Misspecification-Robust Shrinkage and Selection for VAR Forecasts and IRFs

Oriol González-Casasús* Frank Schorfheide

University of Pennsylvania *University of Pennsylvania*,

CEPR, PIER, NBER

This Version: February 5, 2025

Abstract

VARs are often estimated with Bayesian techniques to cope with model dimensionality. The posterior means define a class of shrinkage estimators, indexed by hyperparameters that determine the relative weight on maximum likelihood estimates and prior means. In a Bayesian setting, it is natural to choose these hyperparameters by maximizing the marginal data density. However, this is undesirable if the VAR is misspecified. In this paper, we derive asymptotically unbiased estimates of the multi-step forecasting risk and the impulse response estimation risk to determine hyperparameters in settings where the VAR is (potentially) misspecified. The proposed criteria can be used to jointly select the optimal shrinkage hyperparameter, VAR lag length, and to choose among different types of multi-step-ahead predictors; or among IRF estimates based on VARs and local projections. The selection approach is illustrated in a Monte Carlo study and an empirical application. (JEL C11, C32, C52, C53)

Key words: Forecasting, Hyperparameter Selection, Local Projections, Misspecification, Multi-step Estimation, Shrinkage Estimators, Vector Autoregressions

* Correspondence: O. González-Casasús and F. Schorfheide: Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297. Email: oriolgc@sas.upenn.edu (González-Casasús) and schorf@ssc.upenn.edu (Schorfheide). We thank Frank Diebold, Mikkel Plagborg-Møller, Christian Wolf, and seminar participants at the Penn Econometrics Lunch, the Federal Research Bank of Philadelphia, and the 2024 EC^2 Meetings for helpful comments and suggestions.

1 Introduction

Bayesian vector autoregressions (VARs) have been widely used for macroeconomic forecasting and to analyze the dynamic effects of economic shocks since the early 1980s. They combine the VAR likelihood function with a prior distribution that shrinks the distance between the maximum likelihood estimator (MLE) and prior mean, thereby reducing the variability of the posterior mean estimator in settings where the number of parameters is large relative to the number of available observations. In a typical implementation, one or more hyperparameters control the precision of the prior distribution and hence the relative weight assigned to the prior mean in the construction of the posterior mean. For the posterior mean to deliver accurate forecasts and estimates of impulse response functions (IRFs), a data-driven hyperparameter determination is very important. Early work on forecasting with Bayesian VARs, e.g., Doan, Litterman, and Sims (1984), Todd (1984), and Litterman (1986), calibrated the hyperparameters to optimize forecast performance in a pseudo-out-of-sample setting. More recently, researchers used the Bayesian marginal data density (MDD) to select, e.g., Del Negro and Schorfheide (2004), or integrate out, e.g., Giannone, Lenza, and Primiceri (2015), the hyperparameters. In this paper, we derive novel information criteria based on estimates of prediction or IRF estimation risks that can be used for the joint determination of hyperparameters, lag length, and type of estimator, study their large sample properties, and document their performance in simulations and an empirical application. Throughout this paper, we focus on point estimation evaluated under quadratic loss.

The MLE associated with a Gaussian likelihood function minimizes in-sample one-step-ahead forecast errors. If the goal is h-step-ahead forecasting of an n-dimensional vector time series y_t, then one could either iterate the one-step-ahead MLE-based forecasts forward, or one could use a multi-step regression that projects y_t on y_{t-h} and additional lags. We will refer to the resulting estimator as loss function estimator (LFE) where "loss" refers to the h-step-ahead forecast error. Multi-step regressions are also used to estimate IRFs. Jordà (2005) showed that a regression of y_t on y_{t-h} and additional lags as controls provides an estimate of the h-order coefficient matrix of an infinite-order vector moving average (VMA) representation of y_t, which measures the response of y_t to a shock ϵ_{t-h}. The regression is called local projection (LP) and provides a popular alternative to estimating IRFs by first fitting a VAR with p lags using a one-step-ahead (quasi) likelihood objective function and then iterating the VAR forward. In the remainder of this paper we associate the MLE with the VAR IRF estimate, and the LFE with the LP IRF estimate. The main difference between

the forecasting and the IRF application is that in the former case the coefficient estimates for all p lags matter, whereas in the latter application only the coefficient estimates for the first lag are relevant.

In both applications, the multi-step estimation objective function can be interpreted as a quasi-likelihood function that ignores the serial correlation in the sequence of h-step-ahead forecast errors. Just as a standard (one-step) likelihood function, the quasi-likelihood function can also be combined with a prior distribution to obtain a quasi-posterior mean, that is a regularized version of the LFE. The contribution of this paper is to develop information criteria that provide an estimate of the relevant risk, i.e., the h-step prediction risk in the forecasting application and the IRF estimation risk in applications focusing on the dynamic effects of economic shocks. We label the criteria PC (for "prediction") and IRFC (for IRF estimation).

Starting point of our analysis is the local misspecification framework in Schorfheide (2005), henceforth S2005. The paper assumes that y_t is generated by a stationary infinite-order VMA process that drifts toward a VAR(p_*) at rate $T^{-1/2}$. The asymptotic lag length p_* is fixed and finite. Here T is the size of the estimation sample. In the forecasting application T is also the forecast origin and the econometrician uses a VAR(p) to generate h-step-ahead forecasts. For any finite sample of size T, the wedge between the infinite-order VMA generating the data and the finite-order model used by the econometrician for forecasting or IRF analysis is the source of misspecification. The $T^{-1/2}$ drift in the misspecification balances the bias-variance trade-off among the MLE and the LFE asymptotically. In a forecasting application (without the use of a prior distribution) the MLE plug-in predictor is preferable because it relies on a more efficient estimator. On the other hand, if the VAR(p) is misspecified, then the LFE plug-in predictor has the advantage that it converges to the parameter values that are optimal to predict the infinite-order VMA with a VAR of order p.

S2005 proposed a prediction criterion $PC_T(\iota, p)$ that provides an asymptotically unbiased estimate of the h-step-ahead prediction risk and can be used to select between the predictor $\iota \in \{mle, lfe\}$ and p the VAR lag length based on information available at the forecast origin T. PC is a modification of Shibata (1980)'s final prediction error criterion. The contribution in the current paper is twofold. First, in the context of the forecasting application we extend the class of predictors to shrinkage estimators that are indexed by a hyperparameter λ. This hyperparameter controls the relative weight on MLE/LFE versus prior mean in the computation of the posterior. In turn, the modified prediction criterion is a function of three arguments: $PC_T(\iota, \lambda, p)$. Second, we derive a novel criterion that provides an asymptotically

unbiased estimate of the IRF estimation risk, denoted by $IRFC_T(\iota, \lambda, p)$. This is the first criterion that allows empirical researchers to simultaneously choose among VAR and LP IRF estimates, determine the degree of shrinkage, and the number of lags. The criteria are robust to many empirically relevant types of dynamic misspecification (see S2005 and Montiel Olea, Plagborg-Møller, Qian, and Wolf (2024), henceforth MPQW), which makes selection non-trivial.

We derive formulas for the asymptotic prediction and IRF estimation risk of our shrinkage estimators and the risk estimates of the proposed information criteria. Using a data generating process (DGP) that is calibrated to match key features of a VAR estimated on U.S. data, we numerically evaluate the asymptotic formulas and illustrate their implications about the ranking of MLE and LFE based forecasting and IRF estimation, the optimal degree of shrinkage and number of lags under various degrees of misspecification. The finite-sample properties of the PC-based model determination and predictor choice are illustrated in a Monte Carlo simulation. The performance of the PC-based predictor is close to the better of the MLE and LFE predictors. We also show that once the VAR model is misspecified PC hyperparameter selection works significantly better than an MDD-based selection.

Finally, we evaluate the proposed information criterion for IRF estimation on two hundred empirical VARs constructed from the FRED-QD database. The fraction of samples for which our criterion selects the LP instead of the VAR IRF estimate varies with response horizon. If the degree of shrinkage and number of lags are also selected using the criterion, then it ranges from 60 to 85%. In general, the criterion suggests to use less shrinkage in combination with LP estimates than for VAR estimates and it prefers the use of large lag lengths for the majority of samples. The bottom line is that, from a mean-squared error (MSE) perspective, whether VAR or LP estimation is preferable is sample and horizon dependent. There is no clear winner, discrediting with widespread belief that LPs are *always* preferred under misspecification.

Our paper is connected to three broad strands of the econometrics literature: multi-step forecasting, IRF estimation, and model selection. The currently most active branch is the one on IRFs, partly due to the increasing popularity of LPs. Our LP IRF estimator is similar to Bayesian LP estimator proposed by Miranda-Agrippino and Ricco (2021). However, they use a quasi-MDD for hyperparameter determination (averaging rather than selection), whereas we propose to use a criterion that directly targets estimation risk. Targeting risk works equally well than using MDD under correct specification, but here it is shown to yield large improvements under misspecification.

Plagborg-Møller and Wolf (2021) show that LPs and VARs estimate the same IRFs in population if the number of lags is unrestricted. In finite samples, there is however a bias-variance trade-off that is illustrated in a large-scale simulation study in Li, Plagborg-Møller, and Wolf (2022), henceforth LPW. This trade-off is similar, but not identical, to the bias-variance trade-off between the MLE and LFE shrinkage predictors. While in our local misspecification framework, in the absence of shrinkage toward a prior the LFE based predictor always has lower bias than the MLE based predictor, it is not true that the standard LP IRF estimator *always* has lower bias than the VAR estimator. Once data-driven shrinkage is introduced, there are two sources of bias: a variance-reducing bias generating by the prior distribution and the misspecification bias. This creates a complicated trade-off that our information criteria are designed to resolve.

An important insight in the LP literature developed in the papers by Montiel Olea and Plagborg-Møller (2021) and MPQW, is that the use of "additional" lags in LPs, called lag augmentation, can alleviate inference problems caused by serial correlation in LP regression errors. In particular, using the same drifting DGP framework of S2005, MPQW show that under lag augmentation, which in the S2005 framework means that the number of lags in the VAR exceed the asymptotic lag order of the DGP, the LP estimator is correctly centered up to order $O(T^{-1/2})$. MPQW exploit this centering property to construct IRF confidence intervals that have better coverage properties than the corresponding VAR confidence intervals and are hence robust to misspecification. In our analysis the correct centering eliminates misspecification bias that contributes to the MSE of the IRF estimator. Our IRFC can be used to determine how many lags are necessary in view of the observed data to achieve a correct centering of the LP estimator. However, in our setting the overall bias-variance trade-off is more complicated, because the shrinkage induces an additional variance-reducing bias that affects the MSE.

Ludwig (2024) derived a finite-sample equivalence result between VAR and LP regressions, showing based on the Frisch-Waugh-Lovell Theorem that an iteration of increasingly larger order VARs can exactly replicate an LP, and similarly a multi-step VAR prediction can be replicated by LPs of equal and lower orders. He makes the case that a fair comparison of the two types of IRF estimators should not use the same number of lags for VAR and LP, but instead equalize model size by accounting for the result that a collection of VARs can replicate an LP and vice versa. Our IRFC model determination does this by optimizing jointly over estimator and lag length.

We now turn to the multi-step forecasting literature, where LFEs are also called multi-

step or direct estimators and have been studied by, among others, Findley (1983), Weiss (1991), Bhansali (1997), Clements and Hendry (1998), Ing (2003). Marcellino, Stock, and Watson (2006) undertake a large-scale empirical comparison of MLE versus LFE plug-in predictors using data on more than 150 monthly macroeconomic time series. They find that MLE plug-in predictions tend to yield smaller forecast errors, in particular in high-order autoregressions and for long forecast horizons. For series measuring wages, prices, and money, on the other hand, LFE plug-in predictors improve upon MLE plug-in predictors in low-order autoregressions.

In regard to the theoretical analysis, to capture misspecification it has been typically assumed in the literature on prediction with autoregressive models that the DGP is fixed and the class of candidate forecasting models is increasing with sample size, e.g., Shibata (1980), Speed and Yu (1993), Bhansali (1996), Ing and Wei (2003). Thus, the discrepancy between the best estimated forecasting model and the DGP vanishes asymptotically. We follow the opposite approach. We keep the class of forecasting models fixed and let the degree of misspecification asymptotically vanish. In our setup the degree of misspecification is "too small" to be consistently estimable. Hence, PC and IRFC provide only an asymptotically unbiased, but not consistent estimate of the final prediction or IRF estimation risk.

With regard to the literature on model determination and information criteria, we mentioned previously that PC is a modification of Shibata (1980)'s final prediction error criterion. In the empirical Bayes literature, the use of an unbiased risk estimate as an objective function for hyperparameter selection dates back to Stein (1981). In the recent Bayesian time series literature hyperparameter determination based on MDD dominates. In this regard, Giannone, Lenza, and Primiceri (2015) has been a very influential paper, albeit not the first to propose the use of MDDs. There is other work proposing objective functions that target the estimation risk of VAR coefficient estimators and transformations, e.g., impulse response functions, thereof. Examples of such work include Hansen (2016) and Lohmeyer, Palm, Reuvers, and Urbain (2018). However, assumptions about model misspecification are different from our setting, which leads to different risk estimates.

The remainder of the paper is organized as follows. Rather than developing the theory for multi-step prediction and IRF estimation in parallel, we begin with the multi-step estimation problem and later modify the analysis to study IRF estimation. Because the notation for the general case is cumbersome, we start with an analysis of the restricted case of $p = p_* = q = 1$ before we generalize it to multiple lags and an unknown asymptotic lag order. Section 2 describes the DGP, the shrinkage estimators and predictors, and the

prediction risk associated with them. Section 3 discusses hyperparameter selection based on an asymptotically unbiased risk estimate and a (quasi) marginal data density as an alternative to MDD-based hyperparameter selection. An extension to a setting in which the true asymptotic lag order p_* of the DGP is unknown and p needs to be determined by the empirical researcher is provided in Section 4. The presentation is based on companion form representations of DGP and forecasting models. In Section 5 we turn to IRF estimation with VARs and LPs and derive IRFC. Section 6 provides a numerical illustration of the asymptotic formulas derived previously, Section 7 presents results from Monte Carlo experiments, and an empirical application is conducted in Section 8. Finally, Section 9 concludes. Proofs, derivations, and additional simulation results are relegated to the Online Appendix.

2 Multi-step Forecasting with a VAR(1)

An econometrician considers MLE and LFE shrinkage predictors to forecast an infinite-order VMA process. The predictors are posterior means derived from hierarchical models described in Section 2.1. The distributional assumptions therein are only used to define the class of estimators considered, and no further reference is made in our theory to any of the distributional assumptions imposed by the hierarchical model. The degree of shrinkage is determined by a hyperparameter that controls the weight on the prior information. Setting this hyperparameter to zero leads to the least squares estimators/predictors studied in S2005. The DGP is described in Section 2.2. It takes the form of a VAR but the innovations are distorted by an infinite-dimensional linear process that vanishes at rate $T^{-1/2}$. In Section 2.3 we derive the limit distribution of the predictors and the associated prediction risk. To keep the exposition relatively simple, we first analyze forecasts from a locally misspecified VAR(1). The extension to multiple lags and an unknown lag order p is provided in Section 4. The results presented in this section generalize those from S2005 (Theorems 1 to 3) to shrinkage estimators.

2.1 MLE and LFE Shrinkage Predictors

To generate h-step-ahead forecasts, an econometrician considers a possibly misspecified VAR(1) of the form

$$y_t = \Phi y_{t-1} + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma_{uu}), \tag{1}$$

where y_t is a $n \times 1$ vector. The forecasts are evaluated under the quadratic prediction error loss function

$$L(y_{T+h}, \hat{y}_{T+h}) = \text{tr}\big[W(y_{T+h} - \hat{y}_{T+h})(y_{T+h} - \hat{y}_{T+h})'\big] = \|y_{T+h} - \hat{y}_{T+h}\|^2_W. \tag{2}$$

W is a symmetric and positive-definite weight matrix.

If Φ were known then the optimal h-step-ahead point predictor at forecast origin T would be $\Phi^h y_T$. This raises the question of how to estimate Φ^h. We consider two alternatives: a likelihood-based estimator of Φ that is plugged into the prediction function $\Phi^h y_T$; and a direct estimate of Φ^h obtained by regressing y_t on y_{t-h}. We refer to the latter estimator as loss-function based because the estimation objective function is the loss function under which the forecasts are evaluated. Rather than using these two estimators directly, we combine their estimation objective functions with a prior distribution to obtain a posterior mean estimator that can be interpreted as a shrinkage estimator. The degree of shrinkage is controlled by a hyperparameter that we determine in Section 3.

MLE Shrinkage Predictor. Define $S_{T,kl} = \sum_{t=1}^{T} y_{t-k} y'_{t-l}$. The MLE can be expressed as

$$\hat{\Phi}_T(mle) = S_{T,01} S_{T,11}^{-1}. \tag{3}$$

The likelihood-based shrinkage estimator of Φ is defined as the posterior mean obtained by combining the likelihood function associated with (1) with the following prior:

$$\Phi|\Sigma_{uu} \sim N\big(\underline{\Phi}_T,\ (\tilde{\lambda}_T \underline{P}_\Phi)^{-1} \otimes \Sigma_{uu}\big). \tag{4}$$

The prior are indexed by the hyperparameter $\tilde{\lambda}_T$ that controls the degree of shrinkage. The mean and the hyperparameter of the prior distribution are indexed by the sample size T for a reason that will become clear below. Using standard calculations, the posterior mean can be expressed as the matrix-weighted average of the prior mean and the MLE:

$$\bar{\Phi}_T(mle, \tilde{\lambda}_T) = \big[\tilde{\lambda}_T \underline{\Phi}_T \underline{P}_\Phi + \hat{\Phi}_T(mle) S_{T,11}\big] \bar{P}_\Phi^{-1}(\tilde{\lambda}_T), \quad \bar{P}_\Phi(\tilde{\lambda}_T) = \tilde{\lambda}_T \underline{P}_\Phi + S_{T,11}. \tag{5}$$

Note that for $\tilde{\lambda}_T = 0$ we obtain that $\bar{\Phi}_T(mle, \tilde{\lambda}_T) = \hat{\Phi}_T(mle)$. Moreover, $\bar{\Phi}_T(mle, \tilde{\lambda}_T) = \underline{\Phi}_T$ if $\tilde{\lambda}_T = \infty$. Let $\Psi = \Phi^h$ and we can define the likelihood-based (plug-in) shrinkage estimator of Φ^h as[1]

$$\bar{\Psi}_T(mle, \tilde{\lambda}_T) = \bar{\Phi}^h_T(mle, \tilde{\lambda}_T). \tag{6}$$

[1]We are using a plug-in estimator $\bar{\Phi}^h$ rather than the posterior mean of Φ^h which would also depend on higher-order moments of the posterior distribution. However, these moments would be negligible in our asymptotic analysis.

The MLE shrinkage predictor is then defined as

$$\hat{y}_{T+h}(mle, \tilde{\lambda}_T) = \bar{\Psi}_T(mle, \tilde{\lambda}_T) y_T. \tag{7}$$

LFE Shrinkage Predictor. The loss function-based predictor is based on the multi-step regression

$$y_t = \Psi y_{t-h} + v_t, \quad v_t \sim \mathcal{N}(0, \Sigma_{vv}), \tag{8}$$

ignoring the serial correlation in v_t implied by the VAR(1) in (1). The rationale behind this estimator is that it directly targets the h-step-ahead forecast error covariance matrix. Define

$$\hat{\Psi}_T(lfe) = S_{T,0h} S_{T,hh}^{-1}. \tag{9}$$

Using the prior

$$\Psi | \Sigma_{vv} \sim N\big(\underline{\Psi}_T,\ (\tilde{\lambda}_T \underline{P}_\Psi)^{-1} \otimes \Sigma_{vv}\big), \tag{10}$$

we obtain the quasi-posterior

$$\bar{\Psi}_T(lfe, \tilde{\lambda}_T) = \big[\tilde{\lambda}_T \underline{\Psi}_T \underline{P}_\Psi + \hat{\Psi}_T(lfe) S_{T,hh}\big] \bar{P}_\Psi^{-1}(\tilde{\lambda}_T), \quad \bar{P}_\Psi(\tilde{\lambda}_T) = \tilde{\lambda}_T \underline{P}_\Psi + S_{T,hh}. \tag{11}$$

This leads to the LFE shrinkage predictor

$$\hat{y}_{T+h}(lfe, \tilde{\lambda}_T) = \bar{\Psi}_T(lfe, \tilde{\lambda}_T) y_T. \tag{12}$$

2.2 Drifting DGP and Prior

We assume that the sample has been generated from a covariance stationary DGP with an infinite-dimensional VMA representation. While the sample size T is fixed in practice, we use $T \longrightarrow \infty$ asymptotics to approximate the prediction risk. If the DGP and the lag length of the misspecified forecasting model are fixed then the variance of the estimators of Φ^h vanishes at rate T^{-1} whereas the misspecification bias does not disappear. Thus, eventually, the loss-function-based predictor dominates the likelihood-based predictor along this asymptote, even if the misspecification is small.

To generate asymptotics that better reflect the finite-sample trade-offs faced by the forecaster one has two choices: either increase the dimensionality of the forecasting model with sample size or let the DGP drift toward the forecasting model. As in S2005, we pursue the

latter approach and assume that the DGP takes the form of a drifting VMA process that is local to the VAR in (1):

$$y_t = Fy_{t-1} + \epsilon_t + \frac{\alpha}{\sqrt{T}} \sum_{j=1}^{\infty} A_j \epsilon_{t-j}, \quad \epsilon_t \sim (0, \Sigma_{\epsilon\epsilon}). \tag{13}$$

This means that misspecification bias of the MLE of Φ in (1) relative to the "true" F in (13) is of order $O(T^{-1/2})$. The contribution of parameter estimation to prediction loss can be represented as the sum of a squared bias and a variance term. The $O(T^{-1/2})$ drift guarantees that these two terms are asymptotically of the same order.

Posterior means combine information from the likelihood and the prior. Typically, the information about the unknown parameters contained in the likelihood grows with the sample size, and the information in the prior distribution is held constant. To develop an asymptotic framework that yields non-trivial shrinkage decisions, we need to let the information in the prior distribution grow at the same rate as the likelihood information. In the present stationary environment, each observation adds information at the parametric $T^{-1/2}$ rate. To balance the informational content of likelihood and prior, we assume that the prior means approach F and F^h, respectively, at rate $T^{-1/2}$:

$$\underline{\Phi}_T = F + T^{-1/2}\underline{\phi}, \quad \underline{\Psi}_T = F^h + T^{-1/2}\underline{\psi}. \tag{14}$$

For the subsequent analysis, it is convenient to re-scale the precision hyperparameter as follows:

$$\tilde{\lambda}_T = \lambda T. \tag{15}$$

In slight abuse of notation, we replace the $\tilde{\lambda}_T$ argument of the shrinkage estimators $\bar{\Psi}_T(\cdot)$ by the re-scaled hyperparameter λ. Taken together, the drift and the re-scaling ensure that the bias induced by placing non-zero weight on the prior mean is of the same order as the misspecification bias of MLE and LFE and that prior precision and the information in the likelihood function are of the same order asymptotically.

To understand the assumptions on the drift rates, consider the expressions in (5). Using (15) we can write the posterior precision as

$$\bar{P}_\Phi(\lambda) = T \cdot \big(\lambda \underline{P}_\Phi + S_{T,11}/T\big),$$

where $S_{T,11}/T$ is convergent. Thus, for any fixed λ the prior precision makes a non-trivial contribution to the posterior precision. If the eigenvalues of F are less than one in absolute value and the A_js satisfy a summability condition that will be stated more formally below,

the MLE behaves asymptotically as $\hat{\Phi}_T(mle) = F + T^{-1/2}\xi_T + O_p(T^{-1})$, where ξ_T is an $O_p(1)$ random variable. Thus,

$$\bar{\Phi}_T(mle, \lambda) = F + T^{-1/2} \cdot \left[\lambda \underline{\phi} \underline{P}_\Phi + \xi_T(S_{T,11}/T)\right]\left(\lambda \underline{P}_\Phi + S_{T,11}/T\right)^{-1} + O_p(T^{-1}).$$

Our assumptions on the drifts ensure that we subsequently can focus on the $O_p(T^{-1/2})$ term in the prediction risk calculations. By construction the relative weights on MLE and prior mean are no longer sample size dependent. This captures the fact that in practice prior distributions play an important role in regularizing VAR parameter estimates to obtain good forecasting performance.

2.3 Prediction Risk

Risk and Optimal Prediction. As is common in the literature, to streamline the theoretical derivations we assume that there are two independent processes, $\{y_t\}$, and $\{\tilde{y}_t\}$, both generated from the DGP in (13); see, for instance, Baillie (1979), Reinsel (1980), Shibata (1980), and Lewis and Reinsel (1985, 1988). The former is used for parameter estimation and the latter is the process to be forecast. This assumption removes the (asymptotically negligible) correlation between the parameter estimates and the lagged value at the forecast origin. The optimal predictor of a future observation $\tilde{y}_{T+h}$ generated from the DGP is the conditional mean

$$\hat{y}_{T+h}^{opt} = \mathbb{E}_T[\tilde{y}_{T+h}], \tag{16}$$

where the expectation is taken conditional on the (infinite) history of the process up to time T and the parameters α, F and $A(L)$. The expected loss of $\hat{y}_{T+h}^{opt}$ provides a lower bound for the frequentist risk of any estimator. We normalize the prediction risk $\mathcal{R}(\hat{y}_{T+h})$ of a predictor $\hat{y}_{T+h}$ as follows

$$\mathcal{R}(\hat{y}_{T+h}) = \mathbb{E}\left[\|\tilde{y}_{T+h} - \hat{y}_{T+h}\|_W^2\right] - \mathbb{E}\left[\|\tilde{y}_{T+h} - \hat{y}_{T+h}^{opt}\|_W^2\right] = \mathbb{E}\left[\|\hat{y}_{T+h} - \hat{y}_{T+h}^{opt}\|_W^2\right]. \tag{17}$$

Pseudo-optimal value. To characterize the pseudo-optimal value (pov) for Ψ in the VAR(1)-based prediction function $\Psi\tilde{y}_T$ we define $A_0 = 0$ and $A(L) = \sum_{j=0}^{\infty} A_j L^j$. Moreover, we let $z_t = A(L)\epsilon_t$ and

$$\begin{aligned}
\Gamma_{yy,h} &= \lim_{T\longrightarrow\infty} \mathbb{E}[y_{T+h}y_T'] = \sum_{j=0}^{\infty} F^{j+h}\Sigma_{\epsilon\epsilon}F^{j'} \\
\Gamma_{zy,h} &= \lim_{T\longrightarrow\infty} \mathbb{E}[z_{T+h}y_T'] = \sum_{j=0}^{\infty} A_{j+h}\Sigma_{\epsilon\epsilon}F^{j'}.
\end{aligned}$$

It was shown in S2005 that the pov takes the form

$$\tilde{\Psi}_T(pov) = F^h + \alpha T^{-1/2}\mu(pov) + \alpha O(T^{-1}), \quad \mu(pov) = \sum_{j=0}^{h-1} F^j \Gamma_{zy,h-j}\Gamma_{yy,0}^{-1}. \tag{18}$$

Limit Distribution. As an intermediate step in the calculation of the prediction risk the limit distributions for $\bar{\Psi}_T(mle,\lambda)$ and $\bar{\Psi}_T(lfe,\lambda)$ are derived. To do so, we state some regularity conditions:

Assumption 1

(i) The largest eigenvalue of F is less than one in absolute value.

(ii) The sequence of $n\times n$ matrices $\{A_j\}_{j=0}^{\infty}$ satisfies the following summability condition: $\sum_{j=0}^{\infty} j^2\|A_j\| < \infty$.

(iii) $\{\epsilon_t\}$ is a sequence of independent, n-dimensional, mean zero random variates with $\mathbb{E}[\epsilon_t\epsilon_t'] = \Sigma_{\epsilon\epsilon}$.

(iv) The ϵ_t's are uniformly Lipschitz over all directions, that is, there exist $K > 0$, $\delta > 0$, and $\nu > 0$ such that for all $0 \le w - u \le \delta$,

$$\sup_{\nu'\nu=1} \mathbb{P}\{u < \nu'\epsilon_t < w\} \le K(w-u)^{\nu}.$$

(v) There exists an $\eta > 0$ such that

$$\mathbb{E}\left[\|\epsilon_t'\epsilon_t\|^{3h+\eta}\right] < \infty.$$

Assumptions 1(i) and (ii) guarantee that for any fixed T the DGP is stationary. Assumptions (iii) to (v) ensure that the finite sample moments of the two predictors eventually exist. The following theorem characterizes the limit distribution of the likelihood and loss function based shrinkage estimators for a fixed λ. We use $\Longrightarrow$ to denote convergence in distribution.

Theorem 1 *Suppse that the DGP satisfies Assumption 1. Then, for $\iota \in \{lfe, mle\}$ and $\lambda \ge 0$:*

$$\bar{\Psi}_T(\iota,\lambda) = F^h + T^{-1/2}[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda)] + o_p(T^{-1/2}), \tag{19}$$

where $\zeta_T(\iota,\lambda) \Longrightarrow \zeta(\iota,\lambda) \sim \mathcal{N}(0, V(\iota,\lambda))$.

The formulas for the bias terms $\delta(\iota,\lambda)$ and $\mu(\iota,\lambda)$ and the asymptotic covariance matrix $V(\iota,\lambda)$ are provided in the proof of Theorem 1 in the Online Appendix. $\bar{\Psi}_T(\iota,\lambda)$ converges to F^h in probability as $T \longrightarrow \infty$. The important terms for the subsequent prediction risk calculation are those premultiplied by $T^{-1/2}$. Consider the case $\lambda = 0$. Then the weight on the prior mean is zero and $\delta(\iota,\lambda) = 0$. The bias term $\mu(\iota,\lambda)$ arises from the covariance between $z_t = A(L)\epsilon_t$ and y_t. Importantly, it can be shown that

$$\mu(lfe,0) = \mu(pov), \tag{20}$$

i.e., in the absence of shrinkage, the LFE is centered at the pov. For $\lambda > 0$ there is a second bias term, $\delta(\iota,\lambda)$, which captures the effect of the prior distribution. At $\lambda = \infty$, $\delta(\iota,\lambda)$ equals the local prior mean and $\mu(\iota,\lambda) = 0$. Finally, $\zeta(\iota,\lambda)$ is a mean-zero Normal random variable.

It is well-known that $V(mle,0) < V(lfe,0)$; see, for instance, S2005. The LFE is inefficient, because it ignores the serial correlation of h-step-ahead forecast errors in its estimation objective function. Shrinkage, i.e., $\lambda > 0$ not just affects the asymptotic bias, but also the variance of the estimators. The larger the precision, the smaller the sampling variance of the shrinkage estimator.

Prediction Risk. The next theorem characterizes the asymptotic prediction risk

$$\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota,\lambda)) = \lim_{T\longrightarrow\infty} T\mathcal{R}(\hat{y}_{T+h}(\iota,\lambda)) \tag{21}$$

of the MLE and LFE shrinkage predictors based on their limit distribution. Because of the normalization in (17) the prediction risk is determined by the bias and variance of the Ψ estimators. Assumptions 1 (iii) to (v) ensure that the finite-sample moments of the estimators are eventually finite and converge to the moments of the limit distribution.

Theorem 2 *Suppose Assumption 1 is satisfied. Then, for $\iota \in \{mle, lfe\}$ and $\lambda \geq 0$:*

$$\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota,\lambda)) = \underbrace{\|\delta(\iota,\lambda) - \alpha(\mu(pov) - \mu(\iota,\lambda))\|^2_{W\otimes\Gamma_{yy,0}}}_{=:\,\bar{\mathcal{R}}_B(\iota,\lambda)} + \underbrace{tr\Big\{(W\otimes\Gamma_{yy,0})V(\iota,\lambda)\Big\}}_{=:\,\bar{\mathcal{R}}_V(\iota,\lambda)} + C, \tag{22}$$

where

$$C = \alpha^2\mathbb{E}\left[\left\|\sum_{j=0}^{h-1} F^j\mathbb{E}_T[z_{T+h-j}] - \mu(pov)\tilde{y}_T\right\|^2_W\right].$$

A proof of Theorem 2 is provided in the Online Appendix. The proof eventually replaces the finite-sample second moments of $\zeta_T(\iota,\lambda)$ by the moments of the limit random variable $\zeta(\iota,\lambda)$ of Theorem 1. This requires that the sequence $\zeta_T(\iota,\lambda)$ is uniformly integrable. A formal proof using Assumptions 1(iv) and (v) for $\lambda = 0$ was provided in S2005. The proof can be extended to $\lambda > 0$ by noting that the Bayes estimators considered in this paper are weighted averages of the least squares estimators in S2005 and (non-stochastic) prior means.

Recall that in (17) we defined a normalized risk that can be interpreted as the discrepancy between a predictor $\hat{y}_{T+h}$ and the conditional mean $\hat{y}^{opt}_{T+h}$ associated with infinite-dimensional DGP in (13). According to the calculations underlying Theorem 2, the conditional mean $\hat{y}^{opt}_{T+h}$ can be replaced, without any consequences for hyperparameter selection, by a pseudo-optimal predictor from the VAR model $\tilde{\Psi}(pov)$; see (18). The constant C arises from this replacement, and since it does not depend on the predictor (ι,λ) it is irrelevant for rankings.

The prediction risk is decomposed in a bias term $\bar{\mathcal{R}}_B(\iota,\lambda)$ and a variance term $\bar{\mathcal{R}}_V(\iota,\lambda)$. Consider the LFE which corresponds to $\iota = lfe$. Recall that for $\lambda = 0$ the prior-induced bias $\delta(lfe,\lambda) = 0$ and the regression-induced bias term $\mu(lfe,\lambda) = \mu(pov)$. Thus, $\bar{\mathcal{R}}_B(\iota,\lambda) = 0$, but the variance term $\bar{\mathcal{R}}_V(\iota,\lambda)$ is large. Raising λ generates some bias, but also reduces the variance contribution to the prediction risk. The same logic applies to the MLE shrinkage predictor, i.e., $\iota = mle$, except that $\mu(pov) - \mu(mle,0) \neq 0$. For $\lambda > 0$ the $\delta(\iota,\lambda)$ term generated by the prior could either increase or decrease the estimation bias component. The smaller the misspecification α, the less important is the bias term, and the more important becomes the variance component of the risk, $\bar{\mathcal{R}}_V(\iota,\lambda)$, when choosing between the MLE and LFE shrinkage predictors.

3 Hyperparameter Determination

We now turn to the derivation of a selection criterion that is based on an asymptotically unbiased (prediction) risk estimate (URE). The URE objective function constructed in Section 3.1 generalizes the PC criterion proposed in S2005 so that it can also be used for the determination of hyperparameter λ. The modification required to target IRF estimation risk will be provided in Section 5.2 below. Because in the Bayesian VAR literature it is common so select prior hyperparameters using a (quasi) MDD, we present an MDD criterion for the

hyperparameter selection in Section 3.2.[2] It is important to note that the PC criterion can also be used to choose between MLE and LFE, whereas the MDD cannot.

3.1 Asymptotically Unbiased Risk Estimation

In-sample Prediction Loss. The in-sample mean squared h-step ahead forecast error matrix is given by

$$MSE(\iota,\lambda) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{\Psi}_T(\iota,\lambda)y_{t-h})(y_t - \bar{\Psi}_T(\iota,\lambda)y_{t-h})'. \tag{23}$$

We normalize the forecast error by the MSE of the unshrunk loss function predictor, which gives the smallest in-sample MSE, and define the loss differential

$$\Delta_{L,T}(\iota,\lambda) = T\left(tr\left\{W\cdot MSE(\iota,\lambda)\right\} - tr\left\{W\cdot MSE(lfe,0)\right\}\right) \geq 0. \tag{24}$$

Using the asymptotic representation of $\bar{\Psi}(\iota,\lambda)$ given in Theorem 1, and the facts that $\delta(lfe,0) = 0$ and $\mu(lfe,0) = \mu(pov)$, we show in the Online Appendix that the asymptotic behavior of the risk differential can be characterized as follows:

Theorem 3 *Suppose that Assumption 1 is satisfied. Then, for $\iota \in \{mle, lfe\}$ and $\lambda \geq 0$:*

(i) The in-sample forecast error loss differential has the following limit distribution

$$\begin{aligned}\Delta_{L,T}(\iota,\lambda) \implies & \left\|\delta(\iota,\lambda)\right\|^2_{W\otimes\Gamma_{yy,0}} + \alpha^2\left\|\mu(pov)-\mu(\iota,\lambda)\right\|^2_{W\otimes\Gamma_{yy,0}} \\ & + \left\|\zeta(lfe,0)-\zeta(\iota,\lambda)\right\|^2_{W\otimes\Gamma_{yy,0}} \\ & +2\alpha tr\left\{W\left[\mu(pov)-\mu(\iota,\lambda)\right]\Gamma_{yy,0}\left[\zeta(lfe,0)-\zeta(\iota,\lambda)\right]'\right\} \\ & -2\alpha tr\left\{W\delta(\iota,\lambda)\Gamma_{yy,0}\left[\mu(pov)-\mu(\iota,\lambda)\right]'\right\} \\ & -2tr\left\{W\delta(\iota,\lambda)\Gamma_{yy,0}\left[\zeta(lfe,0)-\zeta(\iota,\lambda)\right]'\right\}.\end{aligned}$$

(ii) The expected in-sample forecast error differential converges to

$$\begin{aligned}\mathbb{E}\left[\Delta_{L,T}(\iota,\lambda)\right] \longrightarrow & \ \bar{\mathcal{R}}_B(\iota,\lambda) + \bar{\mathcal{R}}_V(\iota,\lambda) - \left(\bar{\mathcal{R}}_B(lfe,0) + \bar{\mathcal{R}}_V(lfe,0)\right) \\ & +2\bar{\mathcal{R}}_V(lfe,0) - 2tr\left\{(W\otimes\Gamma_{yy,0})Cov(lfe,0;\iota,\lambda)\right\}.\end{aligned}$$

[2]Bayesian procedures that are based on prior distributions indexed by hyperparameters which have been estimated in a preliminary step from the data are called empirical Bayes (EB) procedures; see Robbins (1955).

A formula for $Cov(lfe, 0; \iota, \lambda)$ is provided in the Online Appendix. It is important to note that in our local analysis, the loss differential $\Delta_{L,T}(\iota, \lambda)$ converges in distribution to a random variable and not a constant.

From In-Sample to Out-of-Sample Prediction Risk. The limit random variables $\zeta(\iota, \lambda)$ are defined in Theorem 1 and the asymptotic risk components $\bar{\mathcal{R}}_B(\iota, \lambda)$ and $\bar{\mathcal{R}}_V(\iota, \lambda)$ are given in Theorem 2. Theorem 3 shows that the expected forecast error loss differential converges to the sum of the risk differential, the risk of the LFE with $\lambda = 0$, and the covariance term $tr\left\{(W \otimes \Gamma_{yy,0})\mathbb{E}\left[(\zeta(\iota, \lambda) - \zeta(lfe, 0))(\zeta(\iota, \lambda) - \zeta(lfe, 0))'\right]\right\}$. Because $\bar{\mathcal{R}}_V(lfe, 0)$ is irrelevant for comparisons across different (ι, λ), the formula suggests to correct the MSE by twice the covariance component of the asymptotic risk to obtain an asymptotically unbiased estimate of the (normalized) prediction risk $\mathcal{R}(\hat{y}_{T+h}(\iota, \lambda))$ that can be used as a selection criterion.

Definition 1 *Define the $PC_T(\iota, \lambda)$ criterion for the joint selection of prior shrinkage and type of estimator as*

$$PC_T(\iota, \lambda) = Ttr[W \cdot MSE(\iota, \lambda)] + 2\hat{\mathcal{R}}_{Cov}(lfe, 0; \iota, \lambda),$$

where $\hat{\mathcal{R}}_{Cov}(lfe, 0; \iota, \lambda)$ has the property that

$$\mathbb{E}[\hat{\mathcal{R}}_{Cov}(lfe, 0; \iota, \lambda)] \longrightarrow \text{tr}\left\{(W \otimes \Gamma_{yy,0})Cov(lfe, 0; \iota, \lambda)\right\}.$$

The term $2\hat{\mathcal{R}}_{Cov}(lfe, 0; \iota, \lambda)$ in the definition of the PC criterion can be viewed as a penalty term that turns the in-sample fit measured by $MSE(\iota, \lambda)$ into a measure of out-of-sample fit. It must fulfill an asymptotic unbiasedness condition to guarantee that PC provides an URE for the asymptotic prediction risk.[3] After combining Definition 1 of the selection criterion with the MSE differential formula in (24) and Theorem 3 we can deduce that

$$\begin{aligned} &\mathbb{E}\big[PC_T(\iota, \lambda) - PC_T(\iota', \lambda')\big] \\ &\quad = \quad \mathbb{E}\big[\Delta_{L,T}(\iota, \lambda) - \Delta_{L,T}(\iota', \lambda')\big] + 2\mathbb{E}\big[\hat{\mathcal{R}}_{Cov}(lfe, 0; \iota, \lambda) - \hat{\mathcal{R}}_{Cov}(lfe, 0; \iota', \lambda')\big] \\ &\quad \longrightarrow \quad \bar{\mathcal{R}}(\hat{y}_{T+h}(\iota, \lambda)) - \bar{\mathcal{R}}(\hat{y}_{T+h}(\iota', \lambda')) \end{aligned} \tag{25}$$

as $T \longrightarrow \infty$. $PC_T(\iota, \lambda)$ can be used to choose between MLE and LFE based shrinkage and to select the hyperparameter λ. Note that in our local misspecification framework it

[3] In the implementation of the PC criterion in Sections 7 and 8 we replace F, $\Sigma_{\epsilon\epsilon}$, and $\Gamma_{yy,j}$ in the covariance formula by consistent estimates to construct $\hat{\mathcal{R}}_{Cov}(lfe, 0; \iota, \lambda)$.

is not possible to consistently estimate the end-of-sample h-step-ahead prediction risk due to the rescaling in (21). One can only construct estimates that remain noisy in the limit, and so would be any putative "oracle/best" estimator. PC provides such an estimate and has the property that it is unbiased for the risk. In Section 7.1 we provide a numerical illustration that compares draws from the distribution of $PC_T(\iota, \lambda)$ to the asymptotic risk function $\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota, \lambda))$.

An Alternative Selection Criterion. One could define an alternative selection criterion by replacing the goodness-of-fit term $Ttr[W \cdot MSE(\iota, \lambda)]$ with the difference between $\bar{\Psi}_T(\iota, \lambda)$ and $\bar{\Psi}_T(lfe, 0)$, which leads to the following definition:

Definition 2 *Define the $PC^*_T(\iota, \lambda)$ criterion for the joint selection of prior shrinkage and type of estimator as*

$$PC^*_T(\iota, \lambda) = T \left\| \bar{\Psi}_T(\iota, \lambda) - \bar{\Psi}_T(lfe, 0) \right\|^2_{W \otimes \Gamma_{yy,0}} + 2\hat{\mathcal{R}}_{Cov}(lfe, 0; \iota, \lambda).$$

The rationale for the formula in Definition 2 is that the difference $\bar{\Psi}_T(\iota, \lambda) - \bar{\Psi}_T(lfe, 0)$ can be expanded to

$$\begin{aligned} &\sqrt{T}\big(\bar{\Psi}_T(\iota, \lambda) - \bar{\Psi}_T(lfe, 0)\big) \\ &= \sqrt{T}\big(\bar{\Psi}_T(\iota, \lambda) - F^h - \alpha\mu(pov)\big) - \sqrt{T}\big(\bar{\Psi}_T(lfe, 0) - F^h - \alpha\mu(pov)\big). \end{aligned}$$

Using Theorem 1 and calculations underlying the proofs of Theorems 2 and 3 one can show that

$$\begin{aligned} &\lim_{T \longrightarrow \infty} T\mathbb{E}\left[\left\| \bar{\Psi}_T(\iota, \lambda) - \bar{\Psi}_T(lfe, 0) \right\|^2_{W \otimes \Gamma_{yy,0}} \right] \\ &= \bar{\mathcal{R}}_B(\iota, \lambda) + \bar{\mathcal{R}}_V(\iota, \lambda) + \bar{\mathcal{R}}_V(lfe, 0) - 2\,\mathrm{tr}\left\{ (W \otimes \Gamma_{yy,0}) Cov(lfe, 0; \iota, \lambda) \right\}. \end{aligned} \tag{26}$$

Thus, $PC^*_T(\iota, \lambda)$ also provides, up to the constant $\bar{\mathcal{R}}_V(lfe, 0)$, an asymptotically unbiased risk estimate. It turns out that $PC^*_T(\iota, \lambda)$ is very closely connected to the $IRFC_T(\iota, \lambda)$ criterion proposed in Section 5.2. In unreported simulations we find that both PC_T and PC^*_T perform equally well in multi-step forecasting applications.

3.2 MDD Based Hyperparameter Selection

In the VAR literature, hyperparameters are often selected using the MDD. We will derive a quasi MDD for the multi-step regression (8), which can be written in matrix form as

$$Y = X\Psi' + V. \tag{27}$$

Here Y, X, and V are the $T \times n$ matrices with rows y_t' and y_{t-h}', and v_t'. The quasi MDD derived subsequently ignores the VAR-implied autocorrelation in v_t and mechanically uses the formulas for a multivariate regression model. We combine the conditional prior for $\Psi|\Sigma_{vv}$ in (10) with a marginal distribution for Σ_{vv}:

$$\Sigma_{vv} \sim IW(\underline{\nu}, \underline{S}). \tag{28}$$

Define

$$\bar{S} = \underline{S} + (\lambda T)\underline{\Psi}_T \underline{P}_\Psi \underline{\Psi}_T' + Y'Y - \bar{\Psi}_T \bar{P}_\Psi \bar{\Psi}_T', \quad \bar{\nu} = \underline{\nu} + T. \tag{29}$$

It can be shown that the MDD takes the form

$$p(Y|\iota,\lambda) = \int\int p(Y|\Psi,\Sigma_{vv})p(\Psi,\Sigma_{vv})d\Psi d\Sigma_{vv} = (2\pi)^{-nT/2}\frac{|\lambda T\underline{P}_\Psi|^{n/2}}{|\bar{P}_\Psi|^{n/2}}\frac{\underline{C}_{IW}}{\bar{C}_{IW}}, \tag{30}$$

where

$$\frac{\underline{C}_{IW}}{\bar{C}_{IW}} = \frac{|\underline{S}|^{\underline{\nu}/2}\, 2^{n\bar{\nu}/2}\prod_{i=1}^n \Gamma((\bar{\nu}+1-i)/2)}{|\bar{S}|^{\bar{\nu}/2}\, 2^{n\underline{\nu}/2}\prod_{i=1}^n \Gamma((\underline{\nu}+1-i)/2)}.$$

We included (ι,λ) as a conditioning argument for the MDD. The hyperparameter enters the formula directly and indirectly through $\bar{S}$, $\bar{\Psi}_T$, and $\bar{P}_\Psi$. The estimator type $\iota \in \{mle, lfe\}$ is controlled through the definition of X. If the matrix X stacks y_{t-h}', the formula yields the quasi MDD for the multi-step regression in (8). On the other hand, if one redefines X as the matrix with rows y_{t-1}, one obtains the MDD associated with the VAR in (1).

It is convenient to take log MDD differentials. Because the log density is not defined for $\lambda = 0$, we consider deviations from $\lambda = \infty$. We also multiply the differential by -1, so that the hyperparameter determination is based on the minimization of the differential, just as in the case of PC_T. Define

$$\begin{aligned} MDD_T(\iota,\lambda) &= 2\big[\ln p(Y|\iota,\infty) - \ln p(Y|\iota,\lambda)\big] \\ &= \bar{\nu}\big\{\ln|\bar{S}_T(\iota,\lambda)| - \ln|\bar{S}_T(\iota,\infty)|\big\} + n\big\{\ln|\lambda\underline{P}_\Psi + X'X/T| - \ln|\lambda\underline{P}_\Psi|\big\}, \end{aligned} \tag{31}$$

where

$$\bar{S}_T(\iota,\infty) = \underline{S} + (Y - X\underline{\Psi}_T')'(Y - X\underline{\Psi}_T').$$

The formula highlights dependence of $\bar{S}$ on (ι,λ). The first term in the second line of (31) is a goodness of in-sample fit differential which is scaled so that it can be shown to converge in distribution to a stochastic process indexed by λ. Thus, just as $PC_T(\iota,\lambda)$, the MDD function remains stochastic in the $T \longrightarrow \infty$ limit. The second term is a penalty differential that is ∞ for $\lambda = 0$ and 0 for $\lambda = \infty$. For values of $\lambda > 0$ it converges to a non-stochastic

function of λ. Hyperparameter selection is based on the minimization of $MDD_T(\iota, \lambda)$ with respect to λ. Note that by construction the MDD cannot be used to choose among lfe and mle.

It is well known in the EB literature that the MDD-based hyperparameter selection is less robust to general model misspecifications than the URE-based hyperparameter selection. Recent illustrations of this point in panel settings can be found, for instance, in Kwon (2023) and Cheng, Ho, and Schorfheide (2024). Under the MDD-EB approach hyperparameters are tuned using specific distributional and dynamic assumptions of a hierarchical model and the risk properties of the resulting procedures are inherently sensitive to these assumptions. In our framework, these assumptions are violated for the MLE-based predictor as soon as the VAR is dynamically misspecified and they are violated for the LFE-based predictor even if the DGP is a VAR because the derivation of the MDD criterion ignores the serial correlation of multi-step forecast errors. The PC-EB approach, on the other hand, only uses the VAR model to define a class of estimators and predictors and then chooses the hyperparameter by directly targeting an estimate of the risk function of interest. Thus, it is more robust.

4 Multiple Lags and Lag Length Selection

Companion Form. So far, we considered a VAR in (1) with a single lag. To extend the analysis to multiple lags, we write the VAR in qth-order companion form and then restrict the lag length to $p \leq q$. Thus, q can be considered as the maximum number of lags considered by the researcher. The q-companion form is given by

$$Y_t = \Phi Y_{t-1} + U_t, \quad \Sigma_{UU} = M \Sigma_{uu} M', \tag{32}$$

where

$$\underbrace{Y_t}_{nq\times 1} = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-q+1} \end{bmatrix}, \quad \underbrace{\Phi}_{nq\times nq} = \begin{bmatrix} \phi_1 & \cdots & \phi_{q-1} & \phi_q \\ I_n & \cdots & 0_n & 0_n \\ \vdots & \ddots & \vdots & \vdots \\ 0_n & \cdots & I_n & 0_n \end{bmatrix}, \quad \underbrace{U_t}_{nq\times 1} = \begin{bmatrix} u_t \\ 0_{n\times 1} \\ \vdots \\ 0_{n\times 1} \end{bmatrix}, \quad \underbrace{M}_{nq\times n} = \begin{bmatrix} I_n \\ 0_n \\ \vdots \\ 0_n \end{bmatrix}.$$

Here M is a selection matrix such that $y_t = M' Y_t$. The q-companion form looks identical to (1), except that we replaced lower-case by upper-case variables.

In addition, we define

$$\underbrace{\phi}_{n\times nq} = [\phi_1, \ldots, \phi_q], \quad \underbrace{\Upsilon_q}_{n(q-1)\times nq} = \begin{bmatrix} I_n & \cdots & 0_n & 0_n \\ \vdots & \ddots & \vdots & \vdots \\ 0_n & \cdots & I_n & 0_n \end{bmatrix}, \quad \underbrace{M_{\Upsilon_q}}_{nq\times(n-1)q} = \begin{bmatrix} 0_n & \cdots & 0_n \\ I_n & \cdots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \cdots & I_n \end{bmatrix}. \tag{33}$$

Using this notation, the q-companion form matrix Φ has the following two properties

$$M'\Phi = \phi, \quad M'_{\Upsilon_q}\Phi = \Upsilon_q. \tag{34}$$

Imposing Lag-length Restrictions. The q-companion form matrix of a VAR(p) takes the form

$$\underbrace{\Phi}_{nq\times nq} = \begin{bmatrix} \phi_1 \ \cdots \ \phi_{p-1} \ \phi_p & 0_{n\times n(q-p)} \\ \Upsilon_p & 0_{n(p-1)\times n(q-p)} \\ 0_{n\times n} \ \cdots \ 0_{n\times n} \ I_n & 0_{n\times n(q-p)} \\ 0_{n(q-p-1)\times np} & \Upsilon_{q-p} \end{bmatrix}. \tag{35}$$

To impose the restriction that the coefficient matrices on lags $p+1, \ldots, q$ are equal to zero, we define the selection matrices

$$\underbrace{R_p}_{nq\times n(q-p)} = \begin{bmatrix} 0_{np\times n(q-p)} \\ I_{n(q-p)} \end{bmatrix}, \quad \underbrace{R_{p\perp}}_{nq\times np} = \begin{bmatrix} I_{np} \\ 0_{n(q-p)\times np} \end{bmatrix}. \tag{36}$$

Then the lag length restriction can be expressed as

$$M'\Phi R_p = 0. \tag{37}$$

Moreover, the p companion form coefficient matrix of the VAR(p) is given by $R'_{p\perp}\Phi R_{p\perp}$.

Prior. To construct the MLE shrinkage predictor, we use a prior mean that shares the restrictions of the q-companion form of a VAR(p) in (35), and for $p > 1$ can be written as

$$\underbrace{\underline{\Phi}_T}_{nq\times nq} = \begin{bmatrix} \underline{\phi}_{1,T} \ \cdots \ \underline{\phi}_{p-1,T} \ \underline{\phi}_{p,T} & 0_{n\times n(q-p)} \\ \Upsilon_p & 0_{n(p-1)\times n(q-p)} \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix}, \tag{38}$$

such that it satisfies

$$M'\underline{\Phi}_T = \underline{\phi}_T, \quad \underline{\phi}_T R_p = 0. \tag{39}$$

For $p = 1$ the second line in (38) drops out. The submatrices in the positions marked by $\cdot$ are irrelevant for the subsequent analysis. They would determine the dynamics of lags $p+1, \ldots, q$, which are irrelevant in the forward iteration of a VAR(p). We complete the prior mean specification by simply using the q-companion form entries in (35).

As in the VAR(1) case, we use a prior covariance matrix with a Kronecker structure. Thus, one only needs to specify the $nq \times nq$ hyperparameter-scaled precision matrix $\underline{P}_\phi$. In fact, it suffices to specify the $np \times np$ submatrix that corresponds to the precision of the coefficients for lags 1 through p. As will be shown later, the prior precision for the coefficients associated with lags $p+1$ to q does not affect the first np rows of the posterior mean $\bar{\Phi}_T(mle, \tilde{\lambda}, p)$ and can be set to zero.

To construct the LFE shrinkage predictor we use a prior that satisfies

$$M'\underline{\Psi}_T = \underline{\psi}_T, \quad \underline{\psi}_T R_p = 0. \tag{40}$$

The prior mean for the coefficients $M'_{\Upsilon_q}\Psi$ is irrelevant, because these coefficients are not used in the subsequent analysis.

Posterior Mean. The exposition focuses on $\bar{\Psi}_T(lfe, \lambda, p)$. The posterior mean $\bar{\Phi}_T(mle, \lambda, p)$ can be obtained by setting $h = 1$ and replacing the corresponding prior objects in the following calculations. Define the sums

$$S_{T,0h} = \sum_{t=1}^{T} Y_t Y'_{t-h}, \quad S_{T,hh} = \sum_{t=1}^{T} Y_{t-h} Y'_{t-h}, \quad \bar{S}_{T,0h} = S_{T,0h} + T\lambda \underline{\Psi}_T \underline{P}_\psi, \quad \bar{S}_{T,hh} = S_{T,0h} + T\lambda \underline{P}_\psi.$$

If the coefficients on lags $p+1, \ldots, q$ are restricted to be zero, then one can use the formula for restricted least squared to express the posterior mean for a VAR(p) in q-companion form as

$$\bar{\Psi}_T(lfe, \lambda, p) = \bar{S}_{T,0h}\bar{S}^{-1}_{T,hh}\big[I_{nq} - R_p(R'_p\bar{S}^{-1}_{T,hh}R_p)^{-1}R'_p\bar{S}^{-1}_{T,hh}\big]. \tag{41}$$

In the special case of $p = q$ we have $R_p = 0$ and $\bar{\Psi}_T(lfe, \lambda, q) = \bar{S}_{T,0h}\bar{S}^{-1}_{T,hh}$. The following Lemma summarizes a few key properties of $\bar{\Psi}_T(\cdot)$.

Lemma 1 *(i) The companion form posterior mean has the following property:*

$$\bar{\Psi}_T(lfe, \lambda, p) = \begin{bmatrix} R'_{p\perp}\bar{S}_{T,0h}R_{p\perp}\left(R'_{p\perp}\bar{S}_{T,hh}R_{p\perp}\right)^{-1} & 0 \\ R'_p\bar{S}_{T,0h}R_{p\perp}\left(R'_{p\perp}\bar{S}_{T,hh}R_{p\perp}\right)^{-1} & 0 \end{bmatrix},$$

and $\bar{\Phi}_T(mle, \lambda, p)$ is obtained by replacing $\bar{S}_{T,0h}$ and $\bar{S}_{T,hh}$ by $\bar{S}_{T,01}$ and $\bar{S}_{T,11}$, respectively. (ii) For $p > 1$, rows $n+1$ to np of $\bar{\Phi}_T(mle, \lambda, p)$ take the form $\begin{bmatrix} \Upsilon_p & 0_{n(p-1)\times n(q-p)} \end{bmatrix}$; see (38).

The first part of the Lemma implies that the restrictions (40) also hold for the posterior mean. Moreover, it states that $R_{p\perp}\bar{\Psi}_T(lfe,\lambda,p)R_{p\perp}$ is identical to the posterior that is obtained to estimate a VAR(p) in p-companion form. For the case of $p > 1$, the second part of the Lemma implies the additional result that for the MLE based shrinkage estimator rows $n+1$ to np contain the Υ_p and zeros, i.e., the posterior mean inherits the form of the prior mean in (38). This is important, because the plug-in predictor will be constructed as $[\bar{\Phi}_T(mle,\lambda,p)]^h$; see (6).

Companion-form DGP. The DGP in q-companion form is given by:

$$Y_t = FY_{t-1} + M\epsilon_t + \frac{\alpha}{\sqrt{T}}\sum_{j=1}^{\infty} A_j M\epsilon_{t-j}, \tag{42}$$

where F and $A_1, A_2, \ldots$ are $nq \times nq$ matrices. F has the companion form structure, whereas the A_j matrices are unrestricted. Recall that the asymptotic lag order of the DGP is denoted by p_*. For reasons that become apparent in Section 5 we will assume that p_* is strictly less than q

$$p_* < q. \tag{43}$$

Because of the companion form restrictions, $M'FR_p = 0$ if and only if $p \geq p_*$ from which we deduce the following lemma:

Lemma 2 *For horizons $h \geq 1$, we have $M'F^hR_p = 0$ if and only if $p \geq p_*$.*

Lag Length Selection. The problem of lag length selection for the forecasting application has been discussed in S2005. In principle, one could use the PC criterion or the MDD to jointly select the hyperparameter λ and the number of lags p. Thus, $PC_T(\iota,\lambda)$ in Definition 1 becomes $PC_T(\iota,\lambda,p)$ and $MDD(\iota,\lambda)$ in (31) becomes $MDD(\iota,\lambda,p)$. If $p < p_*$, then the posterior mean has the property that $M'\bar{\Phi}_T R_p = 0$, whereas the DGP, according to Lemma 2, has the property that $MF^hR_p \neq 0$. This creates $O_p(T)$ distortions in the goodness-of-fit components of PC and the MDD. In turn, a lag length of $p < p_*$ will never be selected asymptotically. The likelihood of choosing $p > p_*$ depends on how strongly model dimensionality is penalized. Holding λ fixed, the MDD penalizes model dimension more strongly than PC, and will asymptotically select p_* lags. Just as, for instance, the Akaike criterion, PC has the tendency to favor more complex models and may keep $\hat{p} > p_*$ asymptotically, if the misspecification is large enough to justify the inclusion of additional lags to reduce the asymptotic bias in the forecast error.

5 IRF Estimation

In the companion form VAR in (32), impulse responses to shocks that occurred h periods ago are functions of the hth coefficient matrix of the VMA representation, given by Φ^h. IRF estimates can be obtained in two ways. First, one can estimate a VAR using a likelihood-based (shrinkage) estimator and iterate the estimated VAR forward to trace out the effect of a shock. In our notation, this corresponds to $M'\bar{\Psi}_T(mle, \lambda, p)M$. The companion form coefficient estimate is pre- and post-multiplied by M to guarantee that we focus on the response of y_{t+h} to a shock in period t. Second, LPs estimate Φ^h directly using an h-step-ahead regression. In our notation, this corresponds to $M'\bar{\Psi}_T(lfe, \lambda, p)M$. In Section 5.1 we define the population IRFs and present the large sample distribution of the VAR and LP-based IRF estimators, and discuss their rankings. In Section 5.2 we derive a model selection criterion for (ι, λ, p) that has the property that it provides an asymptotically unbiased estimate of the IRF estimation risk. We continue the notation $\iota \in \{mle,\ lfe\}$ with the understanding that $\iota = mle$ leads to the VAR IRF estimate and $\iota = lfe$ generates the LP IRF estimate.

5.1 Population IRFs and Asymptotics of IRF Estimates

Population IRFs. The DGP in (42) can be rewritten as an MA(∞) process of the form

$$Y_t = \sum_{s=0}^{\infty} F^s M \epsilon_{t-s} + \frac{\alpha}{\sqrt{T}} \left(\sum_{s=0}^{\infty} F^s L^s \right) \left(\sum_{j=1}^{\infty} A_j L^j \right) M \epsilon_t. \tag{44}$$

The true effect of a shock ϵ_{t-h} on y_t is given by the MA coefficient matrix

$$\frac{\partial y_t}{\partial \epsilon'_{t-h}} = M' F^h M + \frac{\alpha}{\sqrt{T}} \mu(irf), \quad \mu(irf) = \sum_{j=0}^{h-1} M' F^j A_{h-j} M, \quad h \geq 1. \tag{45}$$

To simplify the notation, we dropped the IRF horizon h from the $\mu(\cdot)$ argument.

LP IRF Estimates. Iterating the companion form DGP in (42) h periods forward, we can write:

$$S_{T,0h} = F^h S_{T,hh} + \alpha T^{-1/2} \left(\sum_{j=0}^{h-1} \sum_{t=1}^{T} F^j Z_{t-j} Y'_{t-h} \right) + \left(\sum_{j=0}^{h-1} \sum_{t=1}^{T} F^j (M \epsilon_{t-j}) Y'_{t-h} \right).$$

In turn, the standardized LP estimate of the hth-order moving average coefficient matrix, after re-arranging terms, is given by

$$\begin{aligned}
&\sqrt{T}\big(M'\bar{\Psi}_T(lfe,\lambda,p)M - M'F^hM\big) \qquad (46)\\
&= -\sqrt{T}M'F^hR_p(R_p'\bar{S}_{T,hh}^{-1}R_p)^{-1}R_p'\bar{S}_{T,hh}^{-1}M\\
&\quad +M'\big[\delta(lfe,\lambda,p)+\alpha\mu(lfe,\lambda,p)+\zeta_T(lfe,\lambda,p)\big]M+o_p(T^{-1/2}),
\end{aligned}$$

where

$$\begin{aligned}
\mu(lfe,\lambda,p) &= \sum_{j=0}^{h-1}F^j\Gamma_{ZY,h-j}\bar{Q}_p\\
\delta(lfe,\lambda,p) &= \lambda\underline{\Psi}\,\underline{P}_\Psi\bar{Q}_p\\
\zeta_T(lfe,\lambda,p) &= \frac{1}{\sqrt{T}}\left(\sum_{j=0}^{h-1}\sum_{t=1}^{T}F^j(M\epsilon_{t-j})Y_{t-h}'\right)\bar{Q}_p\\
\bar{Q}_p &= (\Gamma_{YY,0}+\lambda\underline{P}_\Psi)^{-1}\left[I_{nq}-R_p(R_p'\left(\Gamma_{YY,0}+\lambda\underline{P}_\Psi\right)^{-1}R_p)^{-1}R_p'\left(\Gamma_{YY,0}+\lambda\underline{P}_\Psi\right)^{-1}\right].
\end{aligned}$$

The terms $\delta(\cdot)$, $\mu(\cdot)$, and $\zeta_T(\cdot)$ are the companion form generalizations of the terms in Theorem 1. According to Lemma 2 the term $\sqrt{T}M'F^hR_p(R_p'\bar{S}_{T,hh}^{-1}R_p)^{-1}R_p'\bar{S}_{T,hh}^{-1}M$ will generate an $O_p(1)$ bias in the estimate $\bar{\Psi}_T(lfe,\lambda,p)$ whenever $p<p_*$. Thus, it is crucial to use a lag length selection criterion with the property that the probability of $\hat{p}<p_*$ goes to zero as $T\longrightarrow\infty$. As in the forecasting application, $\delta(\cdot)$ represents the bias induced by shrinkage, and $\zeta_T(\cdot)$ is a random variable that asymptotically has mean zero and determines the limit variance of the LP estimator. At last, the term $\mu(\cdot)$ captures the misspecification bias. For the multi-step estimation problem we established in (20) that $\mu(lfe,0,p)=\mu(pov,p)$ for $p=1$, which has a straightforward generalization to $p>1$. The following theorem establishes the relationship between $M'\mu(lfe,0,p)M$ and $\mu(irf)$:

Theorem 4 *$M'\mu(lfe,0,p)M=\mu(irf)$ if and only if $p\geq p_*+1$.*

The work by Montiel Olea and Plagborg-Møller (2021) and MPQW showed that it is beneficial to increase the number of lags in the LP by one, compared to a VAR setting. The authors termed this approach lag-augmentation. Theorem 4 reproduces this result in the context of the drifting coefficient DGP considered in this paper. Importantly, Theorem 4 is explicit about the benchmark relative to which lags have to be added to properly center the LP estimate: it is the asymptotic lag order p_* of the drifting DGP. The theorem implies

that in terms of bias there is neither a cost nor a benefit to increasing the number of lags beyond $p_* + 1$.

VAR IRF Estimates. The VAR based impulse response estimate for horizon h is given by $M'\bar{\Psi}_T(mle, \lambda, p)M$. Its limit can be decomposed as in (46), where $\mu(\cdot)$, $\delta(\cdot)$, and $\zeta_T(\cdot)$ are the companion form generalizations of the MLE terms in Theorem 1.

Shock Identification. The structural VAR literature is typically not interested in responses to the one-step-ahead forecast errors, denoted by u_t in the VAR models in (1) and (32) or by ϵ_t in the DGPs (13) and (42). Instead, it focuses on structural shocks e_t that are related to the one-step-ahead forecast errors by, say, $\epsilon_t = \Sigma_{\epsilon\epsilon}^{tr}\Omega e_t$, where $\Sigma_{\epsilon\epsilon}^{tr}$ is the lower-triangular Cholesky factor of $\Sigma_{\epsilon\epsilon}$, Ω is an orthogonal matrix, and $e_t \sim (0, I)$. It can be shown that the error due to misspecification when using the VAR(p) residuals to estimate the reduced-form covariance matrix $\Sigma_{\epsilon\epsilon}$ is $O_p(T^{-1})$. This implies that if the structural VAR is point or set-identified based only on restrictions on the contemporaneous effect of the structural shocks, then the effect of the local misspecification on the determination of Ω is of smaller order than the $O(T^{-1/2})$ bias of $\bar{\Psi}(\cdot)$, and hence negligible. Alternatively, if the choice of Ω is also based on the dynamic effects of the structural shocks, as is the case for point identification based on long-run restrictions or set-identification based on sign restrictions imposed over multiple horizons, then the asymptotic bias of $\bar{\Psi}(\cdot)$ would create an $O_p(T^{-1/2})$ distortion of Ω.

In the remainder of this section, we abstract from the contamination effect and ignore possible $O_p(T^{-1/2})$ errors in the determination of Ω. In many applications the object of interest is the response to a subset of the structural shocks, rather than all structural shocks. Without loss of generality, we assume that the n_{sh} shocks of interest are ordered first and denote the first n_{sh} columns of the matrix $\Sigma_{\epsilon\epsilon}^{tr}\Omega$ by the $n \times n_{sh}$ matrix Ξ and simply treat Ξ as known.

IRF Estimation Risk. The asymptotic IRF estimation risk takes the form

$$\begin{aligned}\bar{\mathcal{R}}_{IRF}(\iota, \lambda, p) &= \lim_{T\longrightarrow\infty} T \cdot \mathbb{E}\left[\left\| M'\bar{\Psi}_T(\iota, \lambda, p)M\Xi - \left[M'F^hM + \frac{\alpha}{\sqrt{T}}\mu(irf)\right]\Xi\right\|_W^2\right] \quad (47)\\ &= \left\| M'\delta(\iota, \lambda, p)M\Xi + \alpha(M'\mu(\iota, \lambda, p)M - \mu(irf))\Xi\right\|_W^2 \\ &\quad + \mathrm{tr}\Big\{\left[(MWM')\otimes(M\Xi\Xi'M')\right]V(\iota, \lambda, p)\Big\}.\end{aligned}$$

Specific formulas for variances and covariances $V(\cdot)$, generalizing Theorem 1 to $p > 1$, are provided in (A.19) of the Online Appendix. The second equality is based on (46) and Lemma A-2 in the Online Appendix.

Remarks. First, consider the case $\lambda = 0$ and $p = p_*$. The generalizations of the variance formulas in Theorem 1 to $p > 1$ imply that the LP-based estimate always has higher variance than the VAR-based estimate; see S2005. Moreover, using the asymptotic lag length p_* for VAR and LP, there is no clear ranking of $\|M'\bar{\Psi}_T(lfe, \lambda, p_*)M - \mu(irf)\|$ and $\|M'\bar{\Psi}_T(mle, \lambda, p_*)M - \mu(irf)\|$, which means that even under large α misspecification, the VAR-based IRF estimator may dominate the LP estimator. This is broadly consistent with the simulation experiments reported by LPW (Figures 2 and 3) where the authors illustrate that their LP estimates dominate the VAR estimates in terms of bias, but the VAR estimates have lower variance, leading to ambiguous rankings. LPW emphasize that the preferred estimator largely depends on how much one trades off variance and bias and conclude that for researchers who also care about precision, VAR methods are the most attractive. We will come back to this point when we consider a numerical illustration of the asymptotic risk formulas.

Second, MPQW emphasize that lag augmentation, beyond the order that a consistent model selection criterion would recommend, is important to reduce the bias of LP estimators. Suppose that $\lambda = 0$ and $p > p_*$. Proposition 3.1 in MPQW applies also to our setting and implies that the LP estimation risk is identical for all $p > p_*$. We saw in Theorem 4 that the bias does not change by increasing the number of lags. In addition, the variance remains constant as well. Moreover, according to Corollary 3.2 of MPQW the VAR-based IRF estimation risk does not depend on p and is identical to the LP risk for $\lambda = 0$ and horizons $h \leq p - p_*$. These two properties of the LP IRF estimates are quite different from the properties of the LFE-based predictor. In the foreasting application additional lags always weakly reduce the bias relative to the true conditional expectation,[4] but also increase the variance of the predictor.

Third, once $\lambda > 0$, the ranking in the overall bias of the two types of IRF estimators becomes ambiguous again:

$$\text{asymptotic bias} = M'\delta(\iota, \lambda, p)M + \alpha\big(M'\mu(\iota, \lambda, p)M - \mu(irf)\big), \tag{48}$$

[4]This is not immediately apparent from the equations provided in Sections 2 and 4. But notice that as p is increased, the constant C in Theorem 2 weakly decreases because the true conditional expectation of $\tilde{y}_{T+h}$ is projected onto a larger space.

because it is no longer clear that choosing ι to set the second term to zero is optimal. Furthermore, Ludwig (2024) warns that comparing LPs and VARs of the same order p may lead to misleading rankings. All in all, the interplay between (ι, λ, p) is crucial, and our joint selection is able to balance the various trade-offs. This contrasts with the problem of constructing confidence intervals for IRFs, studied in MPQW, where the incorrect centering of the VAR-based estimates can lead to substantial distortions of coverage probabilities.

5.2 Asymptotically Valid URE for the IRF Estimation Risk

Given that LP estimates do not uniformly (in terms of misspecification) dominate VAR-based IRF estimates, it is natural to employ a method like the PC criterion in Definition 1 to find the preferred IRF estimator, akin to what has been proposed above for multi-step point prediction. However, the criterion needs to be carefully tailored toward the IRF estimation risk. By generalizing the calculations in the proof of Theorem 1 to the restricted companion form representation, one can show that for $\iota \in \{mle, lfe\}$, $\lambda \geq 0$, $p_* < q$, and $p_* \leq p \leq q$,

$$\begin{aligned} &\sqrt{T}\left(\bar{\Psi}_T(\iota,\lambda,p) - \bar{\Psi}_T(lfe,0,q)\right) \\ &\implies \mathcal{N}\Big(\delta(\iota,\lambda,p) + \alpha(\mu(\iota,\lambda,p) - \mu(lfe,0,q)), \\ &\qquad\qquad V(\iota,\lambda,p) + V(lfe,0,q) - 2Cov(lfe,0,q;\iota,\lambda,p)\Big). \end{aligned} \tag{49}$$

We proceed by following the calculations for PC_T^* in Section 3.1, following Definition 2. Theorem 4 in combination with the assumption that $p_* < q$ lets us replace $M'\mu(lfe,0,q)M$ by $\mu(irf)$. Thus, the expectation of the coefficient norm difference for the impulse response matrix behaves as follows:

$$\begin{aligned} &\lim_{T\longrightarrow\infty} T\cdot\mathbb{E}\left[\left\|M'\left(\bar{\Psi}_T(\iota,\lambda,p) - \bar{\Psi}_T(lfe,0,q)\right)M\Xi\right\|_W^2\right] \\ &= \left\|M'\delta(\iota,\lambda,p)M\Xi + \alpha(M'\mu(\iota,\lambda,p)M - \mu(irf))\Xi\right\|_W^2 \\ &\quad + \mathrm{tr}\Big\{\left[(MWM')\otimes(M\Xi\Xi'M')\right]\left(V(\iota,\lambda,p) + V(lfe,0,q) - 2Cov(lfe,0,q;\iota,\lambda,p)\right)\Big\}. \end{aligned} \tag{50}$$

This equation resembles (26). In fact, the only major difference is the weighting applied to the $\bar{\Psi}_T(\cdot)$ differential. A more subtle difference is the need for a strict inequality on $p_* < q$, instead of a weak one. By construction, the squared bias terms in (50) is identical to that in (47). The variance term $V(lfe,0,q)$ is independent of (ι,λ,p) and does not affect the ranking

of estimators. Thus, we deduce that the following criterion provides, up to a constant that does not depend on (ι, λ, p), an asymptotically unbiased estimate of the large sample risk:

Definition 3 *Define the $IRFC_T(\iota, \lambda, p)$ criterion for the joint selection of IRF estimator, shrinkage, and lag length, with impact vector Ξ, as*

$$IRFC_T(\iota, \lambda, p) = T \left\| M' \left(\bar{\Psi}_T(\iota, \lambda, p) - \bar{\Psi}_T(lfe, 0, q) \right) M\Xi \right\|_W^2 + 2\hat{\mathcal{R}}_{Cov}(lfe, 0, q; \iota, \lambda, p),$$

where $\hat{\mathcal{R}}_{Cov}(lfe, 0, q; \iota, \lambda, p)$ has the property that

$$\mathbb{E}[\hat{\mathcal{R}}_{Cov}(lfe, 0, q; \iota, \lambda, p)] \longrightarrow \operatorname{tr}\left\{ (MWM' \otimes M\Xi\Xi'M')\, Cov(lfe, 0, q; \iota, \lambda, p) \right\}.$$

After combining Definition 3 with (47) and (50) we can deduce that

$$\mathbb{E}\big[IRFC_T(\iota, \lambda, p) - IRFC_T(\iota', \lambda', p')\big] \longrightarrow \bar{\mathcal{R}}_{IRF}(\iota, \lambda, p) - \bar{\mathcal{R}}_{IRF}(\iota', \lambda', p'). \tag{51}$$

This result extends (25) to IRF estimation with an unknown number of lags. Note that if $p < p_*$ the $IRFC_T$ will diverge. Intuitively, one would expect that $IRFC_T$ will on average decrease if p is raised to $p_* + 1$ because the asymptotic bias is reduced. A further increase in the number of lags cannot improve the asymptotic bias component. We explore these trade-offs numerically by evaluating the asymptotic estimation risk formulas for various configurations of the DGP.

6 Numerical Illustration of Asymptotic Risks

We now provide a numerical illustration of the asymptotic risk formulas derived in Sections 2 to 5. We first specify a DGP for the numerical analysis in Section 6.1, then discuss asymptotic forecasting risk in Section 6.2, and examine the IRF estimation risk in Section 6.3.

6.1 DGP and Prior

Data Generating Process. The DGP is given by (13). We consider an $n = 6$ variable VAR. The coefficient matrix F and error variance matrix Σ_ϵ are calibrated to an estimated VAR(1) on the same variables as those used in Carriero, Clark, and Marcellino (2015). The entries of the MA drift matrices $\{A_j\}_{j=1}^{10}$ are drawn independently from a standard normal distribution. MA matrices of order $j > 10$ are set equal to zero. Most important for the

interpretation of the results is that we maintain the drifting structure of the DGP as we vary the sample size T. The expected loss calculation is frequentist: we keep the parameters of the DGP fixed as we repeatedly generate data and evaluate the loss associated with various prediction procedures.

Prior Distributions. For the interpretability of the results it is important that we align the local prior means $\underline{\phi}$ and $\underline{\psi}$ in (38) and (40) with respect to their implication about the h-step-ahead prediction function. We start by setting the prior mean $\underline{\psi}$ of the local deviation from F^h equal to a multiple of the pov:

$$\underline{\psi} = \varphi\mu(pov, p_*). \tag{52}$$

The parameter φ controls the distance between the prior center and the pov under misspecification. Using a first-order Taylor expansion of $\Phi^h - F^h$, we obtain

$$\Phi^h - F^h \approx \sum_{j=0}^{h-1} F^j(\Phi - F)F^{h-1-j}.$$

In turn, we would like to choose the (local) prior mean for the MLE such that it satisfies

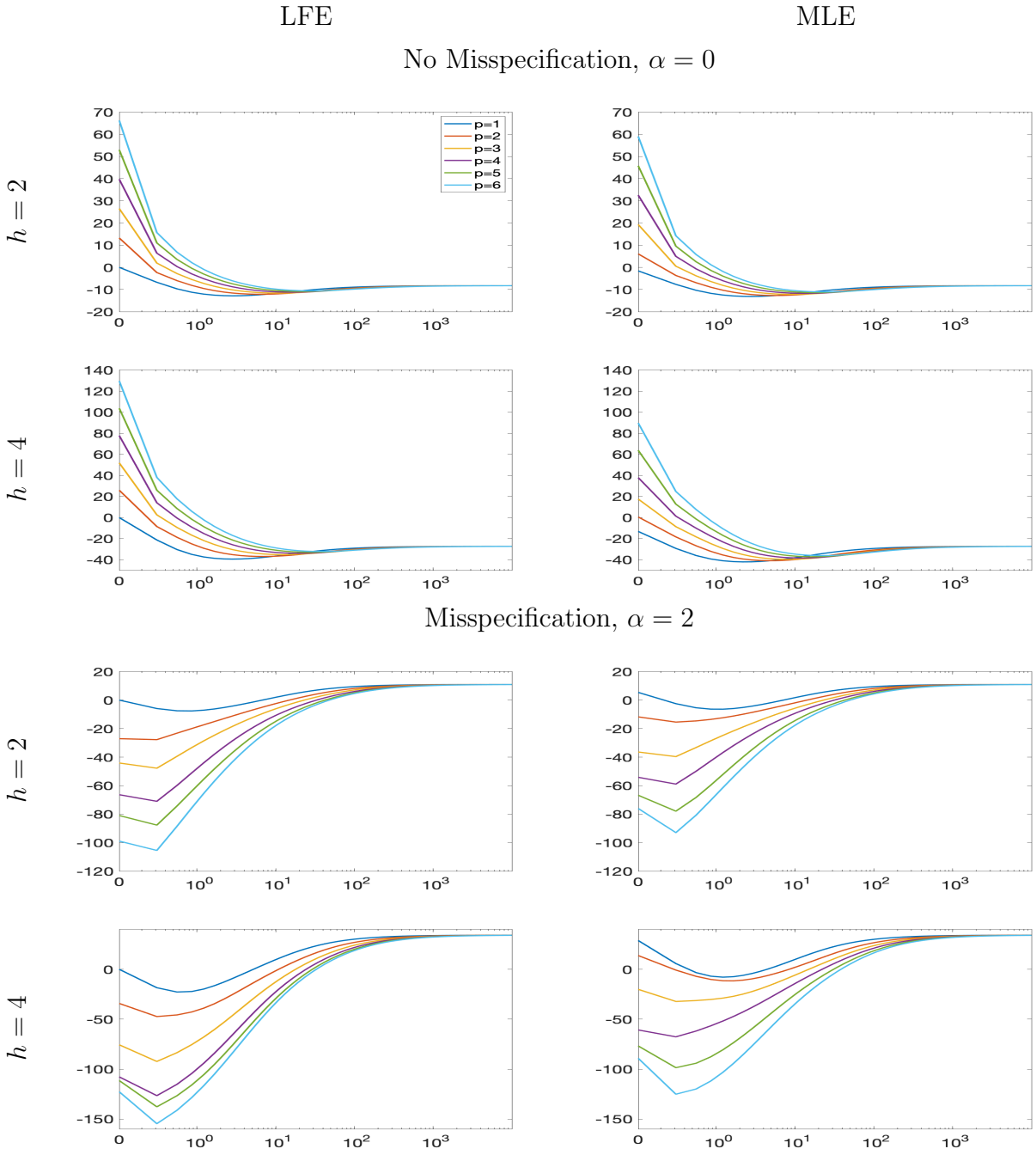
$$\underline{\psi} = \sum_{j=0}^{h-1} F^j \underline{\phi} F^{h-1-j}. \tag{53}$$

Experimental Designs. We consider two different experimental designs. Under Design 1 $\alpha = 0$ and there is no misspecification. By setting $\varphi = 1$ we ensure that the prior is not centered at the "true" value, which in the case of correct specification would correspond to $\varphi = 0$. Under Design 2 $\alpha = 2$, the VAR is misspecified, and we center the prior at $\varphi = 0.5$ to keep it away from the pov.

6.2 Multi-Step-Ahead Forecasting

In the forecasting exercise, normalization of the risk is needed to align the expression in Theorem 2, which has a constant C, with the PC criterion in Definition 1. Without loss of generality, we normalize the risk with respect to the $\lambda = 0$ LFE predictor at the true $p_* = 1$, such that the risk corresponding to $\hat{y}_{T+h}(lfe, 0, 1)$ is zero. Figure 1 depicts the asymptotic risk differentials $\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota, \lambda, p)) - \bar{\mathcal{R}}(\hat{y}_{T+h}(lfe, 0, 1))$ as a function of λ for $\iota \in \{lfe, mle\}$, respectively. We consider two levels of misspecification α and two horizons h.

Figure 1: Asymptotic Forecasting Risk



Notes: The x-axis is the hyperparameter λ on a logarithmic scale with zero as the left endpoint. On the y-axis we plot $\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota, \lambda, p)) - \bar{\mathcal{R}}(\hat{y}_{T+h}(lfe, 0, 1))$. Different lines correspond to different lag orders p.

For $\lambda = \infty$, MLE and LFE are equal to the prior mean values. Because of (53), the resulting predictors are equivalent up to first order and have identical risks. Moreover, we

keep the means fixed as we increase the lag length. Thus, the risk of any given predictor converges to the same value across different p as $\lambda \to \infty$.

The top two rows of Figure 1 contain results for the no-misspecification case $\alpha = 0$. The MLE dominates the LFE conditional on (λ, p) because it is more efficient and there is no misspecification bias. As the horizon increases from $h = 2$ to $h = 4$, the benefit of using the MLE also increases. For $\alpha = 2$ the prediction-risk-based ordering of the estimators changes: the LFE visibly dominates the MLE at all horizons for small and moderate values of λ at a given lag length. As the precision λ increases further the parameter estimates are dominated by the prior and the risk differential vanishes.

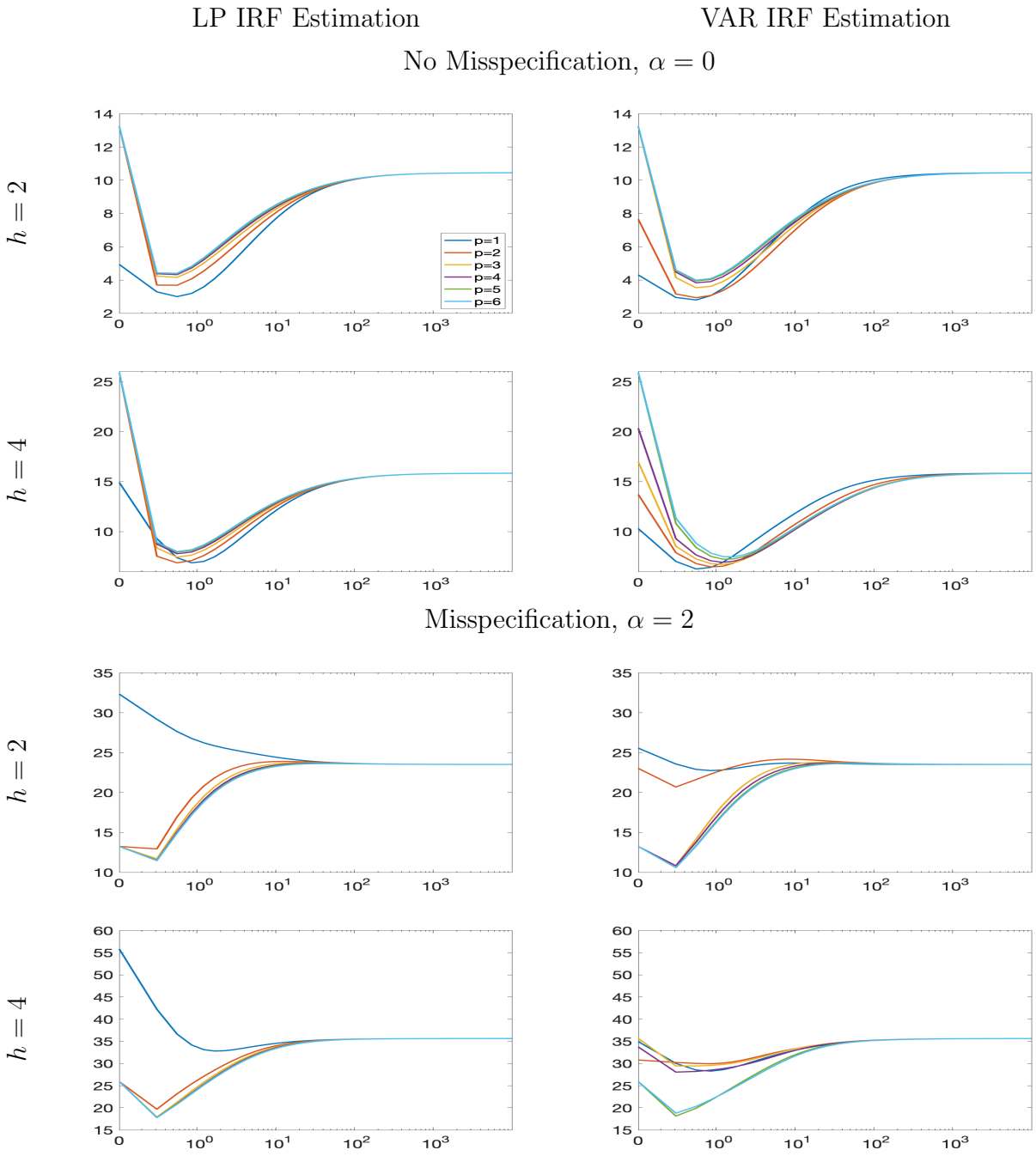
Under the two designs, the optimal level of shrinkage is always non-trivial in the sense that the minimum risk is obtained for an interior value of λ. The benefit of shrinkage is dependent on the lag order, but substantial throughout. For $\alpha = 0$ the degree of shrinkage increases with lag p. Increasing the lag order is undesirable for both predictors in the absence of misspecification. The lowest prediction risk is obtained for $p = p_* = 1$. On the other hand, if the VAR is misspecified with $\alpha = 2$, increasing the lag order above $p_* = 1$ reduces the misspecification bias.

6.3 IRF Estimation

We proceed by comparing the asymptotic risk of VAR and LP IRF estimators, constructed for various lag lengths. The parametrization is the same as for the forecasting exercise in Section 6.2, and we set $\Xi = I_n$. Because a normalization is not needed in the IRF case, we present the actual asymptotic risk values instead of differentials in Figure 2. As before, the prior means $\underline{\Psi}$ and $\underline{\Phi}$ are chosen such that they imply identical IRFs. In turn, as $\lambda \longrightarrow \infty$ the estimation risk is identical for all (ι, p).

The two top rows of Figure 2 show $\bar{\mathcal{R}}_{IRF}(\iota, \lambda, p)$ for the no-misspecification case $\alpha = 0$. At $\lambda = 0$ the choice of $p = 1 = p_*$ is optimal. Because the VAR estimation is more efficient than multi-step regressions, the IRF estimation risk for $p = 1$ is smaller for the VAR than for the LP. The VAR-LP risk differential increases with horizon h. As the number of lags is increased, the asymptotic risk also increases, both for the VAR and the LP estimation. The LP estimation risk is identical for all $p > p_*$ because bias does not change (Theorem 4) and the variance does not increase (Proposition 3.1 in MPQW). The VAR IRF estimation risk is initially increasing in p. Once $p \geq p_* + h$ the risk stays constant and is identical to the LP risk (Corollary 3.2 of MPQW).

Figure 2: Asymptotic IRF Estimation Risk



Notes: The x-axis is the hyperparameter λ on a logarithmic scale with zero as the left endpoint. On the y-axis we plot $\bar{\mathcal{R}}_{IRF}(\iota, \lambda, p)$. Different lines correspond to different lag orders p.

For both LFE and MLE, some shrinkage is always desirable. Conditional on choosing $\hat{\lambda}(p)$, the level of risk is reduced substantially compared to $\lambda = 0$, and the risks associated

with the different lag lengths $p \in \{1, \ldots, 6\}$ become very similar. For $h = 2$ a single lag remains optimal for the LP and VAR estimates. The minimum risk is attained for $\hat{\lambda} \approx 0.5$. For $h = 4$ the risk associated with the LP IRF estimator $(\iota = lfe, \lambda = 0.5, p = 2)$ is approximately equal to the risk of $(\iota = lfe, \lambda = 1, p = 1)$. The VAR IRF estimate $(\iota = mle, \lambda = 0.6, p = 1)$ leads to a slightly lower risk than the best LP estimates.

We now turn to the bottom two rows of Figure 2 which show asymptotic risk functions for the misspecified case of $\alpha = 2$. In terms of lag length selection the choice of $p = 1$ now leads to the largest risk. This is true for the LP and the VAR IRF estimator and holds for $h = 2$ and $h = 4$. The LP estimator benefits from the lag augmentation effect for $p > p_* = 1$. At $\lambda = 0$ the risk drops substantially as the number of lags is increased from $p = 1$ to 2. While there are no further risk reductions for $p > 2$ at $\lambda = 0$, shrinking toward the prior mean leads to additional performance improvements. In our setting the optimal degree of shrinkage is approximately $\hat{\lambda}(p) \approx 0.2$ for $p > 1$. The overall pattern is independent of the horizon h.

The effect of lag length changes on the risk of the VAR IRF estimator is different from the LP IRF estimator, as can be seen from right panels in the two bottom rows of Figure 2. Starting with $h = 2$ and $\lambda = 2$, increasing the lag length from $p = 1$ to 2 reduces the misspecification bias. For $p > 2$ the condition $p \geq p_* + h$ is satisfied and the risk of $(\iota = mle, \lambda = 0, p)$ is identical to $(\iota = lfe, \lambda = 0, p)$. For $h = 4$ the estimation risk falls if p is increased from 1 to 2, but then increases again for $p = 3, 4$. While additional lags reduce the misspecification bias, they also increase the variance of the estimator. At $p = 5$ the condition $p \geq p_* + h$ is met and the LP risk is achieved. Shrinkage further lowers the risk of all estimators.

In our design, the overall performance of LP and VAR IRF estimators is very similar, once an optimal lag length and an optimal shrinkage parameter λ are chosen. IRF estimation differs from multi-step forecasting in regard to the effect of a lag length increase on the variance. The key difference is that forecasting utilizes the estimated coefficients for all lags, whereas the IRF analysis only depends on the estimated MA coefficient matrix of order h, which is an object that does not grow with p. In the forecasting application, an increase of p always raises the variance of the prediction function, while simultaneously reducing the asymptotic bias, which leads to a clear trade-off. The PC and IRFC criteria are designed to balance the various trade-offs in a data-driven manner.

7 Multi-Step Forecasting with Simulated Data

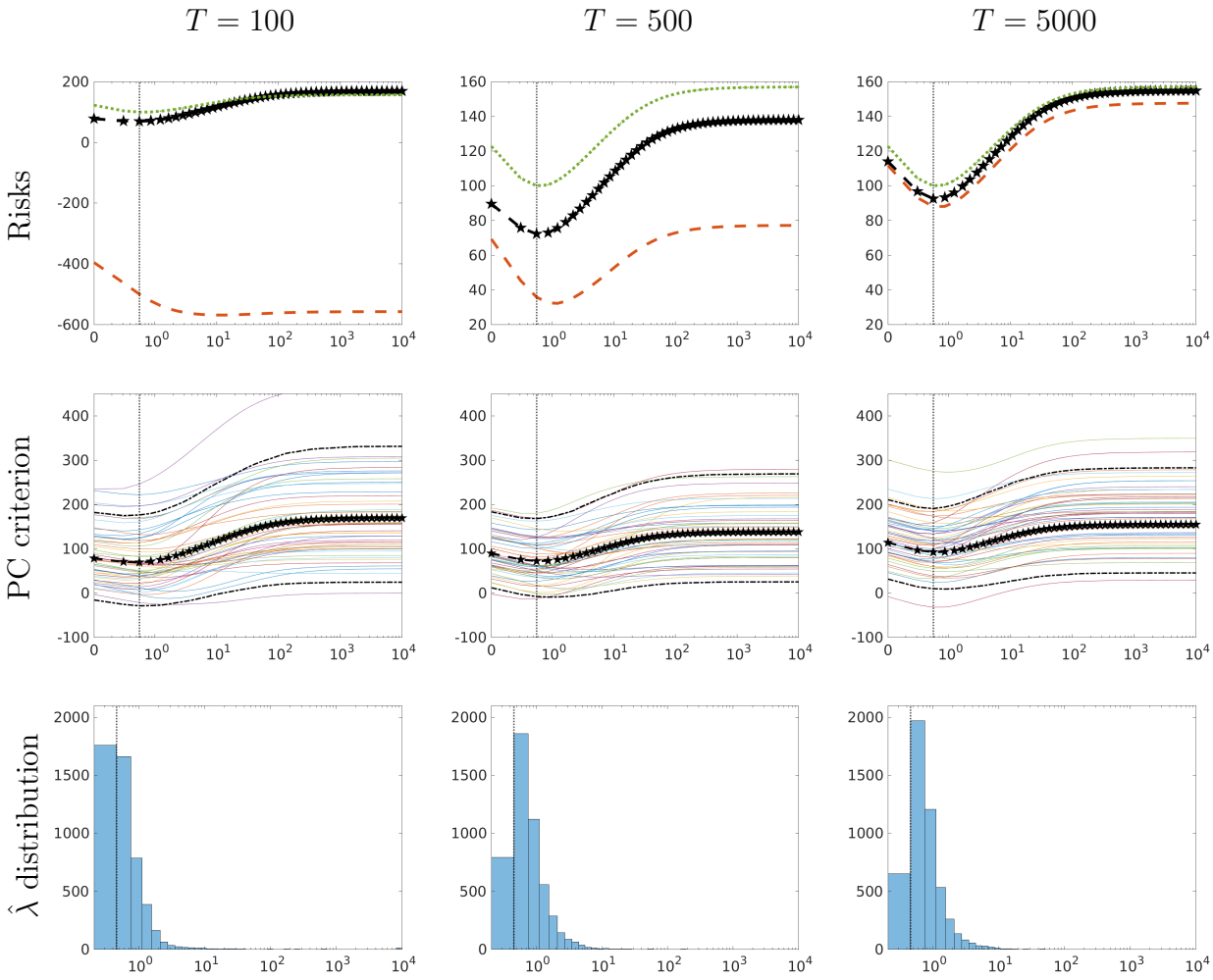
To document the finite-sample behavior of the proposed econometric procedures we conduct a small-scale Monte Carlo study experiment that illustrates the forecasting performance of the MLE and LFE based shrinkage predictors. A large-scale empirical analysis, albeit without consideration of shrinkage estimation was done by Marcellino, Stock, and Watson (2006). In Section 7.1, we compare the finite-sample risk differentials to the expected value of PC_T. We consider the joint PC-based selection of shrinkage, lag length, and predictor in Section 7.2. Finally, we provide a comparison of the risks associated with PC versus MDD-based model determination in Section 7.3. Throughout, the DGP is identical to the one used to compute the asymptotic risk functions in Section 6.

7.1 Simulated Risk Differentials, Expected PC, and λ Selection

The panels in the top row of Figure 3 depict Monte Carlo risk differentials (dashed red), asymptotic risk differentials (dotted green), and the expected value of PC (starred black), respectively. In this section we again normalize the risks with respect to $\hat{y}_{T+h}(lfe, 0, q)$ to difference out the constant C, as explained in Section 6.2. The figure shows results for $\alpha = 2$, $h = 4$, and $p = p_* = 1$. The sample size varies across columns. Starting with $T = 5,000$ in the right column, the two risk measures and the expected value of PC are well aligned. As a function of λ the three criteria attain their respective minima at an interior value $\hat{\lambda} \approx 0.7$. The vertical lines indicate the values of λ that minimize the asymptotic risk. The asymptotic risk differential curve of the LFE is decreasing between $\lambda = 0$ and the minimum of 0.7, and then increases strongly as λ approaches 100. For values of $\lambda > 100$ the curve is fairly flat. For smaller sample sizes the shape of the three functions remains very similar, but there is a level discrepancy. The largest discrepancy arises for $T = 100$ between the Monte Carlo risk on the one hand, and the asymptotic risk and the expected value of PC on the other hand. This discrepancy is caused by a large wedge between the Monte Carlo variance of the estimated coefficients and their asymptotic variance.

In the second row of Figure 3 we plot hairlines of the $PC_T(\iota = lfe, \lambda, p = p_*)$ objective functions. Each hairline corresponds to a particular Monte Carlo repetition. The collection of hairlines illustrates the sampling variation of the PC objective function. The solid black lines with the star symbols represent the pointwise expected values of the objective function and hence approximate $\mathbb{E}[PC_T]$ as a function of λ. They are identical to the black lines

Figure 3: PC versus Finite Sample Risk, LFE, $\alpha = 2$, $h = 4$, $p = p_* = 1$



Notes: The x-axis is the hyperparameter λ on a logarithmic scale with zero as the left endpoint. Top row: asymptotic risk (dotted green), MC risk (dashed red), and $\mathbb{E}[PC_T]$ (starred black). Center row: hairlines represent PC objective function across MC replications; starred black line is $\mathbb{E}[PC_T]$ (same as in top row); dashed black lines are 90% bands. Bottom row: distribution of PC-selected shrinkage hyperparameter. The vertical lines indicate the value of λ that minimizes the asymptotic risk.

in the top row. Overall, the hairline pattern is broadly consistent with the asymptotic risk differential function. Most of the hairlines attain their minimum between $\lambda = 0.5$ and $\lambda = 10$ and are monotonically increasing to the right of the minimum. However, in particular for $T = 100$ there are hairlines that are monotonically increasing over the entire domain of λ, which leads to $\hat{\lambda} = 0$. In the bottom row of Figure 3 we plot histograms of the PC-selected $\hat{\lambda}$ distribution across Monte Carlo repetitions. Most of the mass concentrates near the argmin of the asymptotic risk function, but the distribution is skewed to the right with a small

Table 1: Finite Sample Risk Differentials for $\hat{y}_{T+h}(\iota, \hat{\lambda}, p)$, $T = 500$.

	$\alpha = 0$				$\alpha = 2$			
p	LFE	MLE	Joint	π	LFE	MLE	Joint	π
Horizon $h = 2$								
1	-91	-91	-91	39	55	59	55	94
2	-90	-91	-90	25	37	46	37	100
4	-88	-89	-89	21	-1	10	-1	100
6	-87	-88	-88	21	-26	-15	-26	99
$\hat{p}$	-89	-89	-89	29	-26	-15	-26	99
Horizon $h = 4$								
1	-194	-197	-195	39	39	51	40	90
2	-190	-196	-193	30	25	43	26	96
4	-185	-191	-189	27	-40	10	-40	100
6	-182	-187	-185	27	-58	-33	-58	97
$\hat{p}$	-188	-191	-189	25	-57	-32	-57	99
Horizon $h = 6$								
1	-300	-310	-305	26	-16	7	-15	92
2	-293	-308	-302	19	-36	4	-35	92
4	-284	-301	-297	15	-85	-26	-82	94
6	-277	-296	-290	13	-87	-57	-82	86
$\hat{p}$	-289	-299	-294	13	-84	-54	-82	93

Notes: The finite sample risk differentials are computed relative to $\hat{y}_{T+h}(lfe, 0, q = 6)$. π is the percentage of times that LFE is selected by the PC criterion.

pointmass at $\lambda = \infty$ which vanishes as the sample size T increases.

7.2 Joint PC-Based Determination of ι, λ, and p

We now examine the Monte Carlo risk of LFE and MLE predictors based on data-driven hyperparameter and lag length choice. Table 1 shows risk differentials between $\hat{y}_{T+h}(\iota, \hat{\lambda}, p)$ and $\hat{y}_{T+h}(lfe, 0, q = 6)$ for $T = 500$, where $\hat{\lambda}$ is determined by minimizing $PC_T(\iota, \lambda, p)$ with respect to λ. We report results for $\iota = lfe$, $\iota = mle$ and also use PC to choose among LFE and MLE and report the resulting risks in the column "Joint." We consider the lag lengths

1, 2, 4, and 6. In the rows labeled $\hat{p}$ PC is also used to select the lag length. Negative numbers indicate risk reductions relative to the benchmark predictor $\hat{y}_{T+h}(lfe, 0, q)$.

Columns 2 to 5 of the table contain results for the no-misspecification case of $\alpha = 0$. As expected, the MLE predictor attains a lower risk than the LFE predictor. The asymptotic bias for both predictors is equal to zero and the likelihood-based estimation is more efficient than the loss-function-based estimation. The differential is small for $h = 2$, but increases with the forecast horizon. The lowest risk is attained for $p = 1$ because that equals the true asymptotic lag length p_*. The PC criterion selects the MLE predictor between 60% and 85% of the time and the risk under "joint" selection is in between the LFE and MLE risk, often closer to the latter. The risk values associated with $\hat{p}$, i.e., the case in which PC is also used to choose the lag length, are between the $p = 2$ and $p = 4$ values, indicating that the criterion tends to select models that are somewhat larger than p_*. However, stronger shrinkage to some extent compensates for the additional lags and reduces risk differentials which was also apparent from Figure 1.

Under misspecificaton $\alpha = 2$ in columns 6 to 9 the ranking of LFE and MLE predictors is reversed: LFE clearly dominates and for most settings is selected by PC in more than 95% of the Monte Carlo repetitions. For both MLE and LFE it is desirable to include more than $p_* = 1$ lags to offset the dynamic misspecification of the VAR. The $\hat{p}$ risk differentials as are close to the $p = 6$ risk differentials, indicating that PC selects six lags with high probability. Because of the pronounced risk differentials between LFE and MLE the risk associated with "joint" selection of predictor and hyperparameter is essentially identical to the LFE risk. Overall, the results reported here are consistent with the simulation results in S2005 and the empirical results in Marcellino, Stock, and Watson (2006) in that the use of the LFE is only justified if the misspecification is sufficiently large. The key takeaway is that PC helps the forecaster to adapt to the level of misspecification by tuning the level of shrinkage λ, selecting the number of lags p, and choosing between LFE and MLE predictor.

7.3 PC versus MDD-Based Model Determination

In Table 2 we compare the risk associated with PC versus MDD based (λ, p) selection. Negative numbers are risk reductions of PC selection relative to MDD selection in percent. In the absence of VAR misspecification, i.e., $\alpha = 0$ (correct specification) the performance of PC and MDD selection is very similar and the percentage improvements or deteriorations are in

Table 2: Finite Sample Risk Differentials, PC vs. MDD Selection, $T = 500$

	Horizon $h = 2$				Horizon $h = 4$			
	$\alpha = 0$		$\alpha = 2$		$\alpha = 0$		$\alpha = 2$	
p	LFE	MLE	LFE	MLE	LFE	MLE	LFE	MLE
1	-3	0	-5	-5	-7	0	-13	-18
2	2	2	-4	-11	0	4	34	-3
4	2	0	-102	-86	5	0	-368	-86
6	2	1	-137	-122	5	1	-181	-143
$\hat{p}$	-1	3	-145	-125	-3	4	-226	-152

Notes: We report risk differentials of PC-based versus MDD-based selection relative to the MDD risk, in percent. A negative number indicates that PC selection yields a lower risk than MDD selection.

the single digits. For $\alpha = 2$, on the other hand, the use of PC for hyperparameter determination leads to drastic risk reductions. The likelihood function from which the marginal data density is derived is now misspecified and the use of a risk estimate or targeted information criterion for model determination becomes highly desirable.

We also computed the distribution of the selected lag length $\hat{p}$ for PC and MDD[5] and provide a verbal summary of the results. For $\alpha = 0$ and $\alpha = 2$ the MDD selects $\hat{p} = 1$ in all of the Monte Carlo repetitions. This is the case for MLE and LFE. In case of MLE this is not surprising because MDD model selection has the property that it is "consistent," which implies in the context of the drifting DGP that it selects the asymptotic lag-order p_* with probability approaching one as the sample size increases. The PC tends to select more than p_* lags. Regardless of predictor, if the forecasting model is correctly specified ($\alpha = 0$), PC selects more than one lag in 30% to 50% of the Monte Carlo repetitions. Under misspecification, PC chooses the maximum lag length $q = 6$ in all Monte Carlo repetitions, using the additional lags to reduce the misspecification bias. As mentioned previously, PC can also be used to choose between LFE and MLE which is something MDD is not designed to do.

[5]See Figure A-1 in the Online Appendix.

8 Empirical Application: IRF Estimation

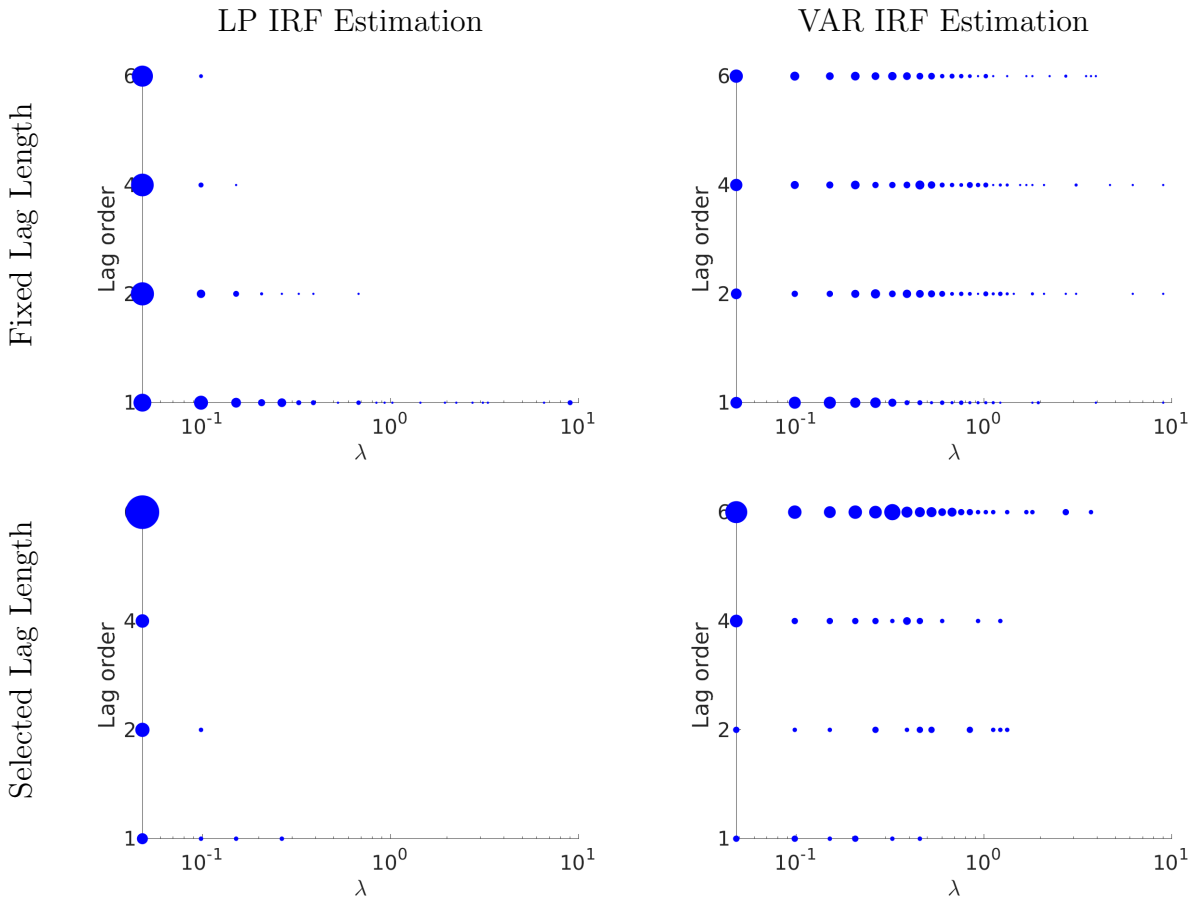
We now estimate IRFs via VARs or LPs using VARs constructed from actual data. While in a pseudo-out-of-sample forecasting application it is possible to compute *ex-post* forecast errors from the actual data, one cannot compute estimation errors in an IRF application because the "true" IRFs are never observed. Thus, we proceed by documenting how often IRFC selects the VAR and LP IRF estimates, respectively, and how much shrinkage and how many lags are used.

Samples for Empirical Estimates. We construct estimation samples by combining time series from the FRED-QD database; see McCracken and Ng (2020). We filter each series using the procedure proposed by Hamilton (2018) to induce stationarity, yet preserve a lot of the persistence, so that shocks can have long-lasting effects. We follow Marcellino, Stock, and Watson (2006) in that we are randomly creating a large number of data sets. We do so by selecting uniformly at random 200 different six-tuples of series, which are demeaned and standardized. For each of the data sets we estimate $n = 6$-dimensional VAR or LP with $p \in \{1, 2, 4, 6\}$ lags. Because many of the time series are persistent, the prior for the VAR coefficients is centered at univariate unit-root representations ("Minnesota" prior). Because the series are combined at random, the degree of misspecification varies across samples. Identification (or shock orthogonalization) is achieved by using the first column of the Cholesky factorization of the one-step-ahead forecast error coveriance matrix (same for VAR and LP IRF estimates). We report results for the estimation sample ranging from 1984:Q1-2006:Q4. We also consider a longer sample, ranging from 1984:Q1 to 2019:Q4. The results are fairly similar and relegated to the Online Appendix.

Empirical Results for IRFC Selection. Figure 4 provides information about the selected degree of shrinkage. The left column contains results for LP IRF estimation and the right column for VAR IRF Estimation. In the top panels we fix the lag length p at the values $\{1, 2, 4, 6\}$. For each p and each of the 200 samples we compute $\hat{\lambda}$. The dots represent $(p, \hat{\lambda}(p))$ and their diameters are proportional to the frequency with which this shrinkage selection occurs among the estimation samples conditional on the four choices of p. In the bottom panels, we set for each of the 200 samples the lag length to the selected value $\hat{p}$, providing information about the pairs $(\hat{p}, \hat{\lambda}(\hat{p}))$. As before, the diameter of the dots is proportional to the frequency.[6]

[6]We previously remarked that for $p \geq p_* + h$ for $\lambda = 0$ LP and VAR IRF estimation risk are identical. In Figure 4 $h = 6$. Thus, the condition would only be satisfied in samples that could be represented by an

Figure 4: Distribution of IRFC Selected Hyperparameter, $h = 6$.

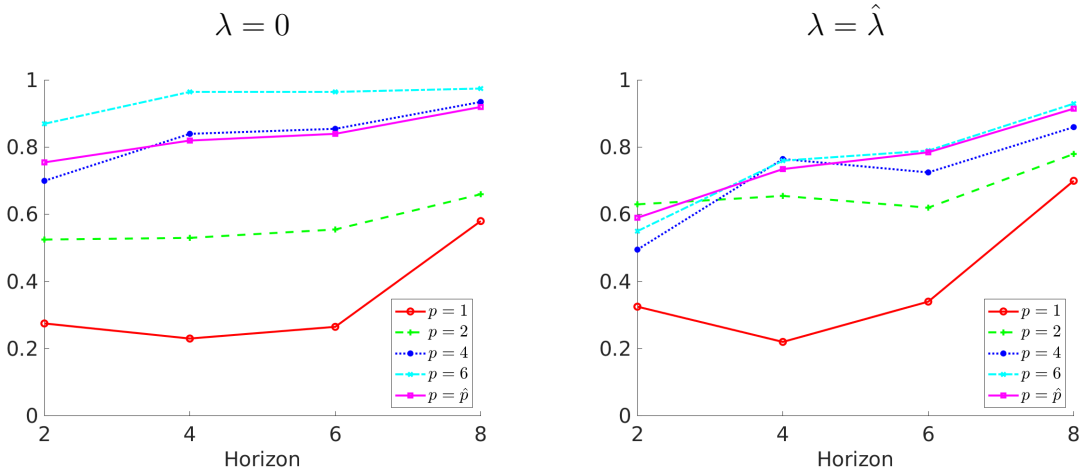


Notes: Grid values of IRFC-selected shrinkage hyperparameters for different fixed lag orders. The diameter of the dots is proportional to the frequency across samples. Fixed lag length refers to $(\hat{\lambda}(p), p)$ and each p-row represents 200 samples. Selected lag length is $(\hat{\lambda}(\hat{p}), \hat{p})$ and the number of samples across the four $\hat{p}$ rows adds up to 200. Estimation sample 1984:Q1-2006:Q4.

For fixed lag lengths three observations stand out. First, for the LP IRF estimator, the larger p, the smaller $\hat{\lambda}(p)$. This pattern is broadly consistent with the $\alpha = 2$ panels for the LP IRF estimation in Figure 2. In the numerical illustration of the asymptotic risk, the shrinkage for $p = 1$ under that specific DGP is substantially larger than for $p > 1$. The key mechanism is that unlike in the case of forecasting, the variance of the LP estimator does not increase with p once $p > p_*$. Thus, there is no need for more shrinkage as the number of lags are increased. Second, VAR shrinkage is generally stronger than LP shrinkage. This is also qualitatively consistent with the patterns in Figure 2, where for $\alpha = 0$ the optimal λ

asymptotic lag length of $p_* = 0$.

Figure 5: IRFC Selection of LP versus VAR IRF Estimate



Notes: Fraction of times the IRFC selects the LP IRF estimator under different lag lengths and across different horizons. Estimation sample 1984:Q1-2006:Q4.

for VAR IRF estimation tends to be slightly larger than for LP IRF estimation. Third, for the VAR IRF estimation the distribution of $\hat{\lambda}(p)$ does not vary as strongly with p as for the LP estimation, suggesting that the bias-variance trade-off in the various samples remains relevant regardless of p.

The bottom row of Figure 4 replaces the fixed lag lengths $p \in \{1, 2, 4, 6\}$ by the selected lag length $\hat{p}$. First, the most frequently selected lag length across the 200 samples is $\hat{p} = 6$. This is the case for LP and VAR IRF estimation. Second, in the vast majority of samples the selected LP shrinkage $\hat{\lambda}(\hat{p})$ is equal to zero. In case of the VAR IRF estimation, in many samples $\hat{\lambda}(\hat{p})$ is between 0.1 and 1.0. As we have seen previously, for LP there is no cost for increasing $\hat{p}$ beyond $p_* + 1$. For the VAR there is a potential benefit of being able to reduce the effect of misspecification by increasing the lag length beyond p_*. However, additional lags are costly in terms of variance and there is a benefit to applying more shrinkage in samples in which $\hat{p}$ is large.

Figure 5 shows the fraction of times the IRFC selects the LP IRF estimator as a function of the horizon h. Each line corresponds to a different lag length p. We also show results for the IRFC-selected lag length $\hat{p}$. The left panel corresponds to $\lambda = 0$, whereas the right panel uses the IRFC-selected $\hat{\lambda}$. Conditional on $p = 1$ the fraction of samples for which LP is selected ranges from 25% to 60%. It is largest for $h = 8$ and attains the lower bound of the range at $h \leq 4$. For $p = 2$ the fraction of LP selection is between 50% and 60% for

$\lambda = 0$ and slightly higher, between 60% and 75%, for the optimal $\hat{\lambda}(p)$. For $p \in \{4, 6\}$ the fraction of samples for which LP is selected is between 70% and 95% at $\lambda = 0$. Here LP may perform relatively well because its ability to achieve correct centering outweighs increases of the variance component of the MSE criterion. If λ is chosen optimally, then for $p \in \{4, 6\}$ the VAR IRF estimator becomes more attractive again and the LP estimate gets selected less often. The $p = \hat{p}$ line is very similar to the case of $p = 6$, which is to be expected from Figure 4. We also computed the norm difference between the VAR and LP IRF estimates (a figure is provided in the Online Appendix). It is generally increasing in the horizon h and the number of lags p.

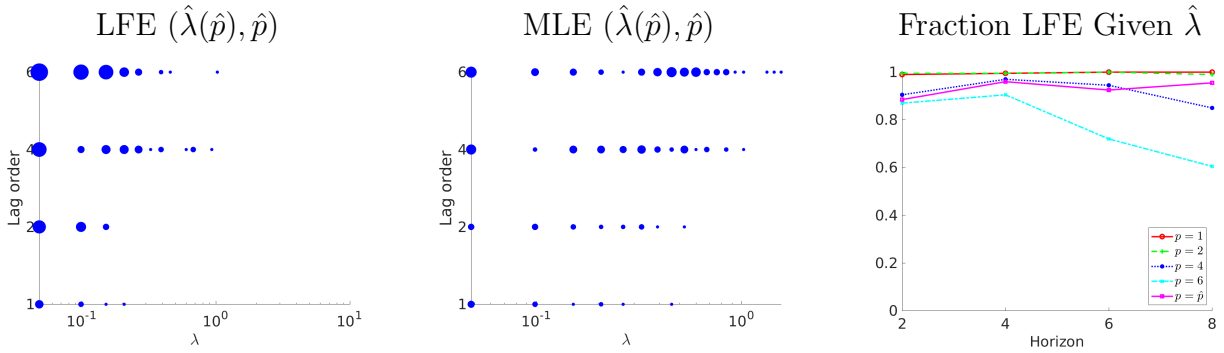
The results in Figure 5 are related to the large-scale simulation experiment by LPW. Our MSE criterion corresponds to equal weights on bias and variance in the LPW paper. While the results in the two papers are not directly comparable, we provide some discussion. The left panel in our Figure 5 could be compared to Figure 7 in LPW, except that they apply a bias correction to the VAR and LP estimators. They find the LP estimator dominates the VAR estimator in fewer than 20% of the samples. Our IRFC-based assessment of LP estimation is more favorable. Depending on the lag length p and the horizon h we select the LP IRF estimator for 20% to 95% of the samples.

In their Figure 6 LPW document that under the MSE weighting and for most of the horizons the BVAR method performs best across their samples. Their BVAR estimator is similar to our VAR estimator with $\hat{\lambda}$.[7] According the right panel in our Figure 5, except for $p = 1$ the LP estimator is selected over the VAR estimator. This might have to do with the fact that we allow for shrinkage of the LP estimator, which according to Figure 2 can improve its performance. The bottom line is that, from an MSE perspective, whether VAR or LP IRF estimation is preferable is sample dependent. There is no clear winner. These findings discredit the widespread idea that LPs are *always* preferred under misspecification. Our IRFC criterion is the first method in the literature to provide a way for empirical researchers to make a data-driven choice between estimation approaches.

PC Selection. We emphasized throughout the paper that model determination under misspecification should be based on a measure of the relevant prediction or estimation risk. We now examine how different a model selection based on multi-step forecast risk is from a choice based on the IRF estimation risk in our two hundred samples. Suppose that a researcher uses PC instead of IRFC to determine (ι, λ, p). The effect on lag length choice

[7] It is similar but not the same. LPW use an MDD-based hyperparameter determination (averaging rather than selection), whereas we use IRFC, which is preferable under misspecification as shown in Table 2.

Figure 6: PC Selection



Notes: Left and center panel: grid values of PC-selected shrinkage hyperparameters $\hat{\lambda}(\hat{p})$ for $h = 6$. The diameter of the dots is proportional to the frequency of the $(\hat{\lambda}(\hat{p}), \hat{p})$ frequency. The number of samples across the four $\hat{p}$ rows add up to 200. Right panel: Fraction of times the PC selects the LP IRF estimator under different lag lengths and across different horizons. Estimation sample 1984:Q1-2006:Q4.

and estimator selection is displayed in Figure 6. The PC results can be compared to the IRFC results shown in the bottom row of Figure 4 and in Figure 5 above. The key differences between IRFC and PC selection are: (i) targeting multi-step prediction risk leads to more shrinkage for MLE and LFE, (ii) the lag length selected by PC tends to be smaller than the IRFC lag length, and (iii) at the horizon $h = 6$ PC selects the LFE/LP more often – in fact in more than 90% of the samples, except for $p = 6$ – than IRFC. This is consistent with the distinct features of multi-step forecasting and IRF estimation discussed previously. The h-step-ahead forecast risk is affected by the estimation uncertainty for *all* coefficients, whereas the IRF estimation risk is only affected by the sampling variability of the coefficient estimates for the first lag. Thus, it is desirable to reduce estimation variance more strongly in the multi-step forecasting setting, by increasing the shrinkage and/or reducing the lag length and thereby number of coefficients that need to be estimated. Thus, an important recommendation for practitioners is to use an IRF estimation risk criterion and not a criterion that measures forecast performance, when choosing the IRF estimator.

9 Conclusion

Multi-step forecasting and IRF estimation with VARs is challenging because the number of parameters may be large relative to the available data, and the VAR may suffer from small but consequential dynamic misspecification. The first issue can be addressed by combining

likelihood information with prior information and selecting the weight in a data-driven manner. The second challenge has led econometricians to replace one-step-ahead regressions by multi-step regressions. In this paper, we propose two criteria, PC and IRFC, that can be used to jointly select the number of lags, the hyperparameters that determine the relative weight on likelihood and prior, and the type of estimator. Based on a quadratic loss function, the criteria trade off bias and variance to minimize MSE. Our simulations and empirical evidence show that the trade-offs are non-trivial. Rather than defaulting to one particular estimation strategy, we recommend practitioners to adopt a data-based selection approach based on the proposed criteria. Our empirical analysis shows that the choice between VAR or LP IRF point estimates should be sample dependent. Throughout this paper we focused on selection, but the objective functions derived in this paper could also be used to determine averaging weights.

References

Baillie, R. T. (1979): "Asymptotic Prediction Mean Squared Error for Vector Autoregressive Models," *Biometrika*, 66(3), 675–678.

Bhansali, R. J. (1996): "Asymptotically Efficient Autoregressive Model Selection for Multistep Prediction," *Annals of the Institute for Statistical Mathematics*, 48(3), 577–602.

——— (1997): "Direct Autoregressive Predictors for Multistep Prediction: Order Selection and Performance Relative to the Plug In Predictors," *Statistica Sinica*, 7, 425–449.

Billingsley, P. (1968): *Probability and Measure*. John Wiley & Sons, New York.

Carriero, A., T. E. Clark, and M. Marcellino (2015): "Bayesian VARs: Specification Choices and Forecast Accuracy," *Journal of Applied Econometrics*, 30(1), 46–73.

Cheng, X., S. C. Ho, and F. Schorfheide (2024): "Optimal Estimation of Two-Way Effects under Limited Mobility," *Manuscript, University of Pennsylvania*.

Clements, M. P., and D. F. Hendry (1998): *Forecasting Economic Time Series*. Cambridge University Press.

Del Negro, M., and F. Schorfheide (2004): "Priors from General Equilibrium Models for VARs," *International Economic Review*, 45(2), 643 – 673.

Doan, T., R. Litterman, and C. Sims (1984): "Forecasting and Conditional Projections Using Realistic Prior Distributions," *Econometric Reviews*, 3(1), 1–100.

Findley, D. F. (1983): "On the Use of Multiple Models for Multi-Period Forecasting," *American Statistical Association: Proceedings of Business and Economic Statistics*, pp. 528–531.

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

Hamilton, J. D. (2018): "Why You Should Never Use the Hodrick-Prescott Filter," *Review of Economics and Statistics*, 100(5), 831–843.

Hansen, B. E. (2016): "Stein Combination Shrinkage for Vector Autoregressions," *Manuscript, University of Wisconsin-Madison.*

Ing, C.-K. (2003): "Multistep Prediction in Autoregressive Processes," *Econometric Theory*, 19(2), 254–279.

Ing, C.-K., and C.-Z. Wei (2003): "On Same-Realization Prediction in an Infinite-Order Autoregressive Process," *Journal of Multivariate Analysis*, 85, 130–155.

Jordà, Ò. (2005): "Estimation and Inference of Impulse Responses by Local Projections," *American Economic Review*, 95(1), 161–182.

Kwon, S. (2023): "Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models," *Manuscript, Brown University.*

Lewis, R., and G. C. Reinsel (1985): "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, 16, 393–411.

Lewis, R. A., and G. C. Reinsel (1988): "Prediction Error of Multivariate Time Series With Mis-specified Models," *Journal of Time Series Analysis*, 9(1), 43–57.

Li, D., M. Plagborg-Møller, and C. Wolf (2022): "Local Projections vs. VARs: Lessons From Thousands of DGPs," *NBER Working Paper*, 30207.

Litterman, R. B. (1986): "Forecasting with Bayesian Vector Autoregressions: Five Years of Experience," *Journal of Business & Economic Statistics*, 4(1), 25–38.

Lohmeyer, J., F. Palm, H. Reuvers, and J.-P. Urbain (2018): "Focused Information Criterion for Locally Misspecified Vector Autoregressive Models," *Econometric Reviews*, 38(7), 763–792.

Ludwig, J. (2024): "Local Projections are VAR Predictions of Different Order," *Manuscript, Texax Tech University.*

Marcellino, M., J. H. Stock, and M. W. Watson (2006): "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics*, 135, 499–526.

McCracken, M. W., and S. Ng (2020): "FRED-QD: A Quarterly Database for Macroeconomic Research," *FRB St. Louis Working Paper*, 005.

Miranda-Agrippino, S., and G. Ricco (2021): "Bayesian Local Projections," *Warwick Economics Research Papers*, 1348.

Montiel Olea, J., and M. Plagborg-Møller (2021): "Local Projection Inference is Easier Than You Think," *Econometrica*, 89(4), 1789–1823.

Montiel Olea, J., M. Plagborg-Møller, E. Qian, and C. Wolf (2024): "Double Robustness of Local Projections and Some Unpleasant VARithmetic," *Manuscript, Princeton University*.

Plagborg-Møller, M., and C. K. Wolf (2021): "Local Projections and VARs Estimate the Same Impulse Responses," *Econometrica*, 89(2), 955–980.

Reinsel, G. C. (1980): "Asymptotic Properties of Prediction Errors for the Multivariate Autoregressive Model Using Estimated Parameters," *Journal of the Royal Statistical Society B*, 42(3), 328–333.

Robbins, H. (1955): "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 157–164. University of California Press, Berkeley and Los Angeles.

Schorfheide, F. (2005): "VAR Forecasting Under Misspecification," *Journal of Econometrics*, 128(1), 99–136.

Shibata, R. (1980): "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *Annals of Statistics*, 8(1), 147–164.

Speed, T., and B. Yu (1993): "Model Selection and Prediction: Normal Regression," *Annals of the Institute of Statistical Mathematics*, 45(1), 35–54.

Stein, C. (1981): "Estimation of the Mean of a Multivariate Normal Distribution," *Annals of Statistics*, 9(6), 1135–1151.

Todd, R. (1984): "Improving Economic Forecasting with Bayesian Vector Autoregressions," *Federal Reserve Bank of Minneapolis Quarterly Review*, 8(4), 18–29.

Weiss, A. A. (1991): "Multi-step Estimation and Forecasting in Dynamic Models," *Journal of Econometrics*, 48, 135–149.

Online Appendix: Misspecification-Robust Shrinkage and Selection for VAR Forecasts and IRFs

Oriol González-Casasús and Frank Schorfheide

This Appendix consists of the following sections:

A. Proofs and Derivations

B. Further Details on the Monte Carlo Simulations

C. Further Details on the Empirical Analysis

A Proofs and Derivations

A.1 Proofs for Section 2

Proof of Theorem 1. The formulas for the bias terms are given by

$$\begin{aligned}
\delta(lfe,\lambda) &= \lambda\underline{\psi}\underline{P}_{\Psi}(\lambda\underline{P}_{\Psi}+\Gamma_{yy,0})^{-1}\\
\delta(mle,\lambda) &= \lambda\sum_{j=0}^{h-1}F^{j}\underline{\phi}\underline{P}_{\Phi}(\lambda\underline{P}_{\Phi}+\Gamma_{yy,0})^{-1}F^{h-1-j}\\
\mu(lfe,\lambda) &= \sum_{j=0}^{h-1}F^{j}\Gamma_{zy,h-j}(\lambda\underline{P}_{\Psi}+\Gamma_{yy,0})^{-1}\\
\mu(mle,\lambda) &= \sum_{j=0}^{h-1}F^{j}\Gamma_{zy,1}(\lambda\underline{P}_{\Phi}+\Gamma_{yy,0})^{-1}F^{h-1-j}.
\end{aligned}$$

The asymptotic covariance matrix between (ι,λ) and (ι',λ') predictors takes the form

$$\mathbf{V}=\begin{bmatrix}
V(mle,\lambda) & & & \\
Cov(lfe,\lambda;mle,\lambda) & V(lfe,\lambda) & & \\
Cov(mle,\lambda';mle,\lambda) & Cov(mle,\lambda';lfe,\lambda) & V(mle,\lambda') & \\
Cov(lfe,\lambda';mle,\lambda) & Cov(lfe,\lambda';lfe,\lambda) & Cov(lfe,\lambda';mle,\lambda') & V(lfe,\lambda')
\end{bmatrix}$$

with the elements defined as

$$\begin{aligned}
Cov(lfe,\lambda;lfe,\lambda') &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^{i}\Sigma_{\epsilon\epsilon}F^{j'})\otimes((\lambda\underline{P}'_{\Psi}+\Gamma_{yy,0})^{-1}\Gamma_{yy,j-i}(\lambda'\underline{P}_{\Psi}+\Gamma_{yy,0})^{-1})\\
Cov(mle,\lambda;mle,\lambda') &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^{i}\Sigma_{\epsilon\epsilon}F^{j'})\otimes(F^{h-1-i'}(\lambda\underline{P}'_{\Phi}+\Gamma_{yy,0})^{-1}\Gamma_{yy,0}(\lambda'\underline{P}_{\Phi}+\Gamma_{yy,0})^{-1}F^{h-1-j})\\
Cov(mle,\lambda;lfe,\lambda') &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^{i}\Sigma_{\epsilon\epsilon}F^{j'})\otimes(F^{h-1-i'}(\lambda\underline{P}'_{\Phi}+\Gamma_{yy,0})^{-1}\Gamma_{yy,h-1-j}(\lambda'\underline{P}_{\Psi}+\Gamma_{yy,0})^{-1}),
\end{aligned}$$

and trivially $V(\iota,\lambda)=Cov(\iota,\lambda;\iota,\lambda)$. Because $\Gamma_{yy,h-1-j}=F^{h-1-j}\Gamma_{yy,0}$, we obtain that

$$Cov(mle,0;lfe,0)=\sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^{i}\Sigma_{\epsilon\epsilon}F^{j'})\otimes(F^{h-1-i'}\Gamma_{yy,0}^{-1}F^{h-1-j})=V(mle,0).$$

Analysis of LFE. First, note that

$$\bar{\Psi}_{T}(lfe,\lambda)-F^{h}=(\underline{\Psi}_{T}-F^{h})\tilde{\lambda}\underline{P}_{\Psi}\bar{P}_{\Psi}^{-1}+(\hat{\Psi}_{T}(lfe)-F^{h})S_{T,hh}\bar{P}_{\Psi}^{-1}.$$

Moreover, the LFE can be written as

$$\hat{\Psi}_T(lfe) = F^h + \alpha T^{-1/2}\left(\sum_{j=0}^{h-1}\sum_{t=1}^{T} F^j z_{t-j} y'_{t-h}\right) S_{T,hh}^{-1} + \left(\sum_{j=0}^{h-1}\sum_{t=1}^{T} F^j \epsilon_{t-j} y'_{t-h}\right) S_{T,hh}^{-1}.$$

Therefore,

$$\begin{aligned}\bar{\Psi}_T(lfe,\lambda) - F^h &= (\underline{\Psi}_T - F^h)\tilde{\lambda}\underline{P}_\Psi \bar{P}_\Psi^{-1} \\ &\quad + \alpha T^{-1/2}\left(\sum_{j=0}^{h-1}\sum_{t=1}^{T} F^j z_{t-j} y'_{t-h}\right)\bar{P}_\Psi^{-1} \\ &\quad + \left(\sum_{j=0}^{h-1}\sum_{t=1}^{T} F^j \epsilon_{t-j} y'_{t-h}\right)\bar{P}_\Psi^{-1}.\end{aligned}$$

By the same steps as in Schorfheide (2005) and equations (14) and (15),

$$T^{1/2}\left(\bar{\Psi}_T(lfe,\lambda) - F^h\right) = \delta(lfe,\lambda) + \alpha\mu(lfe) + \zeta_T(lfe,\lambda),$$

where

$$\begin{aligned}\delta(lfe,\lambda) &= \lambda\underline{\psi}\underline{P}_\Psi(\lambda\underline{P}_\Psi + \Gamma_{yy,0})^{-1} \\ \mu(lfe,\lambda) &= \sum_{j=0}^{h-1} F^j\Gamma_{zy,h-j}(\lambda\underline{P}_\Psi + \Gamma_{yy,0})^{-1} \\ \zeta_T(lfe,\lambda) &\Longrightarrow N(0, V(lfe,\lambda)) \\ V(lfe,\lambda) &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i\Sigma_{\epsilon\epsilon}F^{j'}) \otimes ((\lambda\underline{P}_\Psi + \Gamma_{yy,0})^{-1}\Gamma_{yy,j-i}(\lambda\underline{P}_\Psi + \Gamma_{yy,0})^{-1}).\end{aligned}$$

Analysis of MLE. By a first order Taylor expansion,

$$\Phi^h - F^h = \sum_{j=0}^{h-1} F^j(\Phi - F)F^{h-1-j} + \mathcal{R}(\Phi - F).$$

Note that

$$\bar{\Phi}_T(mle,\lambda) - F = (\underline{\Phi}_T - F)\tilde{\lambda}\underline{P}_\Phi\bar{P}_\Phi^{-1} + (\hat{\Phi}_T(mle) - F)S_{T,11}\bar{P}_\Phi^{-1},$$

so it follows that

$$\begin{aligned}\bar{\Psi}(mle,\lambda) - F^h &= \tilde{\lambda}\sum_{j=0}^{h-1} F^j(\underline{\Phi}_T - F)\underline{P}_\Phi\bar{P}_\Phi^{-1}F^{h-1-j} \\ &\quad + \sum_{j=0}^{h-1} F^j(\hat{\Phi}_T(mle) - F)S_{T,11}\bar{P}_\Phi^{-1}F^{h-1-j} \\ &\quad + \mathcal{R}(\bar{\Phi}_T(mle,\lambda) - F).\end{aligned}$$

By Schorfheide (2005) and equations (14) and (15),

$$T^{1/2}\left(\bar{\Psi}(mle,\lambda) - F^h\right) = \delta(mle,\lambda) + \alpha\mu(mle,\lambda) + \zeta_T(mle,\lambda)$$

where

$$\begin{aligned}
\delta(mle,\lambda) &= \lambda\sum_{j=0}^{h-1} F^j \underline{\phi}\underline{P}_\Phi(\lambda\underline{P}_\Phi + \Gamma_{yy,0})_\Phi^{-1} F^{h-1-j} \\
\mu(mle,\lambda) &= \sum_{j=0}^{h-1} F^j \Gamma_{zy,1}(\lambda\underline{P}_\Phi + \Gamma_{yy,0})^{-1} F^{h-1-j} \\
\zeta_T(mle,\lambda) &\Longrightarrow N(0, V(mle,\lambda)) \\
V(mle,\lambda) &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i \Sigma_{\epsilon\epsilon} F^{j'}) \otimes (F^{h-1-i'}(\lambda\underline{P}'_\Phi + \Gamma_{yy,0})^{-1}\Gamma_{yy,0}(\lambda\underline{P}_\Phi + \Gamma_{yy,0})^{-1} F^{h-1-j}).
\end{aligned}$$

The covariance follows from the same arguments as in Schorfheide (2005). ■

Proof of Theorem 2. The difference between the conditional expectation of y_{T+h} (omitting the tilde) and the predictor $\hat{y}_{T+h}(\iota,\lambda)$ is given by

$$\begin{aligned}
T^{1/2}(\mathbb{E}_T[y_{T+h}] - \hat{y}_{T+h}(\iota,\lambda)) &= \alpha\left(\sum_{j=0}^{h-1} F^j \mathbb{E}_T[z_{T+h-j}] - \mu(pov)y_T\right) \\
&\quad + \alpha[\mu(pov) - \mu(\iota,\lambda)]y_T - \zeta_T(\iota,\lambda)y_T \\
&\quad - \delta(\iota,\lambda)y_T.
\end{aligned}$$

The normalized prediction risk can then be expressed as follows:

$$
\begin{aligned}
& T\mathbb{E}\left[tr\{W(\mathbb{E}_T[y_{T+h}]-\hat{y}_{T+h}(\iota,\lambda))(\mathbb{E}_T[M'Y_{T+h}]-\hat{y}_{T+h}(\iota,\lambda))'\}\right] \qquad \text{(A.1)}\\
=_{(1)}\quad & \alpha^2 tr\left\{W(\mu(pov)-\mu(\iota,\lambda))\Gamma_{YY,0}(\mu(pov)-\mu(\iota,\lambda))'\right\}\\
_{(2)}\quad & +tr\left\{W\mathbb{E}\left[\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right]\right\}\\
_{(3)}\quad & +\alpha^2 tr\left\{W\mathbb{E}\left[\left(\sum_{j=0}^{h-1}F^j\mathbb{E}_T[z_{T+h-j}]-\mu(pov)y_T\right)\right.\right.\\
& \left.\left.\times\left(\sum_{j=0}^{h-1}F^j\mathbb{E}_T[z_{T+h-j}]-\mu(pov)y_T\right)'\right]\right\}\\
_{(4)}\quad & +tr\left\{W\delta(\iota,\lambda)\Gamma_{yy,0}\delta(\iota,\lambda)'\right\}\\
_{(5)}\quad & -2\alpha tr\left\{W\mathbb{E}[\zeta_T(\iota,\lambda)]\mathbb{E}\left[y_T\left(\sum_{j=0}^{h-1}F^j\mathbb{E}_T[z_{T+h-j}]-\mu(pov)y_T\right)'\right]\right\}\\
_{(6)}\quad & +2\alpha^2 tr\left\{W\mathbb{E}\left[\sum_{j=0}^{h-1}F^j\mathbb{E}_T[z_{T+h-j}]y_T'-\mu(pov)y_Ty_T'\right](\mu(pov)-\mu(\iota,\lambda))'\right\}\\
_{(7)}\quad & -2\alpha tr\left\{W\mathbb{E}[\zeta_T(\iota,\lambda)]\Gamma_{yy,0}(\mu(pov)-\mu(\iota,\lambda))'\right\}\\
_{(8)}\quad & -2\alpha tr\left\{W\delta(\iota,\lambda)\mathbb{E}\left[y_T\left(\sum_{j=0}^{h-1}F^j\mathbb{E}_T[z_{T+h-j}]-\mu(pov)y_T\right)'\right]\right\}\\
_{(9)}\quad & -2\alpha tr\left\{W(\mu(pov)-\mu(\iota,\lambda))\Gamma_{yy,0}\delta(\iota,\lambda)'\right\}\\
_{(10)}\quad & +2tr\left\{W\mathbb{E}[\zeta_T(\iota,\lambda)]\Gamma_{yy,0}\delta(\iota,\lambda)'\right\}.
\end{aligned}
$$

Since

$$tr[WABA'] = vecr(A)'(W\otimes B)vecr(A)$$

and $tr[AB]=tr[BA]$ we can rewrite term (2) in (A.1) as

$$tr\left\{W\mathbb{E}\left[\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right]\right\} = tr\left\{(W\otimes\Gamma_{yy,0})\mathbb{E}\left[\zeta_T(\iota,\lambda)\zeta_T(\iota,\lambda)'\right]\right\}$$

with the understanding that on the right-hand side of the equation $\zeta_T(\iota,\lambda)$ is vectorized. Under Assumption 1 in S2005, the sequence $\|\zeta_T(\iota,\lambda)\|^2$ is uniformly integrable. Hence, we can deduce that (see Theorem 3.5 of Billingsley (1968))

$$tr\left\{(W\otimes\Gamma_{yy,0})\mathbb{E}\left[\zeta_T(\iota,\lambda)\zeta_T(\iota,\lambda)'\right]\right\} \longrightarrow tr\left\{(W\otimes\Gamma_{yy,0})V(\iota,\lambda)\right\}.$$

Moreover, uniform integrability of $\|\zeta_T(\iota,p)\|^2$ implies that $\mathbb{E}[\zeta_T(\iota,\lambda)] = o(1)$, and so terms (5), (7), and (10) in (A.1) are $o(1)$. Since

$$\mathbb{E}\left[\sum_{j=0}^{h-1} F^j \mathbb{E}_T[z_{T+h-j}] y_T'\right] = \sum_{j=0}^{h-1} F^j \Gamma_{zy,h-j} = \mu(pov)\Gamma_{yy,0}$$

terms (6) and (8) in (A.1) are $o(1)$, too. The above simplifications allow us to rewrite the normalized prediction risk as

$$\begin{aligned}
&T\mathbb{E}\Big[tr\{W(\mathbb{E}_T[y_{T+h}] - \hat{y}_{T+h}(\iota,\lambda))(\mathbb{E}_T[y_{T+h}] - \hat{y}_{T+h}(\iota,\lambda))'\}\Big] \\
&=_{(1)} \quad \alpha^2 tr\Big\{W(\mu(pov) - \mu(\iota,\lambda))\Gamma_{yy,0}(\mu(pov) - \mu(\iota,\lambda))'\Big\} \\
&\quad_{(2)} \quad + tr\Big\{(W \otimes \Gamma_{yy,0})V(\iota,\lambda)\Big\} \\
&\quad_{(3)} \quad + \alpha^2 tr\left\{W\mathbb{E}\left[\left(\sum_{j=0}^{h-1} F^j \mathbb{E}_T[z_{T+h-j}] - \mu(pov)y_T\right)\left(\sum_{j=0}^{h-1} F^j \mathbb{E}_T[z_{T+h-j}] - \mu(pov)y_T\right)'\right]\right\} \\
&\quad_{(4)} \quad + tr\left\{W\delta(\iota,\lambda)\Gamma_{yy,0}\delta(\iota,\lambda)'\right\} \\
&\quad_{(9)} \quad - 2\alpha tr\left\{W\delta(\iota,\lambda)\Gamma_{yy,0}\left(\mu(pov) - \mu(\iota,\lambda)\right)'\right\} \\
&\qquad + o(1).
\end{aligned}$$

Hence, the desired result follows. ■

A.2 Proofs for Section 3

Proof of Theorem 3. Using the asymptotic representation of $\bar{\Psi}(\iota,\lambda)$ given in Theorem 1, the in-sample loss can be decomposed as follows

$$\begin{aligned}
&T \cdot MSE(\iota,\lambda) \\
&= \sum_{t=1}^{T}(y_t - F^h y_{t-h})(y_t - F^h y_{t-h})' \\
&= -T^{-1/2}\sum_{t=1}^{T}(y_t y_{t-h}' - F^h y_{t-h} y_{t-h}')(\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda) + o_p(1))' \\
&\quad -T^{-1/2}\sum_{t=1}^{T}(\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda) + o_p(1))(y_t y_{t-h}' - F^h y_{t-h} y_{t-h}')' \\
&\quad +(\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda) + o_p(1))\left(T^{-1}\sum_{t=1}^{T} y_{t-h} y_{t-h}'\right) \\
&\quad \times(\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda) + o_p(1))'.
\end{aligned}$$

From the definition of $\zeta_T(lfe,\lambda)$, it follows that

$$\begin{aligned} & T^{-1/2}\sum_{t=1}^{T}(y_t y_{t-h}' - F^h y_{t-h} y_{t-h}') \\ &= \alpha\sum_{j=0}^{h-1}\left(T^{-1}\sum_{t=1}^{T}F^j z_{t-j}y_{t-h}'\right) + \sum_{j=0}^{h-1}\left(F^j T^{-1/2}\sum_{t=1}^{T}\epsilon_{t-j}y_{t-h}'\right) \\ &= \left[\zeta_T(lfe,\lambda) + \alpha\mu(lfe,\lambda) + o_p(1)\right](T\bar{P}_{\Psi}^{-1})^{-1} \end{aligned}$$

for any $\lambda \geq 0$. Without loss of generality, take $\lambda = 0$, whence

$$T^{-1/2}\sum_{t=1}^{T}(y_t y_{t-h}' - F^h y_{t-h} y_{t-h}') = \left[\zeta_T(lfe,0) + \alpha\mu(pov)\right]T^{-1}S_{T,hh}.$$

Therefore,

$$\begin{aligned} & T\cdot tr\left\{W\cdot MSE(\iota,\lambda)\right\} \\ &= tr\left\{W\sum_{t=1}^{T}(y_t - F^h y_{t-h})(y_t - F^h y_{t-h})'\right\} \\ &\quad -2tr\left\{W\left[\zeta_T(lfe,0) + \alpha\mu(pov)\right](T^{-1}S_{T,hh})\left[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda) + o_p(1)\right]'\right\} \\ &\quad +tr\Bigg\{W\left[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda) + o_p(1)\right]\left(T^{-1}S_{T,hh}\right) \\ &\quad \times\left[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda) + o_p(1)\right]'\Bigg\}. \end{aligned}$$

Observe that $T^{-1}S_{T,hh} = \Gamma_{yy,0} + o_p(1)$, hence

$$\begin{aligned} & T\Big(tr\left\{W\cdot MSE(\iota,\lambda)\right\} - tr\left\{W\cdot MSE(lfe,0)\right\}\Big) \\ &= -2tr\left\{W\left[\zeta_T(lfe,0) + \alpha\mu(pov)\right]\Gamma_{yy,0}\left[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda)\right]'\right\} \\ &\quad +tr\left\{W\left[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda)\right]\Gamma_{yy,0}\left[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda)\right]'\right\} \\ &\quad +tr\left\{W\left[\zeta_T(lfe,0) + \alpha\mu(pov)\right]\Gamma_{yy,0}\left[\alpha\mu(pov) + \zeta_T(lfe,0)\right]'\right\} + o_p(1). \end{aligned}$$

Statement (i) now follows from Theorem 1, the Continuous Mapping Theorem and a straightforward rearrangement of terms.

For statement (ii), from part (i) and uniform integrability of the in-sample loss differential it is easy to see that

$$\begin{aligned} & \mathbb{E}\left[\Delta_{\mathcal{R},T}(\iota,\lambda)\right] \\ &\longrightarrow \mathbb{E}\left[\left\|\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta(\iota,\lambda)\right\|_{W\otimes\Gamma_{yy,0}}^2\right] + \mathbb{E}\left[\left\|\alpha\mu(pov) + \zeta(lfe,0)\right\|_{W\otimes\Gamma_{yy,0}}^2\right] \\ &\quad -2\mathbb{E}\left[tr\left\{W\left[\alpha\mu(pov) + \zeta(lfe,0)\right]\Gamma_{yy,0}\left[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta(\iota,\lambda)\right]'\right\}\right]. \end{aligned}$$

Working out the expected values according to Theorem 1 yields

$$\begin{aligned}\mathbb{E}\left[\Delta_{\mathcal{R},T}(\iota,\lambda)\right] \longrightarrow\ & \left\|\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right\|^2_{W\otimes\Gamma_{yy,0}} + tr\left\{(W\otimes\Gamma_{yy,0})V(\iota,\lambda)\right\} \\ & +\alpha^2\left\|\mu(pov)\right\|^2_{W\otimes\Gamma_{yy,0}} + tr\left\{(W\otimes\Gamma_{yy,0})V(lfe,0)\right\} \\ & -2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\} \\ & -2tr\left\{(W\otimes\Gamma_{yy,0})Cov(lfe,0;\iota,\lambda)\right\}.\end{aligned}$$

Using the definitions of $\bar{\mathcal{R}}_B(\iota,\lambda)$ and $\bar{\mathcal{R}}_V(\iota,\lambda)$ in Theorem 2 and recognizing that $\bar{\mathcal{R}}_B(lfe,0) = 0$ we can write the r.h.s. as

$$\begin{aligned}\text{r.h.s} &= \left\|\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right\|^2_{W\otimes\Gamma_{yy,0}} + \alpha^2\left\|\mu(pov)\right\|^2_{W\otimes\Gamma_{yy,0}} \\ &\quad -2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\} \\ &\quad +\bar{\mathcal{R}}_V(\iota,\lambda)+\bar{\mathcal{R}}_V(lfe,0) - 2tr\left\{(W\otimes\Gamma_{yy,0})Cov(lfe,0;\iota,\lambda)\right\} \\ &= \bar{\mathcal{R}}_B(\iota,\lambda)+\bar{\mathcal{R}}_V(\iota,\lambda) - \left(\bar{\mathcal{R}}_B(lfe,0)+\bar{\mathcal{R}}_V(lfe,0)\right) \\ &\quad +2\bar{\mathcal{R}}_V(lfe,0) - 2tr\left\{(W\otimes\Gamma_{yy,0})Cov(lfe,0;\iota,\lambda)\right\}. \quad \blacksquare\end{aligned}$$

A.3 Proofs for Section 4

Proof of Lemma 1. (i) Partition

$$\bar{S}^{-1}_{T,hh} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

such that the partitions conform with the position of zeros and ones in R_p and $R_{p\perp}$. Then:

$$\begin{aligned}
\bar{\Psi}_T(\cdot) &= \bar{S}_{T,0h}\bar{S}_{T,hh}^{-1}\left[I_{nq} - R_p(R_p'\bar{S}_{T,hh}^{-1}R_p)^{-1}R_p'\bar{S}_{T,hh}^{-1}\right] \qquad \text{(A.2)}\\
&= \bar{S}_{T,0h}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\left(\begin{bmatrix}I_{11} & 0\\ 0 & I_{22}\end{bmatrix} - \begin{bmatrix}0 & 0\\ 0 & A_{22}^{-1}\end{bmatrix}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\right)\\
&= \bar{S}_{T,0h}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\left(\begin{bmatrix}I_{11} & 0\\ 0 & I_{22}\end{bmatrix} - \begin{bmatrix}0 & 0\\ A_{22}^{-1}A_{21} & I_{22}\end{bmatrix}\right)\\
&= \bar{S}_{T,0h}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}I_{11} & 0\\ -A_{22}^{-1}A_{21} & 0\end{bmatrix}\\
&= \bar{S}_{T,0h}\begin{bmatrix}A_{11}-A_{12}A_{22}^{-1}A_{21} & 0\\ 0 & 0\end{bmatrix}\\
&= \bar{S}_{T,0h}\begin{bmatrix}\left(R_{p\perp}'\bar{S}_{T,hh}R_{p\perp}\right)^{-1} & 0\\ 0 & 0\end{bmatrix}\\
&= \begin{bmatrix}R_{p\perp}'\bar{S}_{T,0h}R_{p\perp} & R_{p\perp}'\bar{S}_{T,0h}R_p\\ R_p'\bar{S}_{T,0h}R_{p\perp} & R_p'\bar{S}_{T,0h}R_p\end{bmatrix}\begin{bmatrix}\left(R_{p\perp}'\bar{S}_{T,hh}R_{p\perp}\right)^{-1} & 0\\ 0 & 0\end{bmatrix}\\
&= \begin{bmatrix}R_{p\perp}'\bar{S}_{T,0h}R_{p\perp}\left(R_{p\perp}'\bar{S}_{T,hh}R_{p\perp}\right)^{-1} & 0\\ R_p'\bar{S}_{T,0h}R_{p\perp}\left(R_{p\perp}'\bar{S}_{T,hh}R_{p\perp}\right)^{-1} & 0\end{bmatrix}. \quad \blacksquare
\end{aligned}$$

(ii) Part (i) implies that the last $n(q-p)$ columns of the first np rows of $\bar{\Phi}_T(mle,\tilde{\lambda},p)$ are equal to zero as required. Now consider

$$\begin{aligned}
&R_{p\perp}'\bar{S}_{T,01}R_{p\perp}\left(R_{p\perp}'\bar{S}_{T,11}R_{p\perp}\right)^{-1}\\
&= \left[R_{p\perp}'S_{T,01}R_{p\perp}\left(R_{p\perp}'S_{T,11}R_{p\perp}\right)^{-1}\left(R_{p\perp}'S_{T,11}R_{p\perp}\right) + \tilde{\lambda}R_{p\perp}'\underline{\Phi}_T\underline{P}_\phi R_{p\perp}\right]\left(R_{p\perp}'\bar{S}_{T,11}R_{p\perp}\right)^{-1}
\end{aligned}$$

Notice that $R_{p\perp}'S_{T,01}R_{p\perp}\left(R_{p\perp}'S_{T,11}R_{p\perp}\right)^{-1}$ is the OLS estimator of a VAR(p) written in p-companion form. It can be expressed as

$$R_{p\perp}'S_{T,01}R_{p\perp}\left(R_{p\perp}'S_{T,11}R_{p\perp}\right)^{-1} = \begin{bmatrix}M'S_{T,01}R_{p\perp}\left(R_{p\perp}'S_{T,11}R_{p\perp}\right)^{-1}\\ \Upsilon_p\end{bmatrix}.$$

The last $n(p-1)$ rows are equal to Υ_p because $y_{t-1},\ldots,y_{t-p+1}$ lie in the space spanned by $y_{t-1},\ldots,y_{t-p}$.

For the (weighted) prior mean we obtain from (38):

$$\begin{aligned} & R_{p\perp}'\underline{\Phi}_T\underline{P}_\phi R_{p\perp} \\ &= R_{p\perp}'\begin{bmatrix} \underline{\phi}_{1,T} & \cdots & \underline{\phi}_{p-1,T} & \underline{\phi}_{p,T} & 0_{n\times n(q-p)} \\ & & \Upsilon_p & & 0_{n(p-1)\times n(q-p)} \\ & & \cdot & & \cdot \\ & & \cdot & & \cdot \end{bmatrix} \begin{bmatrix} \underbrace{\underline{P}_{\phi,11}}_{np\times np} & \underbrace{\underline{P}_{\phi,12}}_{np\times n(q-p)} \\ \underbrace{\underline{P}_{\phi,21}}_{n(q-p)\times np} & \underbrace{\underline{P}_{\phi,22}}_{n(q-p)\times n(q-p)} \end{bmatrix} R_{p\perp} \\ &= \begin{bmatrix} \underline{\phi}_{1,T} & \cdots & \underline{\phi}_{p-1,T} & \underline{\phi}_{p,T} \\ & & \Upsilon_p & \end{bmatrix} \underline{P}_{\phi,11}. \end{aligned}$$

The result follows by noting that rows $n+1$ to np of $R_{p\perp}'\bar{\Phi}_T(lfe,\tilde{\lambda},p)R_{p\perp}'$ the weighted MLE and prior mean are identical and equal to Υ_p. ■

Proof of Lemma 2. *(i) The "if" part* can be proved as follows. Because $p \geq p_*$ we can partition F as follows:

$$F = \begin{bmatrix} \underbrace{F_{11,1}}_{np\times np} & \underbrace{0}_{np\times n(q-p)} \\ \underbrace{F_{21,1}}_{n(q-p)\times np} & \underbrace{F_{22,1}}_{n(q-p)\times n(q-p)} \end{bmatrix}. \tag{A.3}$$

The "if" part of the lemma follows if F^h has the form

$$F^h = \begin{bmatrix} \underbrace{F_{11,h}}_{np\times np} & \underbrace{0}_{np\times n(q-p)} \\ \underbrace{F_{21,h}}_{n(q-p)\times np} & \underbrace{F_{22,h}}_{n(q-p)\times n(q-p)} \end{bmatrix}, \tag{A.4}$$

which is true for $h = 1$. For $h > 1$ we can use a proof by induction. Suppose that (A.4) is true for h. Then

$$F^{h+1} = FF^h = \begin{bmatrix} F_{11,1} & 0 \\ F_{21,1} & F_{22,1} \end{bmatrix}\begin{bmatrix} F_{11,h} & 0 \\ F_{21,h} & F_{22,h} \end{bmatrix} = \begin{bmatrix} F_{11,1}F_{11,h} & 0 \\ F_{21,1}F_{11,h} + F_{22,1}F_{21,h} & F_{22,1}F_{22,h} \end{bmatrix}. \tag{A.5}$$

Thus, the "if" part of the Lemma holds for any h.

(ii) The "and only if" part can be proved as follows. Because $p < p_*$ we can partition F as follows:

$$F = \begin{bmatrix} \underbrace{F_{11,1}}_{np\times np} & \underbrace{F_{12,1}}_{np\times(p_*-p)} & \underbrace{0}_{np\times n(q-p_*)} \\ \underbrace{F_{21,1}}_{n(p_*-p)\times np} & \underbrace{F_{22,1}}_{n(p_*-p)\times(p_*-p)} & \underbrace{0}_{n(p_*-p)\times n(q-p_*)} \\ \underbrace{F_{31,1}}_{n(q-p_*)\times np} & \underbrace{F_{32,1}}_{n(q-p_*)\times n(p_*-p)} & \underbrace{F_{33,1}}_{n(q-p_*)\times n(q-p_*)} \end{bmatrix}. \tag{A.6}$$

The "only if" part of the lemma follows if F^h has the form

$$F^h = \begin{bmatrix} \underbrace{F_{11,h}}_{np\times np} & \underbrace{F_{12,h}}_{np\times(p_*-p)} & \underbrace{0}_{np\times n(q-p_*)} \\ \underbrace{F_{21,h}}_{n(p_*-p)\times np} & \underbrace{F_{22,h}}_{n(p_*-p)\times(p_*-p)} & \underbrace{0}_{n(p_*-p)\times n(q-p_*)} \\ \underbrace{F_{31,h}}_{n(q-p_*)\times np} & \underbrace{F_{32,h}}_{n(q-p_*)\times n(p_*-p)} & \underbrace{F_{33,h}}_{n(q-p_*)\times n(q-p_*)} \end{bmatrix}, \tag{A.7}$$

where $F_{12,h} \neq 0$. Note that this is true for $h = 1$ because $p < p_*$. We proceed by induction. Suppose that (A.7) is true for h. Then

$$F^{h+1} = FF^h = \begin{bmatrix} F_{11,1} & F_{12,1} & 0 \\ F_{21,1} & F_{22,1} & 0 \\ F_{31,1} & F_{32,1} & F_{33,1} \end{bmatrix} \begin{bmatrix} F_{11,h} & F_{12,h} & 0 \\ F_{21,h} & F_{22,h} & 0 \\ F_{31,h} & F_{32,h} & F_{33,h} \end{bmatrix} \tag{A.8}$$

$$= \begin{bmatrix} \cdot & F_{11,1}F_{12,h} + F_{12,1}F_{22,h} & 0 \\ \cdot & \cdot & 0 \\ \cdot & \cdot & F_{33,1} + F_{33,h} \end{bmatrix}. \tag{A.9}$$

Because $F_{11,1}$ and $F_{12,h}$ are both non-zero, the $(1,2)$ element of F^{h+1} is non-zero. Thus, the "only if" part of the Lemma holds for any h. ■

A.4 Proofs and Derivations for Section 5

Proof of Theorem 4. Write

$$\mu(lfe, 0, p) = \operatorname*{plim}_{T\longrightarrow\infty} \left(\sum_{t=1}^{T} \left(\sum_{j=0}^{h-1} F^j Z_{t-j} \right) Y'_{t-h} \right) Q_{T,p}^{(h)} \tag{A.10}$$

where

$$Q_{T,p}^{(h)} = S_{T,hh}^{-1}\big[I_{nq} - R_p(R_p'S_{T,hh}^{-1}R_p)^{-1}R_p'S_{T,hh}^{-1}\big].$$

Thus, $\mu(lfe, 0, p)$ can be interpreted as the probability limit of the least squares estimate obtained by regressing $\sum_{j=0}^{h-1} F^j Z_{t-j}$ onto Y_{t-h} subject to the restriction that the coefficients on lags $p+1$ to q are equal to zero. To facilitate the subsequent, it is helpful to define the exact VAR(p_*) process

$$Y_t^* = FY_{t-1}^* + M\epsilon_t. \tag{A.11}$$

The difference between Y_t and Y_t^* is that the latter excludes the $T^{-1/2}$ misspecification term. By construction, the $T \longrightarrow \infty$ limit autocovariances of the Y_t process satisfy:

$$\Gamma_{ZY^*,h} = \Gamma_{ZY,h} \quad \text{and} \quad \Gamma_{YY^*,h} = \Gamma_{YY,h}. \tag{A.12}$$

This justifies the replacement of Y_t by Y_t^* in the subsequent analysis.

The statement of the theorem concerns $M'\mu(lfe, 0, p)M$, which corresponds to estimates of the coefficients that relate $M' \sum_{j=0}^{h-1} F^j Z_{t-j}$ to y_{t-h}^*, after controlling for $y_{t-h-1}^*, \ldots, y_{t-h-p+1}^*$. We now apply the FWL theorem to rewrite the coefficient estimates. The application of the FWL theorem involves regressing the right-hand-side variable y_{t-h}^* on $y_{t-h-1}^*, \ldots, y_{t-h-p+1}^*$ and constructing residuals, which we denote by $\tilde{y}_{t-h}^*$. To represent the OLS estimator, we do not have to transform the left-hand-side variable $M' \sum_{j=0}^{h-1} F^j Z_{t-j}$ because the underlying projection is idempotent. Thus, we define

$$\tilde{z}_t^* = \sum_{j=0}^{h-1} M' F^j Z_{t-j}$$

and note that the probability limit of the least squares estimator takes the form

$$M'\mu(lfe, 0, p)M = \operatorname*{plim}_{T\longrightarrow\infty} \left(\frac{1}{T}\sum_{t=1}^{T} \tilde{z}_t^* \tilde{y}_{t-h}^{*\prime}\right) \left(\frac{1}{T}\sum_{t=1}^{T} \tilde{y}_{t-h}^* \tilde{y}_{t-h}^{*\prime}\right)^{-1} = \Gamma_{\tilde{z}\tilde{y},h}\Gamma_{\tilde{y}\tilde{y},0}^{-1}. \tag{A.13}$$

Next, observe that $\tilde{y}_{t-h}^*$ is essentially the residuals from a VAR$(p-1)$. Using the companion form notation, we can write

$$\tilde{y}_{t-h}^* = M'\big(Y_{t-h}^* - \big(S_{T,01}^*(S_{T,11}^*)^{-1}Q_{T,p-1}^{*(1)} + O_p(T^{-1})\big)Y_{t-h-1}^*\big). \tag{A.14}$$

It is straightforward to verify using (A.11) and (A.12) that

$$\Gamma_{\tilde{y}\tilde{y},0} = \operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \tilde{y}_{t-h}^* (\tilde{y}_{t-h}^*)' = \Sigma_{\epsilon\epsilon} \quad \text{if and only if } p-1 \geq p_*, \tag{A.15}$$

and that the moments and autocovariances of $\tilde{y}_t^*$ are asymptotically equivalent to the moments and autocovariances of the error terms ϵ_t.

(i) "If" Part. Recall that for $p > p_*$ $\tilde{y}_{t-h}^* \approx \epsilon_{t-h}$. Using the Ergodic Theorem, we deduce that

$$\Gamma_{\tilde{z}\tilde{y},h} = \operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \tilde{z}_t^* \tilde{y}_{t-h}^* = \mathbb{E}\left[\sum_{j=0}^{h-1} M' F^j Z_{t-j} \epsilon_{t-h}'\right].$$

Plugging in the definition of Z_{t-j}, we obtain

$$\Gamma_{\tilde{z}\tilde{y},h} = \sum_{j=0}^{h-1}\sum_{l=1}^{\infty} M'F^j A_l M \mathbb{E}\big[\epsilon_{t-j-l}\epsilon'_{t-h}\big]. \tag{A.16}$$

Notice that the only terms with a non-zero expected value are the ones for which $j+l=h$. This leads to

$$\Gamma_{\tilde{z}\tilde{y},h} = \sum_{j=0}^{h-1} M'F^j A_{j-h} M \mathbb{E}\big[\Sigma_{\epsilon\epsilon}\big] \tag{A.17}$$

and we deduce that

$$\Gamma_{\tilde{z}\tilde{y},h}\big(\Gamma_{\tilde{y}\tilde{y},0}\big)^{-1} = \mu(irf) \tag{A.18}$$

as required.

(ii) "And Only If" Part. The equality between asymptotic bias and $\mu(irf)$ breaks down for $p-1<p_*$ because by projecting y_{t-h} on fewer than p_* lags in the application of the FWL theorem, the residuals are no longer asymptotically equivalent to ϵ_{t-h}. Instead, because of omitted lags, they depend also on ϵs dated $t-h-1$ and earlier. Following the steps of the calculations for the "if" part, it is straightforward to see that the additional terms create a wedge between between $M'\mu(lfe,0,p)M$ and $\mu(irf)$. ■

Covariance Formulas. In order to express the covariance formulas we need to make the dependence of $\bar{Q}_p$ on (ι,λ) explicit in the notation. We do so by writing $\bar{Q}_p(\iota,\lambda)$. For instance, for $p=q$ we obtain

$$\bar{Q}_q(\iota,\lambda) = \begin{cases} (\Gamma_{YY,0}+\lambda\underline{P}_\Psi)^{-1} & \text{if } \iota = lfe \\ (\Gamma_{YY,0}+\lambda\underline{P}_\Phi)^{-1} & \text{if } \iota = mle \end{cases}.$$

Following steps similar to those in the proof of Theorem 1, it can be shown that the covariance formulas for the companion form model are

$$\begin{aligned}
V(mle,\lambda,p) &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i\Sigma_{EE}F^{j'})\otimes(F^{h-1-i'}\bar{Q}'_p(mle,\lambda)\Gamma_{YY,0}\bar{Q}_p(mle,\lambda)F^{h-1-j}) \\
V(lfe,\lambda,p) &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i\Sigma_{EE}F^{j'})\otimes(\bar{Q}'_p(lfe,\lambda)\Gamma_{YY,j-i}\bar{Q}_p(lfe,\lambda)) \\
Cov(lfe,0,q;lfe,\lambda,p) &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i\Sigma_{EE}F^{j'})\otimes(\Gamma^{-1}_{YY,0}\Gamma_{YY,j-i}\bar{Q}_p(lfe,\lambda)) \\
Cov(lfe,0,q;mle,\lambda,p) &= \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i\Sigma_{EE}F^{j'})\otimes(\Gamma^{-1}_{YY,0}\Gamma'_{YY,h-1-i}\bar{Q}_p(mle,\lambda)F^{h-1-j}).
\end{aligned} \tag{A.19}$$

More specifically, the covariance formulas can be derived as follows. Vectorizing $\zeta_T(\iota, \lambda, p)$ yields

$$\begin{aligned} vecr(\zeta_T(mle, \lambda, p)) &= \sum_{j=0}^{h-1}(F^j \otimes F^{h-1-j'}\bar{Q}_p(mle, \lambda)')vec\left(T^{-1/2}\sum_{t=1}^{T} Y_{t-1}E_t'\right) + o_p(1) \\ vecr(\zeta_T(lfe, \lambda, p)) &= \sum_{j=0}^{h-1}(F^j \otimes \bar{Q}_p'(lfe, \lambda))vec\left(T^{-1/2}\sum_{t=1}^{T} Y_{t-h+j}E_t'\right) + o_p(1). \end{aligned}$$

Based on Assumption 1, the terms

$$vec\left(T^{-1/2}\sum_{t=1}^{T} Y_{t-h+j}E_t'\right), \quad j = 0, \ldots, h-1,$$

jointly satisfy a central limit theorem for vector martingale difference sequences, with covariance matrix $\Sigma_{EE} \otimes \Gamma_{YY,j-i}$. To see why, note that

$$\begin{aligned} &\mathbb{E}\left[vec\left(Y_{t-h+j}E_t'\right) vec\left(Y_{t-h+i}E_t'\right)'\right] \\ &=_{(1)} \mathbb{E}\left[(E_t \otimes I_{nq})vec(Y_{t-h+j})vec(Y_{t-h+i})'(E_t' \otimes I_{nq})\right] \\ &=_{(2)} \mathbb{E}\left[(E_t \otimes I_{nq})Y_{t-h+j}Y_{t-h+i}'(E_t' \otimes I_{nq})\right] \\ &=_{(3)} \mathbb{E}\left[(M\epsilon_t \otimes I_{nq})Y_{t-h+j}Y_{t-h+i}'(\epsilon_t'M' \otimes I_{nq})\right] \\ &=_{(4)} \mathbb{E}\left[(M\epsilon_t\epsilon_t'M' \otimes Y_{t-h+j}Y_{t-h+i}')\right] \\ &=_{(5)} \Sigma_{EE} \otimes \Gamma_{YY,j-i} + o(1). \end{aligned}$$

The first equality follows because $vec(AB) = (B' \otimes I)vec(A)$. The second and third equalities are by definition. The fourth equality follows by Lemma A-1 below by setting $\nu = M\epsilon$. The covariance formulas in (A.19) can now be generating by noting that $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and $\bar{Q}_q(lfe, 0) = \Gamma_{YY,0}^{-1}$. ■

Lemma A-1 *For any $n \times 1$ vector ν and square matrix C of dimension d,*

$$(\nu \otimes I_d)C(\nu' \otimes I_d) = (\nu\nu') \otimes C.$$

Proof of Lemma A-1. By direct calculation,

$$\begin{aligned}(\nu \otimes I_d)C(\nu' \otimes I_d) &= \begin{bmatrix}\nu_1 I_d \\ \vdots \\ \nu_n I_d\end{bmatrix} C \begin{bmatrix}\nu_1 I_d & \cdots & \nu_n I_d\end{bmatrix} \\ &= \begin{bmatrix}\nu_1^2 C & \cdots & \nu_1\nu_n C \\ \vdots & \ddots & \\ \nu_n\nu_1 C & & \nu_n^2 C\end{bmatrix} \\ &= \begin{bmatrix}\nu_1^2 & \cdots & \nu_1\nu_n \\ \vdots & \ddots & \\ \nu_n\nu_1 & & \nu_n^2\end{bmatrix} \otimes C \\ &= (\nu\nu') \otimes C. \quad \blacksquare\end{aligned}$$

Lemma A-2 *Let M be an $nq \times n$ matrix, A an $nq \times nq$ matrix, Ξ an $n \times n_{sh}$ matrix, where $n_{sh} \leq n$, and W is a $n \times n$ symmetric positive definite weight matrix. Then*

$$\|M'AM\Xi\|_W^2 = \text{tr}\left\{[(MWM') \otimes (M\Xi\Xi'M')]vecr(A)vecr(A)'\right\}.$$

Proof of Lemma A-2. The result can be derived as follows:

$$\begin{aligned}\|M'AM\Xi\|_W^2 &=_{(1)} \text{tr}\left\{W\big(M'AM\Xi)(\Xi'M'A'M\big)\right\} \\ &=_{(2)} \text{tr}\left\{\underbrace{MWM'}_{=C} A \underbrace{M\Xi\Xi'M'}_{=B} A'\right\} \\ &=_{(3)} vecr(A)'[(MWM') \otimes (M\Xi\Xi'M')]vecr(A) \\ &=_{(4)} \text{tr}\left\{[(MWM') \otimes (M\Xi\Xi'M')]vecr(A)vecr(A)'\right\}.\end{aligned}$$

The second equality uses $\text{tr}(AB) = \text{tr}(BA)$. The third equality is based on

$$\text{tr}\{CABA'\} = vecr(A)'(C \otimes B)vecr(A).$$

The last equality uses $a'Ba = \text{tr}[Baa']$. $\blacksquare$

B Further Details on the Monte Carlo Simulations

Parameterization of the DGP: The specific values of the F and A_j matrices are provided in the replication code.

Parameterization of the Prior. We need to solve (53) for $\underline{\phi}$ as a function of $\underline{\psi}$. Note that

$$\text{vec}(ABC) = (C' \otimes A)\text{vec}(B).$$

Thus,

$$\begin{aligned} \text{vec}(\underline{\psi}) &= \sum_{j=0}^{h-1} \text{vec}\big(F^j \underline{\phi} F^{h-1-j}\big) \\ &= \left(\sum_{j=0}^{h-1} \big(F^{h-1-j'} \otimes F^j\big)\right) \text{vec}(\underline{\phi}) \end{aligned}$$

In turn,

$$\text{vec}(\underline{\phi}) = \left(\sum_{j=0}^{h-1} \big(F^{h-1-j'} \otimes F^j\big)\right)^{-1} \text{vec}(\underline{\psi}).$$

Additional Results. Tables A-1 and A-2 are analogous to Table 1 for shorter and larger sample sizes, respectively. The general pattern remains intact: under no misspecification, the performance of LFE and MLE is comparable; under misspecification, LFE performs significantly better, and is almost always picked by our selection criterion. Joint selection is always very close to the best possible fixed-estimand selection.

Tables A-3 and A-4 are analogous to Table 2 for shorter and larger sample sizes, respectively. The same conclusion applies: under no misspecification, MDD and PC selection yield similar results; under misspecification, PC clearly outperforms MDD. In our simulation setup, for small sample sizes, MDD seems to do better than PC.

Figure A-1 plots the selected lag length underlying Table 2. Under no misspecification, MDD always provides a consistent estimate of the true lag length p_*. PC is able to adapt the selected lag length to improve the risk. For this reason, it tends to overselect the lag length to absorb some of the misspecification and reduce the bias.

Table A-1: Finite Sample Risk Differentials for $\hat{y}_{T+h}(\iota, \hat{\lambda}, p)$, $T = 100$.

	$\alpha = 0$				$\alpha = 2$			
p	LFE	MLE	Joint	π	LFE	MLE	Joint	π
Horizon $h = 2$								
1	-1003	-1014	-1003	70	-1002	-1028	-1002	100
2	-975	-998	-977	64	-1002	-973	-1002	100
4	-909	-932	-914	56	-989	-968	-989	99
6	-841	-856	-843	46	-872	-837	-871	94
$\hat{p}$	-851	-864	-850	56	-872	-837	-871	94
Horizon $h = 4$								
1	-2259	-2329	-2262	79	-2427	-2552	-2426	97
2	-2190	-2302	-2204	66	-2224	-2481	-2226	97
4	-2023	-2187	-2050	60	-2087	-2101	-2087	99
6	-1859	-2009	-1909	53	-1836	-1878	-1837	98
$\hat{p}$	-1905	-2026	-1932	65	-1836	-1874	-1835	99
Horizon $h = 6$								
1	-3519	-3709	-3537	74	-4492	-4908	-4488	94
2	-3411	-3684	-3458	63	-4092	-4750	-4093	93
4	-3155	-3531	-3254	54	-3632	-4219	-3640	94
6	-2945	-3320	-3082	49	-3199	-3683	-3208	94
$\hat{p}$	-3061	-3345	-3107	62	-3246	-3683	-3248	96

Notes: The finite sample risk differentials are computed relative to $\hat{y}_{T+h}(lfe, 0, q = 6)$. π is the percentage of times that LFE is selected by the PC criterion.

Table A-2: Finite Sample Risk Differentials for $\hat{y}_{T+h}(\iota, \hat{\lambda}, p)$, $T = 5000$.

	$\alpha = 0$				$\alpha = 2$			
p	LFE	MLE	Joint	π	LFE	MLE	Joint	π
Horizon $h = 2$								
1	-8	-8	-8	30	9	9	9	84
2	-8	-8	-8	21	7	8	7	100
4	-7	-8	-8	19	3	4	3	100
6	-7	-7	-7	22	-1	0	-1	100
$\hat{p}$	-8	-8	-8	25	-1	0	-1	100
Horizon $h = 4$								
1	-17	-17	-17	30	9	11	9	92
2	-16	-17	-17	26	7	10	7	98
4	-16	-17	-16	25	-1	5	-1	100
6	-16	-16	-16	28	-4	-1	-4	99
$\hat{p}$	-16	-17	-17	20	-4	-1	-4	100
Horizon $h = 6$								
1	-25	-26	-25	19	7	12	7	95
2	-24	-25	-25	15	3	11	4	96
4	-23	-25	-24	13	-3	5	-3	96
6	-23	-24	-24	12	-5	0	-5	93
$\hat{p}$	-24	-25	-25	11	-5	0	-5	97

Notes: The finite sample risk differentials are computed relative to $\hat{y}_{T+h}(lfe, 0, q = 6)$. π is the percentage of times that LFE is selected by the PC criterion.

Table A-3: Finite Sample Risk Differentials, PC vs. MDD Selection, $T = 100$

	Horizon $h = 2$				Horizon $h = 4$			
	$\alpha = 0$		$\alpha = 2$		$\alpha = 0$		$\alpha = 2$	
p	LFE	MLE	LFE	MLE	LFE	MLE	LFE	MLE
1	2	4	7	10	0	3	-2	5
2	8	6	15	15	7	4	19	11
4	14	11	14	16	15	7	25	25
6	20	19	24	27	22	15	34	33
$\hat{p}$	17	19	19	26	16	15	23	31

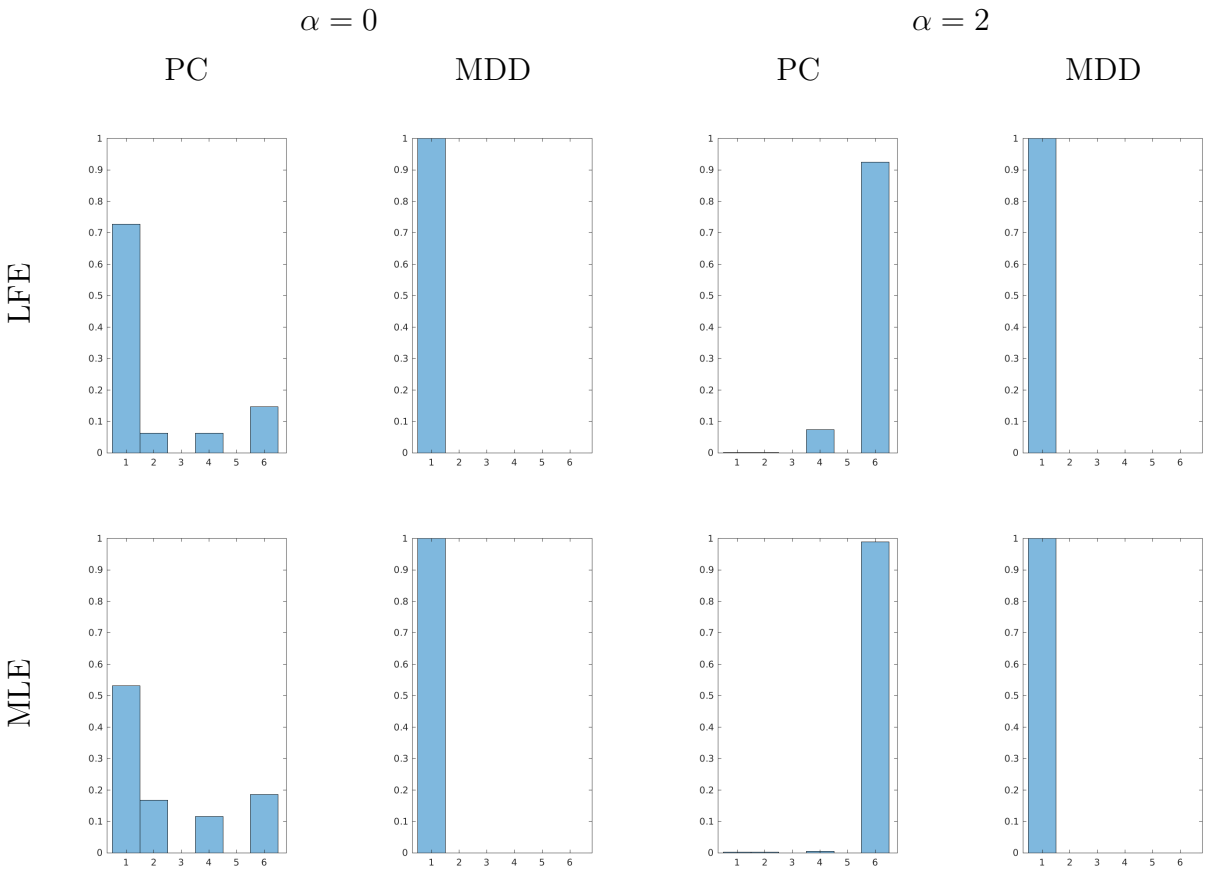
Notes: We report risk differentials of PC-based versus MDD-based selection relative to the MDD risk, in percent. A negative number indicates that PC selection yields a lower risk than MDD selection.

Table A-4: Finite Sample Risk Differentials, PC vs. MDD Selection, $T = 5000$

	Horizon $h = 2$				Horizon $h = 4$			
	$\alpha = 0$		$\alpha = 2$		$\alpha = 0$		$\alpha = 2$	
p	LFE	MLE	LFE	MLE	LFE	MLE	LFE	MLE
1	-3	0	-1	-2	-7	0	0	-5
2	1	2	-1	1	-1	4	9	-6
4	2	0	-52	-54	4	1	-132	-49
6	0	-1	-107	-96	3	0	-164	-105
$\hat{p}$	-2	1	-107	-95	-5	2	-141	-105

Notes: We report risk differentials of PC-based versus MDD-based selection relative to the MDD risk, in percent. A negative number indicates that PC selection yields a lower risk than MDD selection.

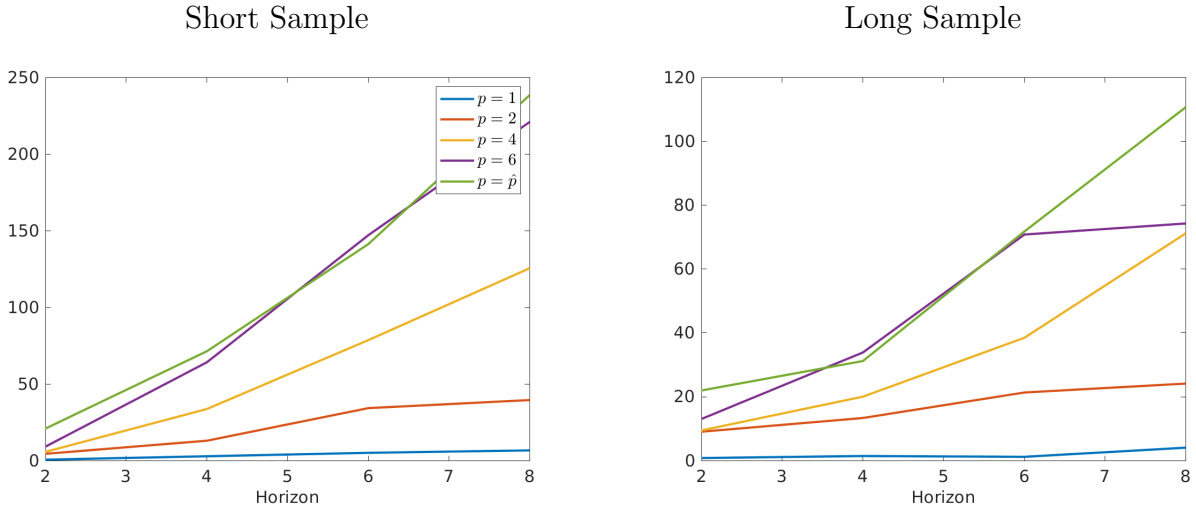
Figure A-1: Distribution of $\hat{p}$, Horizon $h = 4$, $T = 500$



C Additional Empirical Results

Figure A-2 shows the norm difference between LP and VAR IRF estimators for different lag lengths as a function of the horizon. As expected, differences grow large the further the horizon of interest, and as more lags are included in the VAR. While apparently obvious, these patterns suggest that the selection decisions will have important risk implications, reinforcing the relevance of the use of IRFC.

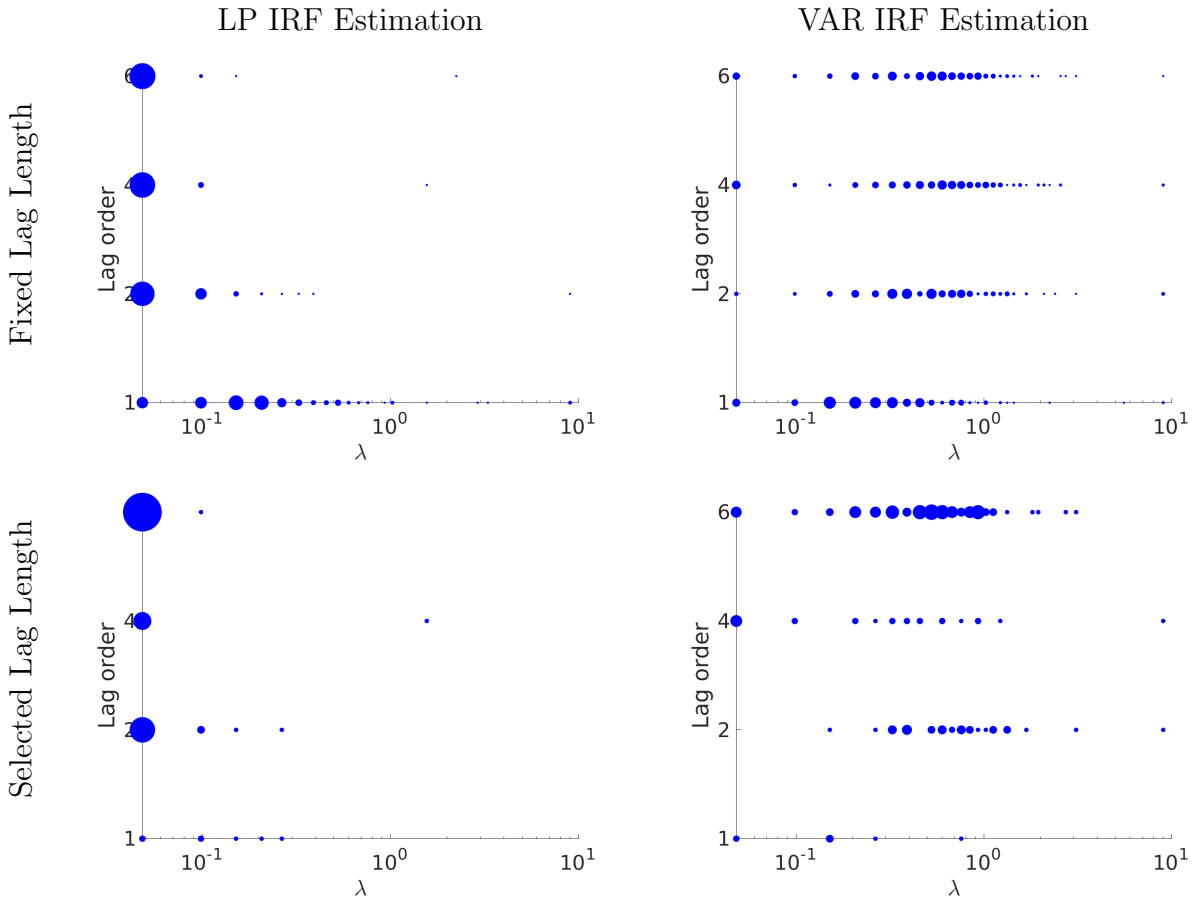
Figure A-2: Norm Difference Between LP and VAR-Based IRF Estimates



Notes: Average normed difference between IRF estimators, $\mathbb{E}[\|\Psi(lfe) - \Psi(mle)\|^2]$, for different lag lengths, across different horizons.

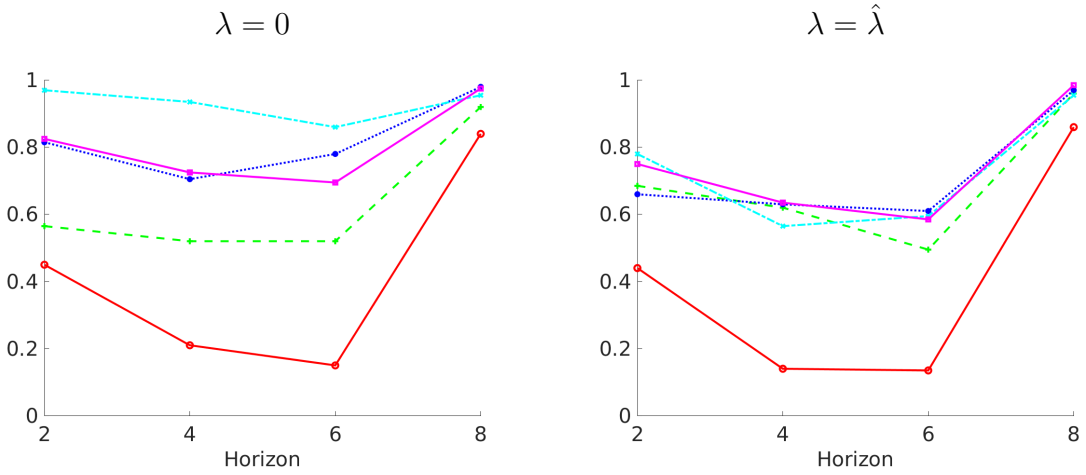
Figures A-3 and A-4 are similar to those reported in the main text, except that they are generated based on the long samples ranging from 1984:Q1 to 2019:Q4.

Figure A-3: Distribution of Selected Hyperparameter, $h = 6$.



Notes: Grid values of IRFC-selected shrinkage hyperparameters $\hat{\lambda}(p)$ for different fixed lag orders p. The diameter of the dots is proportional to the frequency of the $(\hat{\lambda}(p), p)$ frequency. Fixed lag length refers to $(\hat{\lambda}(p), p)$ and each p-row represents 200 samples. Selected lag length is $(\hat{\lambda}(\hat{p}), \hat{p})$ and the number of samples across the four $\hat{p}$ rows add up to 200. Estimation sample 1984:Q1-2019:Q4.

Figure A-4: Selection of LP versus VAR IRF Estimate



Notes: Fraction of times the IRFC selects the LP IRF estimator under different lag lengths and across different horizons. Estimation sample 1984:Q1-2019:Q4.