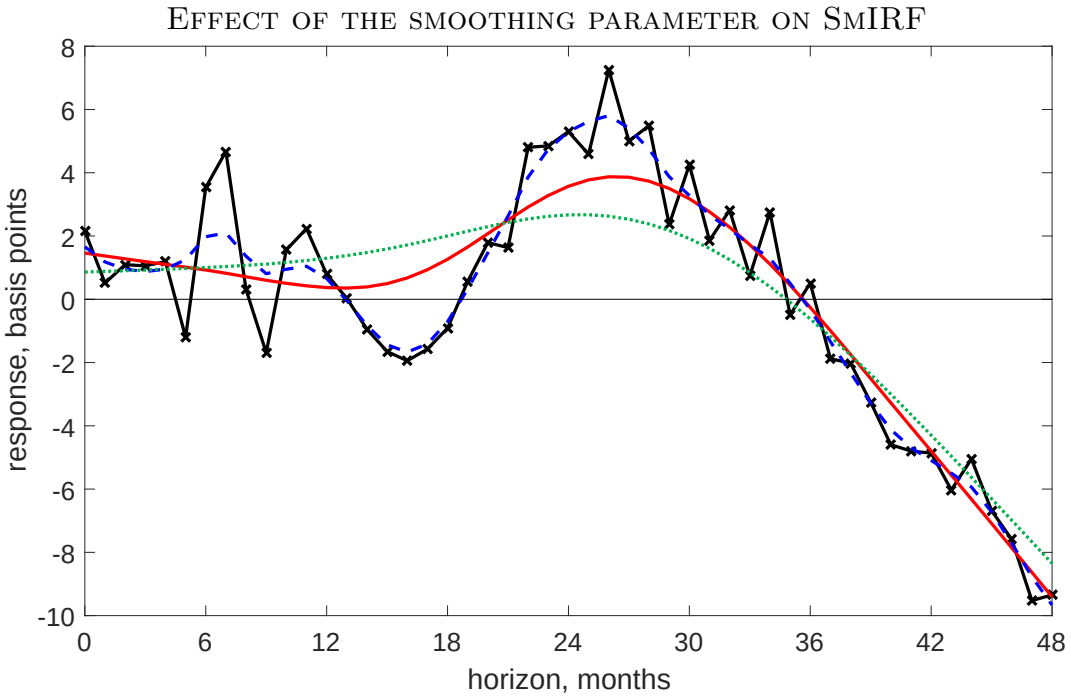

Figure 2: SmIRF estimator for $\lambda = 0$ (solid line, crosses), 2 (dashed), 298 (solid, no crosses), and 2000 (dotted). $\lambda = 298$ is optimal in the sense of Section 2.3. See caption for Figure 1.

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 (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 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 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 2.3 below shows

how to select λ in a data-dependent way to optimally trade off bias and variance.[7]

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

$$\hat{\beta}_{M,W}(\lambda) = \underset{\beta \in \mathbb{R}^n}{\arg\min} \ \|\beta - \hat{\beta}\|_W^2 + \lambda \|M\beta\|^2 = \Theta_{M,W}(\lambda)\hat{\beta}, \tag{3}$$

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

$$\Theta_{M,W}(\lambda) = (I_n + \lambda W^{-1} M'M)^{-1}. \tag{4}$$

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 (2) obtains when $W = I_n$ and

$$M = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{(n-2)\times n}, \tag{5}$$

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) shows.

General shrinkage estimators (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.

[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.

PROJECTION SHRINKAGE CLASS. An analytically convenient subclass of estimators are the projection shrinkage estimators. These estimators are obtained from the general class (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, the projection shrinkage estimator $\hat{\beta}_P(\lambda) := \hat{\beta}_{P,I_n}(\lambda)$ can be written

$$\begin{aligned}\hat{\beta}_P(\lambda) &= \underset{\beta \in \mathbb{R}^n}{\arg\min} \left\{ \|P\beta - P\hat{\beta}\|^2 + \|(I_n - P)\beta - (I_n - P)\hat{\beta}\|^2 + \lambda \|P\beta\|^2 \right\} \\ &= \frac{1}{1+\lambda} P\hat{\beta} + (I_n - P)\hat{\beta}.\end{aligned} \tag{6}$$

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 (6) penalizes deviations of the IRF from functions of the form $l(i)$ (for some 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 the initial IRF estimate is shrunk towards a quadratic function. This procedure does not produce as smooth-looking IRFs as the basic SmIRF estimator (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.

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

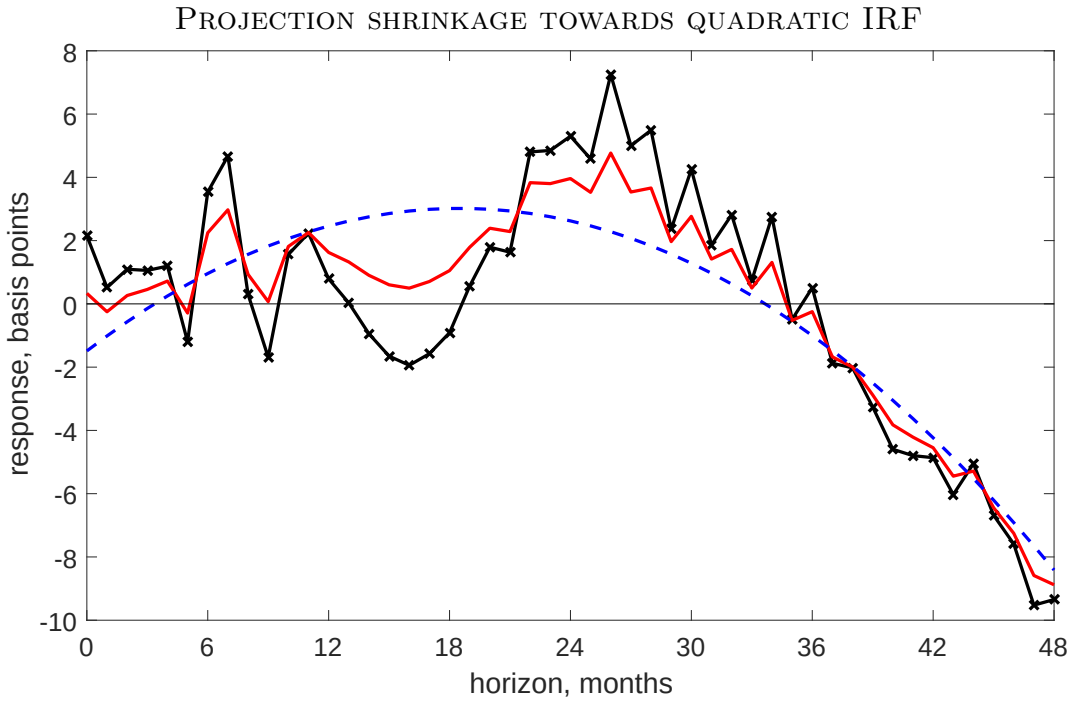

Figure 3: Initial IRF estimate (jagged line, crosses), best-fitting quadratic IRF (dashed), and projection shrinkage estimator with $\lambda = 1$ (solid, no crosses). See caption for Figure 1.

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 3.1, $\hat{\beta}$ 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 3.1:

$$\hat{R}_{M,W,\tilde{W}}(\lambda) = T\|\hat{\beta}_{M,W}(\lambda) - \hat{\beta}\|^2_{\tilde{W}} + 2\,\mathrm{tr}\left\{\tilde{W}\Theta_{M,W}(\lambda)\hat{\Sigma}\right\}, \tag{7}$$

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 (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 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.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 4 plots the URE corresponding to the projection shrinkage estimator in Figure 3. It is straight-forward to compute the URE (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 5 indicate that this estimator 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 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

$$\hat{\beta}_{P,I_n,I_n} = \left(1 - \frac{\mathrm{tr}(\hat{\Sigma}_P)}{T\|P\hat{\beta}\|^2}\right)_+ P\hat{\beta} + (I_n - P)\hat{\beta} = \hat{\beta} - \min\left\{\frac{\mathrm{tr}(\hat{\Sigma}_P)}{T\|P\hat{\beta}\|^2}, 1\right\} P\hat{\beta}, \tag{8}$$

[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).

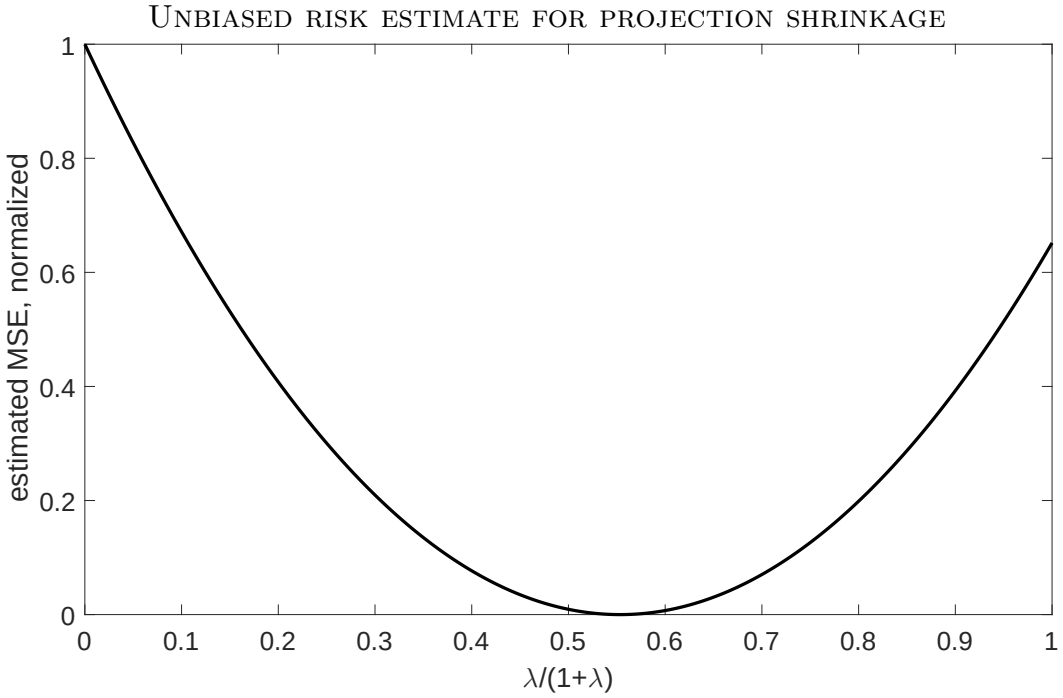

Figure 4: URE criterion to optimally select the smoothing parameter λ for the projection shrinkage estimator in Figure 3. The horizontal axis plots $\lambda/(1+\lambda)$, not λ. MSE estimates on the vertical axis are normalized to $[0,1]$. HAC: Newey-West, 24 lags. See caption for Figure 1.

where $\hat{\Sigma}_P = P\hat{\Sigma}P$ and $x_+ = \max\{x, 0\}$ for any $x \in \mathbb{R}$. Expression (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 3.2 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]

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-

[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).

dependent shrinkage parameter.[10] Here I provide recipes for constructing joint and pointwise confidence bands. As discussed in detail in Sections 4 and 5, the proposed 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]

$$\hat{\theta}_{M,W,\tilde{W}}(\eta, \Sigma) = \Theta_{M,W}(\hat{\lambda}_{M,W,\tilde{W}}(\eta, \Sigma))\eta, \tag{9}$$

$$\hat{\lambda}_{M,W,\tilde{W}}(\eta, \Sigma) = \operatorname*{arg\,min}_{\lambda \geq 0} \left(\|\{\Theta_{M,W}(\lambda) - I_n\}\eta\|^2_{\tilde{W}} + 2 \operatorname{tr}\{\tilde{W}\Theta_{M,W}(\lambda)\Sigma\} \right). \tag{10}$$

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

$$\{s'\hat{\theta}_{M,W,\tilde{W}}(\zeta u + \theta, \Sigma) - s'\theta\}^2, \tag{11}$$

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 estimators because the minimizer (10) is available in closed form, cf. Appendix A.3.

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

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

[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.

[11] If the minimum (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.

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 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 (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 $\hat{\mathcal{I}}_{s,1-\delta}$, 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 5 and 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 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 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

[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 4.

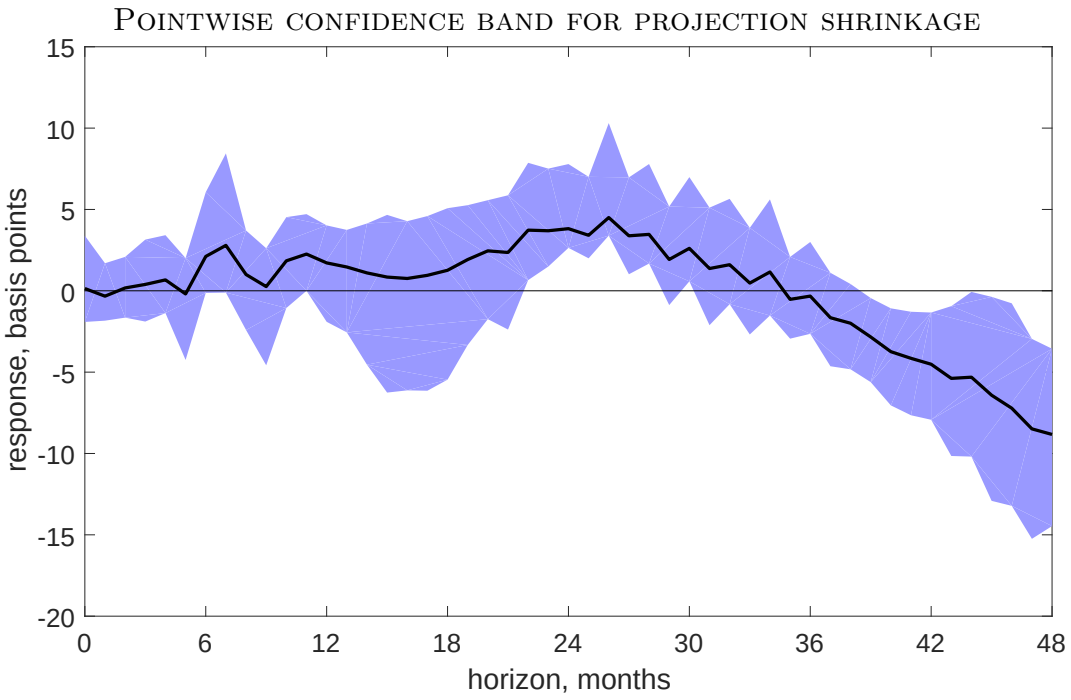

Figure 5: Projection shrinkage point estimate (thick line) and 90% pointwise confidence band $\hat{\mathcal{C}}_{s,0.9,1-\delta}$ for $\delta = 0.01$ (shaded). HAC: Newey-West, 24 lags. See caption for Figure 1.

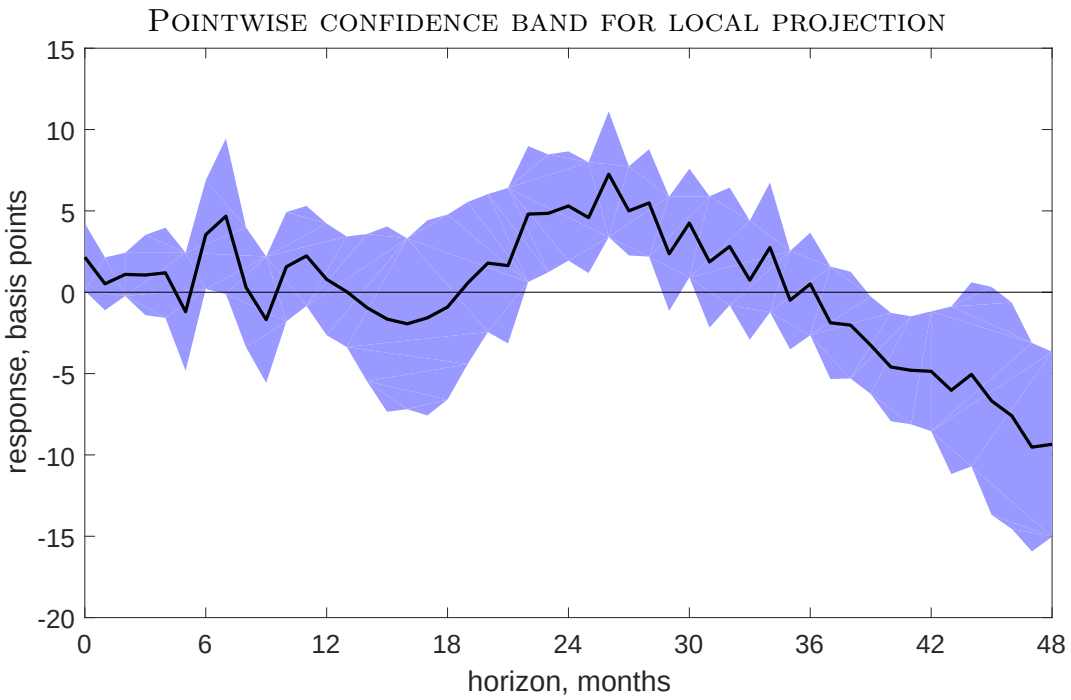

Figure 6: Local projection point estimate (thick line) and 90% pointwise confidence band (shaded). HAC: Newey-West, 24 lags. See caption for Figure 1.

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 $\theta = (\theta_0, \dots, \theta_{n-1})' \in \mathbb{R}^n$ 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 (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, define the joint level $1-\alpha$ shrinkage confidence set[13]

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

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 $\beta = (\beta_0, \dots, \beta_{n-1})'$ 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 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 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

[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.

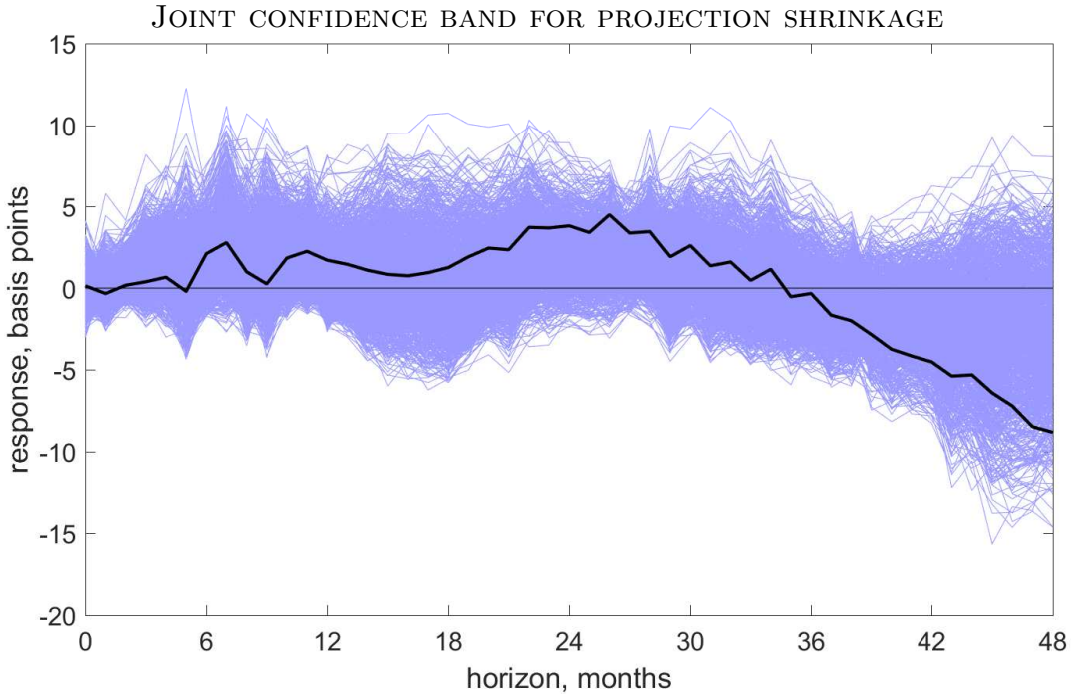

Figure 7: Projection shrinkage point estimate (thick line) and draws from 68% joint confidence band (thin lines). 10,000 t-distributed proposal draws, 5 d.f.; 1,251 draws accepted. HAC: Newey-West, 24 lags. See caption for Figure 1.

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 asymptotically normal, with consistently estimable asymptotic variance.

Assumption 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

3.1 Unbiased risk estimate

I now justify the name "unbiased risk estimate" by proving that the URE criterion (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), 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 1. *Let Assumption 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*

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

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

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

any finite T. The assumption can probably be relaxed at the expense of a more complicated statement in Proposition 1, and a trimmed-risk statement in Proposition 2 similar to Hansen (2016b, Thm. 1).

where I have abused notation slightly relative to definition (6) (note that here I use τ instead of λ for the argument). The optimal projection shrinkage estimator (8) is a member of the above class with $\tau = \text{tr}(\hat{\Sigma}_P)$. 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 2 (Oman, 1982; Hansen, 2016b)**.** *Let Assumption 1 hold. Assume* $\sqrt{T}P\beta_T^\dagger \to h \in \mathbb{R}^n$. *Let* $\hat{\tau} \geq 0$ *be a scalar random variable satisfying* $\hat{\tau} \xrightarrow{p} \tau \leq 2(\text{tr}(\Sigma) - 2\rho(\Sigma))$ *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*

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

The result shows that if $\text{plim}_{T\to\infty}\hat{\tau} \leq 2(\text{tr}(\Sigma_P) - 2\rho(\Sigma_P))$, 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 $\text{span}(I_n - P)$. 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 $\tau = \text{tr}(\Sigma_P) - 2\rho(\Sigma_P)$; 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 1. *Let Assumption 1 hold, and assume* $\sqrt{T}P\beta_T^\dagger \to h \in \mathbb{R}^n$. *If* $\text{tr}(\Sigma_P) > 4\rho(\Sigma_P)$, *the URE-minimizing projection shrinkage estimator* $\hat{\beta}_P(\hat{\tau})$ *with* $\hat{\tau} = \text{tr}(\hat{\Sigma}_P)$ *has smaller limiting MSE than the initial IRF estimator* $\hat{\beta}$, *uniformly in* h.

The sufficient condition $\text{tr}(\Sigma_P) > 4\rho(\Sigma_P)$ in Corollary 1 requires that the number of response horizons of interest is sufficiently high. Since $\text{tr}(\Sigma_P) \leq \text{rk}(P)\rho(\Sigma_P)$, a necessary condition for $\text{tr}(\Sigma_P) > 4\rho(\Sigma_P)$ is that $\text{rk}(P) > 4$. If the projection shrinkage estimator is used to shrink the initial IRF estimate towards a polynomial of order p, as explained in Section 2.2, then $\text{rk}(P) = n - p$, 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 $\text{tr}(\Sigma_P)$ versus $\rho(\Sigma_P)$.

4 Confidence sets

Here I develop novel methods for computing joint and marginal confidence sets based on the shrinkage estimators in Section 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.

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 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 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 2.2 and 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 (4), and also

$$\hat{\lambda}_{M,W,\tilde{W}} = \operatorname*{arg\,min}_{\lambda \geq 0} \left(\|\hat{\theta}_{M,W}(\lambda) - \hat{\theta}\|_{\tilde{W}}^2 + 2\,\mathrm{tr}\{\tilde{W}\Theta_{M,W}(\lambda)\Sigma\} \right). \tag{12}$$

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

$$\hat{\theta}_{P,I_n,I_n} = \hat{\theta} - \min\left\{\frac{\operatorname{tr}(P\Sigma)}{\|P\hat{\theta}\|^2}, 1\right\} P\hat{\theta}, \tag{13}$$

which is the analogue of the optimal projection shrinkage estimator (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 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 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 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 (9) (this coincides with the definition in Section 2.4 for the choice of $g(\cdot,\cdot)$ used there). Then, by definition,

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

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

The proposed shrinkage confidence set has several attractive features, although I have

[15] If the minimum in (12) is attained at $\lambda = \infty$, $\hat{\theta}_{M,W,\tilde{W}}$ is defined as a limit, cf. Footnote 11.

[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.

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 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 (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 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 2.4. Since the jointly Gaussian random variables $s'\hat{\theta}$ and $\hat{\nu}^\theta$ are orthogonal and 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,

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

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

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

[17] The argument is similar to the proof of Proposition 3 below.

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

The first equality uses independence of $s'\hat{\theta}$ and $\hat{\nu}^\theta$, the definition (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 2.4. Then

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

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

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

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 (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 2.4.

[19] In the case of projection shrinkage (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 $c_{1-\delta} \ge \|s\|\sqrt{\operatorname{tr}(P\Sigma)}$. This follows from $\|\hat{\theta}_{P,I_n,I_n} - \hat{\theta}\| \le \sqrt{\operatorname{tr}(P\Sigma)}$, cf. Appendix A.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 1 in Section 3.

Assumption 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* $\underline{c}, \bar{c} > 0$. *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 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 below, define the shrinkage estimator

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

Assumption 3. $f : \mathbb{R}_+ \times \mathbb{S}^n_+ \to \mathbb{R}$ *is continuous, and* $\lim_{x\to\infty} x f(x^2, \Sigma) \to 0$ *for all* $\Sigma \in \mathbb{S}^n_+$.

The choice $f(x, \Sigma) = \min\{\text{tr}(P\Sigma)/x, 1\}$ satisfies Assumption 3 and yields the optimal projection shrinkage estimator (8).

JOINT CONFIDENCE SETS. The next result states that the joint confidence set in Section 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 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. *Let Assumptions 2 and 3 hold, and assume that $g(\cdot,\cdot)$ is continuous. Then*

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\left(\hat{S}(\beta) \le q_{1-\alpha}(\sqrt{T}\beta, \hat{\Sigma})\right) = 1-\alpha. \tag{14}$$

Thus, for sufficiently large sample sizes, the worst-case *finite-sample* coverage probability of the confidence region $\{\beta \in \mathbb{R}^n\colon \hat{S}(\beta) \le q_{1-\alpha}(\sqrt{T}\beta, \hat{\Sigma})\}$ 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 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 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 4. *Let Assumptions 2 and 3 hold. Then*

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\left(\hat{S}_s(s'\beta) \le q_{s,1-\alpha}\left(\sqrt{T}(\hat{\zeta}(s'\beta) + \hat{\nu}), \hat{\Sigma}\right)\right) = 1-\alpha. \tag{15}$$

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

$$\liminf_{T\to\infty} \inf_{(\beta,\Sigma,\gamma)\in\mathbb{R}^n\times\mathcal{S}\times\Gamma} Prob_{F_T(\beta,\Sigma,\gamma)}\Big(\hat{S}_{s,W}(s'\beta) \le \hat{c}^2_{1-\delta}, \\ \hat{S}_s(s'\beta) \le \tilde{q}_{s,\frac{1-\alpha}{1-\delta}}\left(\sqrt{T}(\hat{\zeta}(s'\beta) + \hat{\nu}), \hat{\Sigma}, \hat{c}_{1-\delta}\right)\Big) = 1-\alpha. \tag{16}$$

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 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 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.

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 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 (13), where P is the projection matrix that shrinks towards a second-degree polynomial, cf. Section 2.2. Appendix A.5 contains simulation results for the SmIRF estimator. These are qualitatively similar to the projection shrinkage results.

RESULTS. Table 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

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 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, $\alpha = 0.1$, $\delta = 0.01$.

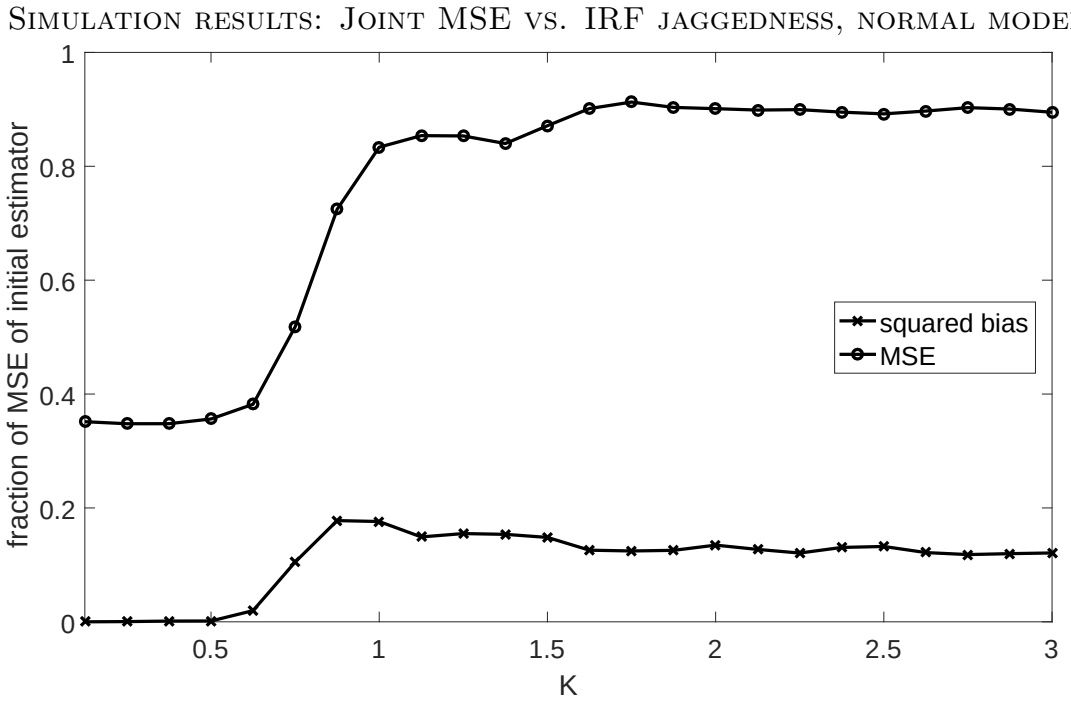

Figure 8: Joint MSE and squared bias of quadratic projection shrinkage estimator as functions of the jaggedness of the true IRF, finite-sample normal model. Horizontal axis shows $K = 0.125, 0.25, \ldots, 3$, with true IRF given by $\theta_i^\dagger = \sin \frac{2\pi K i}{n-1}$. Vertical axis units normalized by joint MSE of $\hat{\theta}$, for each K. Other DGP parameters: $n = 25$, $\kappa = 0.5$, $\sigma_0 = 0.2$, $\varphi = 3$.

shrinkage estimator is only substantial in the DGPs with $K = 1$ or 2, i.e., for very non-quadratic IRFs.[20] Figure 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.

Although shrinkage is not designed to improve individual impulse response estimates, Table 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

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

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 1 shows that joint and marginal shrinkage confidence sets are competitive with 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 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 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.

- 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 correspond-

[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 $\theta^\dagger$.

ing 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.

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 2.

DGP AND ESTIMATORS. I generate data from a VAR calibrated to the Gertler & Karadi (2015) data described in Appendix A.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 $A_1, A_2 \in \mathbb{R}^{4\times4}$ and $b \in \mathbb{R}^4$ are obtained by least-squares regression on the calibration data, with $C \in \mathbb{R}^{4\times4}$ 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.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 (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

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

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 2: Simulation results for quadratic projection shrinkage, time series regression on VAR data. Columns 5–6, 9–12: See caption for Table 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.

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 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 markedly when the shock x_t is t-distributed. The last rows of Table 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.

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).

A Technical appendix

A.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.2 Data and empirical specification

The data used in Sections 1, 2 and 5 is from the replication files for Gertler & Karadi (2015), which are available on the American Economic Association Website.[25] 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 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 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 $\|\beta - \hat{\beta}\|^2 =$

[25] https://www.aeaweb.org/articles?id=10.1257/mac.20130329. Downloaded April 11, 2016.

$\|P(\beta-\hat\beta)\|^2+\|(I_n-P)(\beta-\hat\beta)\|^2$, implying that $\hat\beta_P(\lambda)$ defined in (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 (7) can be written

$$\begin{aligned}\hat R_{P,I_n,I_n}(\lambda)&=T\|P(\hat\beta_P(\lambda)-\hat\beta)\|^2+T\|(I_n-P)(\hat\beta_P(\lambda)-\hat\beta)\|^2\\&\quad+2\operatorname{tr}\{P\Theta_P(\lambda)\hat\Sigma\}+2\operatorname{tr}\{(I_n-P)\Theta_P(\lambda)\hat\Sigma\}\\&=\left(\frac{\lambda}{1+\lambda}\right)^2T\|P\hat\beta\|^2+\left(1-\frac{\lambda}{1+\lambda}\right)\operatorname{tr}(\hat\Sigma_P)+\text{constant},\end{aligned}$$

where $\hat\Sigma_P=P\hat\Sigma P$. The value of $\lambda\ge0$ that minimizes the above quadratic form satisfies

$$\frac{\hat\lambda_P}{1+\hat\lambda_P}=\min\left\{\frac{\operatorname{tr}(\hat\Sigma_P)}{T\|P\hat\beta\|^2},1\right\}.$$

Quantile simulation. For projection shrinkage, the simulation of the quantile functions in Section 2.4 simplifies drastically, since $\hat\theta_{M,W,\tilde W}(\eta,\Sigma)$ defined in (9) reduces to

$$\hat\theta_{P,I_n,I_n}(\eta,\Sigma)=\eta-\min\left\{\frac{\operatorname{tr}(P\Sigma)}{\|P\eta\|^2},1\right\}P\eta.$$

Conditional quantile bound. In the notation of Section 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 2, $\|\hat\theta_{P,I_n,I_n}(\eta,\Sigma)-\eta\|\le\sqrt{\operatorname{tr}(P\Sigma)}$ for all η,Σ. Using $s'\zeta=1$, it follows that

$$|s'\hat\theta_{P,I_n,I_n}(\zeta u+\theta,\Sigma)-s'\theta|\le|s'(\zeta u+\theta)-s'\theta|+\|s\|\sqrt{\operatorname{tr}(P\Sigma)}=|u|+\|s\|\sqrt{\operatorname{tr}(P\Sigma)},$$

implying $\sqrt{q_{s,1-\alpha,P,I_n,I_n}(\theta,\Sigma)}\le\sqrt{(s'\Sigma s)z_{1,1-\alpha}}+\|s\|\sqrt{\operatorname{tr}(P\Sigma)}$ for all θ,Σ.

A.4 URE and the bias-variance tradeoff

The URE (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:

$$\begin{aligned} R_{M,I_n,I_n}(\lambda) &= TE[\hat{\beta}_{M,I_n}(\lambda) - \beta^\dagger]' E[\hat{\beta}_{M,I_n}(\lambda) - \beta^\dagger] \\ &\quad + \operatorname{tr}\left\{TE[(\hat{\beta}_{M,I_n}(\lambda) - E[\hat{\beta}_{M,I_n}(\lambda)])(\hat{\beta}_{M,I_n}(\lambda) - E[\hat{\beta}_{M,I_n}(\lambda)])']\right\} \\ &= T \operatorname{tr}\left\{[I_n - \Theta_{M,I_n}(\lambda)]^2 \beta^\dagger \beta^{\dagger\prime}\right\} + \operatorname{tr}\left\{\Theta_{M,I_n}(\lambda)^2 \Sigma\right\}. \end{aligned}$$

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.5 Supplemental simulation results

Supplementing the analysis in Section 5, I provide simulation results for the SmIRF estimator and plot the true IRF in the VAR DGP.

Table 3 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 4.1, where M is the $(n-2) \times n$ second difference matrix (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 (2) is more lenient toward a sine curve than is the quadratic projection penalty in (6).

Table 4 illustrates the performance of the SmIRF-based confidence sets for three of the DGPs considered in Table 1 in Section 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 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

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 3: Simulation results for MSE of SmIRF estimator. See caption for Table 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 4: Simulation results for SmIRF confidence sets. See caption for Table 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".

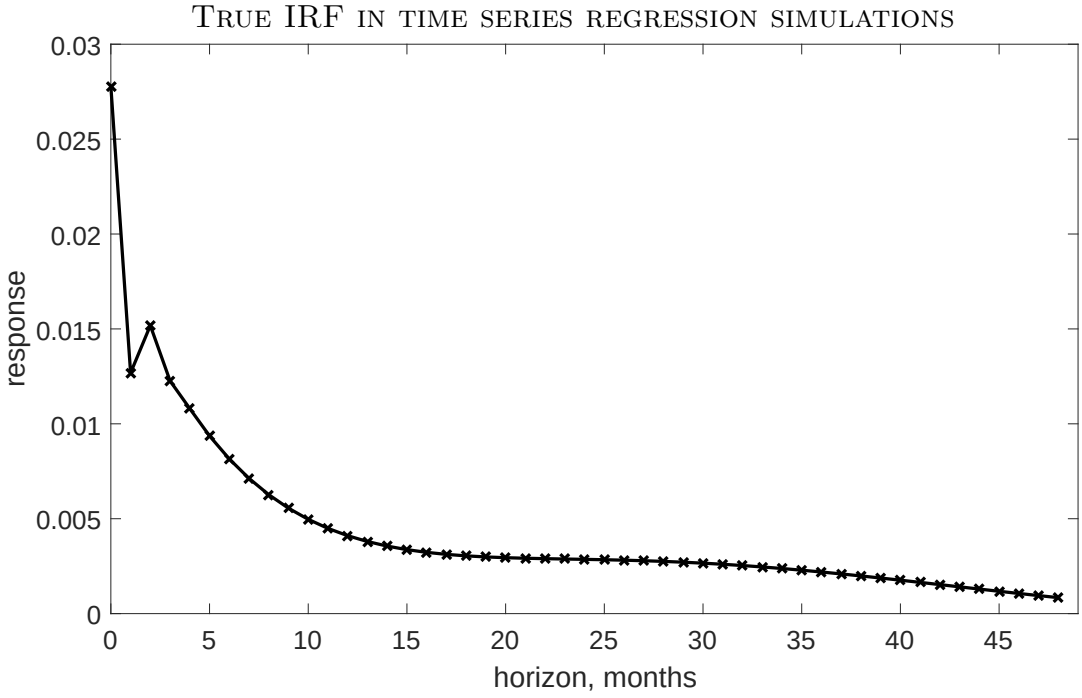

Figure 9: True VAR-implied IRF in time series regression simulations.

shrinkage confidence sets.

Figure 9 shows the true IRF implied by the data generating VAR(2) model used in the simulations in Section 5.2.

B Proofs

B.1 Proof of Proposition 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:

$$\hat{C} = 2\,\mathrm{tr}\left\{\tilde{W}\Theta_{M,W}(\lambda)(I_n - P_M)[T\hat{\beta}(\hat{\beta} - \beta^\dagger)' - \hat{\Sigma}]\right\} + T\|\beta^\dagger\|_{\tilde{W}}^2 - T\|\hat{\beta}\|_{\tilde{W}}^2.$$

We have

$$\left|R_{M,W,\tilde{W}}(\lambda) - E\left(\hat{R}_{M,W,\tilde{W}}(\lambda) + \hat{C}\right)\right| = \left|\operatorname{tr}\left\{\tilde{W}\Theta_{M,W}(\lambda)P_M E[T\hat{\beta}(\hat{\beta}-\beta^\dagger)' - \hat{\Sigma}]\right\}\right|$$
$$\leq \|\tilde{W}\|\|\Theta_{M,W}(\lambda)\| \left\|E[TP_M\hat{\beta}(\hat{\beta}-\beta^\dagger)' - P_M\hat{\Sigma}]\right\|.$$

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 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 1; hence, its expectation converges to $E(hU') = 0$. □

B.2 Proof of Proposition 2

I proceed in three steps.

Step 1. By the continuous mapping theorem,

$$\sqrt{T}(\hat{\beta}_P(\hat{\tau}) - \beta^\dagger) = \sqrt{T}(\hat{\beta}-\beta^\dagger)$$
$$- \min\left\{\frac{\hat{\tau}}{\|P\sqrt{T}(\hat{\beta}-\beta^\dagger) + \sqrt{T}P\beta^\dagger\|^2}, 1\right\}\{P\sqrt{T}(\hat{\beta}-\beta^\dagger) + \sqrt{T}P\beta^\dagger\}$$
$$\xrightarrow{d} V,$$

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

$$\begin{aligned}
E\|\tilde{V}\|^2 - E\|V\|^2 &= E\|P\tilde{V}\|^2 + E\|PV\|^2 \\
&= E\|P\tilde{V}+h\|^2 - E\|PV+h\|^2 - 2E[h'P(\tilde{V}-V)] \\
&\geq -2E[h'P(\tilde{V}-V)] \\
&= 2E[(B-1)h'(PU+h) \mid B>1]\Pr(B>1),
\end{aligned}$$

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 $E\|U\|^2 = \operatorname{tr}(\Sigma)$, we have

$$E\|\tilde{V}\|^2 = \operatorname{tr}(\Sigma) + \tau^2 E\left(\frac{1}{\|PU+h\|^2}\right) - 2\tau E[\eta(U+h)'PU], \tag{17}$$

where $\eta(x) = x/\|Px\|^2$ for $x \in \mathbb{R}^n$. By Stein's Lemma (Hansen, 2016b, Lem. 2),

$$\begin{aligned}
E[\eta(U+h)'PU] &= E\left[\operatorname{tr}\left(\frac{\partial}{\partial x}\eta(U+h)'P\Sigma\right)\right] \\
&= E\left[\operatorname{tr}\left\{\left(\frac{1}{\|PU+h\|^2}I_n - \frac{2}{\|PU+h\|^4}P(U+h)(U+h)'\right)P\Sigma\right\}\right] \\
&= E\left[\frac{\operatorname{tr}(\Sigma_P)}{\|PU+h\|^2} - \frac{2\operatorname{tr}\{(PU+h)'\Sigma_P(PU+h)\}}{\|PU+h\|^4}\right] \\
&\geq E\left[\frac{\operatorname{tr}(\Sigma_P)}{\|PU+h\|^2} - \frac{2\rho(\Sigma_P)\|PU+h\|^2}{\|PU+h\|^4}\right] \\
&= E\left[\frac{\operatorname{tr}(\Sigma_P) - 2\rho(\Sigma_P)}{\|PU+h\|^2}\right].
\end{aligned}$$

Inserting this into equation (17), we obtain

$$\begin{aligned}
E\|\tilde{V}\|^2 &\leq \operatorname{tr}(\Sigma) - \tau E\left(\frac{2(\operatorname{tr}(\Sigma_P) - 2\rho(\Sigma_P)) - \tau}{\|PU+h\|^2}\right) \\
&\leq \operatorname{tr}(\Sigma) - \tau\frac{2(\operatorname{tr}(\Sigma_P) - 2\rho(\Sigma_P)) - \tau}{E\|PU+h\|^2},
\end{aligned}$$

where the last line uses Jensen's inequality and the assumption $0 \leq \tau \leq 2(\operatorname{tr}(\Sigma_P) - 2\rho(\Sigma_P))$. Finally, observe that $E\|PU+h\|^2 = \operatorname{tr}(\Sigma_P) + \|h\|^2$. □

B.3 Proof of Corollary 1

I verify the conditions of Proposition 2. With $\hat{\tau} = \operatorname{tr}(\hat{\Sigma}_P)$, Assumption 1 implies $\tau = \operatorname{plim}_{T\to\infty}\hat{\tau} = \operatorname{tr}(\Sigma_P)$. It remains to show that $\{T\|\hat{\beta} - \beta_T^\dagger\|^2 + \hat{\tau}\}_{T\geq 1}$ is uniformly integrable, which follows from Assumption 1 and $\hat{\tau} = \operatorname{tr}(\hat{\Sigma}P) \leq \|\hat{\Sigma}P\| \leq \|\hat{\Sigma}\|\rho(P) = \|\hat{\Sigma}\|$. □

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 (14) equals

$$\lim_{T\to\infty} \text{Prob}_{F_{k_T}(\beta_{k_T},\Sigma_{k_T},\gamma_{k_T})}\left(\hat{S}(\beta_{k_T}) \leq q_{1-\alpha}(\sqrt{k_T}\beta_{k_T},\hat{\Sigma})\right). \tag{18}$$

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

$$\lim_{T\to\infty} \sqrt{\tilde{k}_T}\tilde{\beta}_{i,\tilde{k}_T} = h_i. \tag{19}$$

In case (b), there exists a further subsequence $\{\tilde{k}_T\}_{T\geq 1}$ of $\{k_T\}_{T\geq 1}$ such that

$$\lim_{T\to\infty} \sqrt{\tilde{k}_T}\tilde{\beta}_{i,\tilde{k}_T} \in \{-\infty,\infty\}. \tag{20}$$

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

$$\lim_{T\to\infty} \Sigma_{\tilde{k}_T} = \tilde{\Sigma}. \tag{21}$$

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 (19) or (20) holds, and such that (21) holds. Since any subsequence of a convergent sequence converges to the same limit, expression (18) equals

$$\lim_{T\to\infty} \text{Prob}_{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}\left(\hat{S}(\beta_{\tilde{k}_T}) \leq q_{1-\alpha}(\sqrt{\tilde{k}_T}\beta_{\tilde{k}_T},\hat{\Sigma})\right). \tag{22}$$

Write

$$\begin{aligned}\hat{S}(\beta_{\tilde{k}_T}) = g\Big(&\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T}) \\ &- f(\|P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2,\hat{\Sigma})\Big\{P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\Big\},\hat{\Sigma}\Big).\end{aligned} \tag{23}$$

There are now two cases to consider.

Case I. Suppose first that (20) holds for some i. Then $\lim_{T\to\infty} \|\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\| = \infty$, and by Assumption 2 and equation (21),

$$\|P\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T}) + \sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\| \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{p} \infty.$$

Hence, by Assumptions 2 and 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 2 on expression (23),

$$\hat{S}(\beta_{\tilde{k}_T}) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{d} g(U, \tilde{\Sigma}), \quad U \sim N(0, \tilde{\Sigma}).$$

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 (22) – and thus the left-hand side of (14) – equals $1-\alpha$.

Case II. Suppose instead that (19) holds for all $i = 1, \ldots, n$. Let $h = (h_1, \ldots, h_n)'$, and note that $Ph = h$. By expression (23) and Assumptions 2 and 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 (22) – and thus the left-hand side of (14) – equals $1-\alpha$. □

B.4 Proof of Proposition 4

I focus on the proof of (15), although I remark at the end how (16) can be obtained from similar arguments. Using the same steps as in the proof of Proposition 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 (15) equals

$$\lim_{T\to\infty} \text{Prob}_{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}\left(\hat{S}_s(s'\beta_{\tilde{k}_T}) \leq q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T})+\hat{\nu}), \hat{\Sigma}\Big)\right), \tag{24}$$

and – for all $i = 1, \ldots, n$ – either (19) or (20) holds, and moreover (21) holds, where $\tilde{\beta}_{i,\tilde{k}_T}$ is the i-th element of $P\tilde{\beta}_{\tilde{k}_T}$. Write

$$\begin{aligned}\hat{S}_s(s'\beta_{\tilde{k}_T}) = \Bigg(&\sqrt{\tilde{k}_T}s'(\hat{\beta}-\beta_{\tilde{k}_T}) \\ &- f(\|P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}\|^2, \hat{\Sigma})\Big\{s'P\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T})+\sqrt{\tilde{k}_T}s'\tilde{\beta}_{\tilde{k}_T}\Big\}\Bigg)^2.\end{aligned} \tag{25}$$

There are two cases to consider.

Case I. Suppose that (20) holds for some $i = 1, \ldots, n$. As in the proof of Proposition 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 (25),

$$\hat{S}_s(s'\beta_{\tilde{k}_T}) \xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{d} (s'\tilde{\Sigma}s)\chi^2(1).$$

Moreover,

$$\begin{aligned}q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T})+\hat{\nu}), \hat{\Sigma}\Big) &= q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{P}\beta_{\tilde{k}_T}+(I_n-\hat{P})\hat{\beta}), \hat{\Sigma}\Big) \\ &= q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}\beta_{\tilde{k}_T}+(I_n-\hat{P})\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T}), \hat{\Sigma}\Big) \\ &= q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T}+P(I_n-\hat{P})\sqrt{\tilde{k}_T}(\hat{\beta}-\beta_{\tilde{k}_T}), \hat{\Sigma}\Big) \\ &\xrightarrow[F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})]{p} (s'\tilde{\Sigma}s)z_{1,1-\alpha},\end{aligned} \tag{26}$$

since $q_{s,1-\alpha}(\theta,\Sigma) = q_{s,1-\alpha}(P\theta,\Sigma)$ for all θ,Σ, and

$$\|\sqrt{\tilde{k}_T}\tilde{\beta}_{\tilde{k}_T} + P(I_n - \hat{P})\sqrt{\tilde{k}_T}(\hat{\beta} - \beta_{\tilde{k}_T})\| \underset{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}{\xrightarrow{p}} \infty.$$

The above displays imply that (24) – and thus the left-hand side of (15) – equals $1-\alpha$.

Case II. Suppose now that (19) holds for all $i = 1,\ldots,n$. Let $h = (h_1,\ldots,h_n)'$, and note that $Ph = h$. Then

$$\hat{S}_s(s'\beta_{\tilde{k}_T}) \underset{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}{\xrightarrow{d}} V,$$

where

$$V = \Big\{s'U - f(\|PU+h\|^2,\tilde{\Sigma})(s'PU + s'h)\Big\}^2, \quad U \sim N(0,\tilde{\Sigma}). \tag{27}$$

Jointly with the above convergence, expression (26) yields

$$q_{s,1-\alpha}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T}) + \hat{\nu}),\hat{\Sigma}\Big) \underset{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}{\xrightarrow{d}} q_{s,1-\alpha}\Big(h + (I_n - \tilde{P})U,\tilde{\Sigma}\Big),$$

where $\tilde{P} = \tilde{\zeta}s'$, $\tilde{\zeta} = (s'\tilde{\Sigma}s)^{-1}\tilde{\Sigma}s$. I will have shown that (24) – and thus the left-hand side of (15) – equals $1-\alpha$ if I show that

$$\text{Prob}\Big(V \le q_{s,1-\alpha}\Big(h + (I_n - \tilde{P})U,\tilde{\Sigma}\Big)\Big) = 1-\alpha. \tag{28}$$

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,

$$\begin{aligned}
&\text{Prob}\Big(V \le q_{s,1-\alpha}\Big(h + (I_n - \tilde{P})U,\tilde{\Sigma}\Big)\,\Big|\,(I_n - \tilde{P})U = \nu\Big) \\
&\quad = \text{Prob}\Big(\Big\{(s'U) - f(\|P\{\tilde{\zeta}(s'U) + h + \nu\}\|^2,\tilde{\Sigma})s'P\{\tilde{\zeta}(s'U) + h + \nu\}\Big\}^2 \\
&\qquad \le q_{s,1-\alpha}\Big(h + \nu,\tilde{\Sigma}\Big)\Big)
\end{aligned}$$

$$= 1 - \alpha,$$

where the last equality holds by definition of $q_{s,1-\alpha}(\cdot,\cdot)$. The above conditional result implies the unconditional statement (28).

PROOF OF (16). As in the above proof of (15), pick a subsequence of parameters such that the left-hand side of (16) equals

$$\lim_T \text{Prob}_{F_{\tilde{k}_T}(\beta_{\tilde{k}_T},\Sigma_{\tilde{k}_T},\gamma_{\tilde{k}_T})}\Big(\hat{S}_{s,W}(s'\beta_{\tilde{k}_T}) \le \hat{c}^2_{1-\delta}, \hat{S}_s(s'\beta_{\tilde{k}_T}) \le \tilde{q}_{s,\frac{1-\alpha}{1-\delta}}\Big(\sqrt{\tilde{k}_T}(\hat{\zeta}(s'\beta_{\tilde{k}_T})+\hat{\nu}), \hat{\Sigma}, \hat{c}_{1-\delta}\Big)\Big). \tag{29}$$

As above, there are two cases. In "Case I", the limit (29) 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 (29) equals

$$\text{Prob}\left((s'U)^2 \le c^2_{1-\delta},\ V \le \tilde{q}_{s,\frac{1-\alpha}{1-\delta}}\Big(h + (I_n - \tilde{P})U, \tilde{\Sigma}, c_{1-\delta}\Big)\right), \tag{30}$$

where $U \sim N(0, \tilde{\Sigma})$ and V is given by (27). The variables $s'U$ and $(I_n - \tilde{P})U$ are independent. Hence, conditional on $(I_n - \tilde{P})U = \nu$, the event in (30) 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 (30) also equals $1 - \alpha$. □

References

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. & 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.

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.

Bock, M. E. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Annals of Statistics* 3(1), 209–218.

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.

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.

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.

Cochrane, J. H. & Piazzesi, M. (2002). The Fed and Interest Rates – A High-Frequency Identification. *American Economic Review* 92(2), 90–95.

Fessler, P. & Kasy, M. (2016). How to use economic theory to improve estimators, with an application to labor demand and wage inequality. Working paper.

Gertler, M. & Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics* 7(1), 44–76.

Giannone, D., Lenza, M. & Primiceri, G. E. (2015). Prior Selection for Vector Autoregressions. *Review of Economics and Statistics* 97(2), 436–451.

Gilchrist, S. & Zakrajšek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *American Economic Review* 102(4), 1692–1720.

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.

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.

Imbens, G. W. & Kolesár, M. (2016). Robust Standard Errors in Small Samples: Some Practical Advice. *Review of Economics and Statistics*. Forthcoming.

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

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: Contributions to the Theory of Statistics, 361–379. University of California Press.

Jordà, Ò. (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.

Leeb, H. & Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory* 21(01), 21–59.

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.

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.

Müller, U. K. (2014). HAC Corrections for Strongly Autocorrelated Time Series. *Journal of Business & Economic Statistics* 32(3), 311–322.

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.

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.

Shiller, R. J. (1973). A Distributed Lag Estimator Derived from Smoothness Priors. *Econometrica* 41(4), 775–788.

Sims, C. A. & Zha, T. (1999). Error Bands for Impulse Responses. *Econometrica* 67(5), 1113–1155.

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. (2012). Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business & Economic Statistics* 30(4), 481–493.

Stock, J. H. & Watson, M. W. (2015). Factor Models and Structural Vector Autoregressions in Macroeconomics. Draft of chapter to appear in *Handbook of Macroeconomics*, Vol. 2.

Tseng, Y.-L. & Brown, L. D. (1997). Good exact confidence sets for a multivariate normal mean. *Annals of Statistics* 25(5), 2228–2258.

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.

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.