VAR Hyperparameter Determination Under Misspecification

Oriol González-Casasús* Frank Schorfheide

University of Pennsylvania *University of Pennsylvania*,

CEPR, PIER, NBER

Preliminary Version: June 26, 2024

Abstract

Prior hyperparameters for Bayesian vector autoregressions (VARs) are often determined by maximization of a marginal data density (MDD). However, if a VAR is misspecified, it is not clear that a MDD based hyperparameter determination is desirable. In this paper we use an asymptotically unbiased estimate of the multi-step forecasting risk to determine the hyperparameters of shrinkage estimators in an environment in which the VAR forecasting model is locally misspecified. We show that due to misspecification the prediction results do not directly carry over to impulse response function estimation and discuss the needed modifications to target impulse response estimation risk. The resulting criterion can be used for hyperparameter determination in local projection applications. The hyperparameter selection approach is illustrated in a Monte Carlo study and an empirical application. (JEL C11, C32, 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 participants of the Penn Econometrics Lunch for helpful comments and suggestions.

1 Introduction

Bayesian vector autoregressions (VARs) have been successfully used for macroeconomic forecasting since the early 1980s. They combine the VAR likelihood function with a prior distribution that shrinks the distance between 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. 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, 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). In this paper, we will consider hyperparameter selection based on an estimate of the prediction risk.

The MLE associated with a Gaussian likelihood function minimizes in-sample one-step-ahead forecast errors. If the goal is h-step-ahead forecasting, then one could either iterate the one-step-ahead MLE-based forecasts forward, or one could use a multi-step regression that projects an $n \times 1$ vector of variables 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. The multi-step estimation objective function could be interpreted as a quasi-likelihood function that ignores the serial correlation in the sequence of h-step-ahead forecast error. This 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 a criterion that provides an estimate of the h-step ahead prediction risk and can be used to choose the prior hyperparameter(s), select between an MLE and LFE based shrinkage estimator, and determine the lag length of the VAR.

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 vector moving average (VMA) process that drifts toward a VAR(p_*), where $p_* \leq q$ at rate $T^{-1/2}$. Here T is the size of the estimation sample and the forecast origin. The forecaster uses a VAR(p) with $p \leq q$ lags to generate h-step-ahead forecasts. Two predictors are considered in S2005: an MLE plug-in predictor that uses the MLE and iterates the VAR(p)

forward for h-periods and an LFE plug-in predictor that directly projects y_t onto y_{t-h} and additional lag. In the absence of misspecification, 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 data generating process (DGP) with a VAR of order p. The $T^{-1/2}$ drift in the misspecification balances the bias-variance trade-off among the two estimators.

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 current paper extends the class of predictors by considering MLE or LFE posterior mean (or shrinkage) estimators indexed by a hyperparameter λ that scales the precision of the prior distribution from which the estimators are derived. PC_T can be viewed as providing an (asymptotically) unbiased risk estimate (URE) along the lines of Stein (1981) that can be minimized with respect to the choice of the predictor type ι and the hyperparameter λ. As a benchmark for the (ι, λ) selection we consider an oracle that can determine the predictor and hyperparameter based on the knowledge of the conditional expectation of y_{T+h}. Unlike in compound decision problems, in our environment it is not feasible to implement a selection that achieves the oracle risk. Instead, we provide simulation evidence on the superior performance of the PC-based hyperparameter selection and compare it to a selection based on a (quasi) MDD. Finally, the procedure is applied in an empirical illustration in which we document its performance across a large set of VARs comprising different sets of macroeconomic variables.

In the context of multi-step ahead forecasting 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.

It is 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 provides only an asymptotically unbiased estimate of the final prediction risk but not a consistent estimate as in Shibata (1980) framework.

While in empirical work the hyperparameter selection based on MDDs dominates, there is theoretical 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.

Multi-step estimators are also used to estimate impulse response functions (IRFs). Jorda (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 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(p) using a one-step-ahead (quasi) likelihood objective function and then iterating the VAR forward. The likelihood-based estimation in the forecasting context corresponds to the VAR estimation in IRF context, whereas the loss-function based estimation corresponds to the LP.

Plagborg-Moller 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-Moller, and Wolf (2022). 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, 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, while it remains the case that the LP estimator has larger variance than the VAR estimator.

Montiel Olea and Plagborg-Moller (2021) propose a lag-augmented estimator to alleviate inference problems caused by serial correlation in LPs. Montiel Olea, Plagborg-Moller, Qian, and Wolf (2024) show that the lag augmentation also is essential for inference under mis-

specification because it properly centers LP confidence intervals. We demonstrate that lag augmentation is also useful for shrinkage estimation and modify PC to generate an asymptotically unbiased estimate of the IRF estimation risk that can be used for hyperparameter selection.

The remainder of the paper is organized as follows: Section 2 describes the data generating process (DGP), the shrinkage estimators and predictors, and the prediction risk associated with them. We initially focus on the case of using a potentially misspecified VAR(1) to generate the forecasts. Section 3 discusses hyperparameter selection based on an asymptotically unbiased risk estimate and a (quasi) marginal data density. Implications for IRF inference based on local projections instead of VARs are discussed in Section 4. Section 5 presents results from Monte Carlo experiments with a VAR(1). Section 6 has an extension to the VAR(p) and the case of unknown lag length. An empirical forecasting application is provided in Section 7. Finally, Section 8 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 vector moving average process. The predictors are described in Section 2.1. The degree of shrinkage is determined by a hyperparameter. Setting this hyperparameter to zero leads to the 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 6. 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

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

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

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

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:

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

The prior are indexed by the hyperparameter $\tilde{\lambda}$ that controls the degree of shrinkage. The mean of the prior distribution is 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:

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

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

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

[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

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

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

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

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

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

we obtain the quasi-posterior

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

This leads to the LFE shrinkage predictor

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

2.2 Drifting DGP and Prior

We assume that the sample has been generated from a covariance stationary data generating process (DGP) with an infinite-dimensional VMA representation. While the sample size T is fixed in practice, we would like to 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 will vanish at rate $O(T^{-1})$ whereas the misspecification bias does not disappear. Thus, eventually, the loss-function-based predictor will dominate along this asymptote, even if the misspecification is small.

To generate asymptotics that better reflect the finite-sample trade-offs faced by the forecaster we have two choices: either increase the dimensionality of the forecasting model with sample size or let the DGP drift toward the forecasting model. As in S2015, 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.

In addition to the DGP, we also assume that the prior means used to construct the shrinkage estimators drift. They are located within a $T^{-1/2}$ radius from F:

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

Moreover, we re-scale the hyperparameter as follows:

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

In slight abuse of notation, we replace the $\tilde{\lambda}$ 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 that 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,

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

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 $0 \leq w - u \leq \delta$,

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

(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.

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

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

where

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

Moreover, $\zeta_T(\iota, \lambda)$ converges weakly to $\zeta(\iota, \lambda)$, where

$$\begin{pmatrix} \zeta_T(mle, \lambda) \\ \zeta_T(lfe, \lambda) \\ \zeta_T(mle, \lambda') \\ \zeta_T(lfe, \lambda') \end{pmatrix} \Longrightarrow \begin{pmatrix} \zeta(mle, \lambda) \\ \zeta(lfe, \lambda) \\ \zeta(mle, \lambda') \\ \zeta(lfe, \lambda') \end{pmatrix} \sim N\left(\mathbf{0}, \boldsymbol{V}(\lambda, \lambda')\right)$$

and

$$\boldsymbol{V}(\lambda,\lambda') = \begin{bmatrix} V(mle,\lambda,\lambda) & & & & \\ V(mle,lfe,\lambda,\lambda) & V(lfe,\lambda,\lambda) & & & \\ V(mle,\lambda',\lambda) & V(mle,lfe,\lambda',\lambda) & V(mle,\lambda',\lambda') & & \\ V(mle,lfe,\lambda',\lambda) & V(lfe,\lambda',\lambda) & V(mle,lfe,\lambda',\lambda') & V(lfe,\lambda',\lambda') \end{bmatrix}.$$

$\bar{\Psi}_T(\iota,\lambda)$ ultimately converges to F^h. 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 $\mu(lfe,0) = \mu(pov)$, 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.

Formulas for the partitions of $\mathbf{V}(\lambda,\lambda')$ are provided in the Online Appendix. The shrinkage also affects variance terms $V(\iota,\iota',\lambda,\lambda')$. The larger the precision hyperparameter λ, the smaller the sampling variance of the shrinkage estimator. Moreover, holding λ fixed, the variance of the likelihood based shrinkage estimator is smaller than the variance of the loss function based shrinkage estimator:

$$V(mle,\lambda,\lambda) < V(lfe,\lambda,\lambda).$$

The latter is inefficient, as it ignores the serial correlation of the h-step-ahead forecast errors in its estimation objective function.

Prediction Risk. The next theorem characterizes the asymptotic prediction risk

$$\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota,\lambda)) = \lim_{T\longrightarrow\infty} T\mathcal{R}(\hat{y}_{T+h}(\iota,\lambda)) \tag{20}$$

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) guarantee the finite-sample moments of the estimators converge to the moments of the limit distribution.

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

$$\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota,\lambda)) = \underbrace{\|\delta(\iota,\lambda) - \alpha(\mu(pov) - \mu(\iota,\lambda))\|^2_{W\otimes\Gamma_{yy,0}}}_{=:\,\bar{\mathcal{R}}_B(\iota,\lambda)} + \underbrace{tr\Big\{(W\otimes\Gamma_{yy,0})V(\iota,\lambda,\lambda)\Big\}}_{=:\,\bar{\mathcal{R}}_V(\iota,\lambda)} + C, \tag{21}$$

where

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

The prediction risk is decomposed in a bias term $\bar{\mathcal{R}}_B(\iota,\lambda)$ and a variance term $\bar{\mathcal{R}}_V(\iota,\lambda)$. The constant C does not depend on the forecast $\hat{y}_{T+h}(\iota,\lambda)$. 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 consider two different methods of determining the hyperparameter λ. The first method relies on an asymptotically unbiased (prediction) risk estimate (URE). The URE objective function is a generalization of the PC criterion proposed in S2005. The second method uses a (quasi) marginal data density (MDD) to select λ. While the first criterion can also be used to choose between the two estimators – LFE based shrinkage versus MLE based shrinkage – the MDD cannot. The PC-based hyperparameter determination is discussed in Section 3.1. As a benchmark we discuss an oracle-based hyperparameter selection in Section 3.2 and contrast the PC objective function with the oracle objective function. At last, we derive an MDD criterion for the hyperparameter selection in Section 3.3 which has been used in practice and will be included in the Monte Carlo experiments of Section 5. 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). Thus, in this paper we are contrasting PC-EB and MDD-EB estimation and forecasting approaches.

3.1 Asymptotically Unbiased Risk Estimation

We proceed by deriving an objective function for the hyperparameter λ and the choice of estimator ι that relies on an asymptotically unbiased estimate of the prediction risk of $\hat{y}_{T+h}(\iota,\lambda)$. We begin by characterizing the in-sample prediction loss associated with $\bar{\Psi}_T(\iota,\lambda)y_{t-h}$.

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

$$MSE(\iota,\lambda) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{\Psi}_T(\iota,\lambda)y_{t-h})(y_t - \bar{\Psi}_T(\iota,\lambda)y_{t-h})'. \tag{22}$$

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

$$\Delta_{L,T}(\iota,\lambda) = T\left(tr\left\{W\cdot MSE(\iota,\lambda)\right\} - tr\left\{W\cdot MSE(lfe,0)\right\}\right) \geq 0. \tag{23}$$

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

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

It is important to note that due to the drifting DGP, the normalization of the prediction risk in (17), and the scaling by T, 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*

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

where $\hat{\bar{\mathcal{R}}}_V(lfe, \iota, 0, \lambda)$ is a consistent estimate of $\operatorname{tr}\left\{(W \otimes \Gamma_{yy,0})V(lfe, \iota, 0, \lambda)\right\}$.

The term $2\hat{\bar{\mathcal{R}}}_V(lfe, \iota, 0, \lambda)$ 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. $\hat{\bar{\mathcal{R}}}_V(lfe, \iota, 0, \lambda)$ can be consistently estimated, for instance, by replacing the matrices F and $\Sigma_{\epsilon\epsilon}$ that appear in the covariance expressions of Theorem 1 with a quasi MLE. Moreover, population autocovariance matrices $\Gamma_{yy,j}$ can be replaced by their sample analogues. $PC_T(\iota, \lambda)$ can be used to choose between MLE and LFE based shrinkage and to select the hyperparameter λ. After combining Definition 1 of the selection criterion with the MSE differential formula in (23) and Theorem 3 we can deduce that

$$\begin{aligned} &\mathbb{E}\big[PC_T(\iota, \lambda) - PC_T(\iota', \lambda')\big] \\ &\quad = \quad \mathbb{E}\big[\Delta_{L,T}(\iota, \lambda) - \Delta_{L,T}(\iota', \lambda')\big] + 2\mathbb{E}\big[\hat{\bar{\mathcal{R}}}_V(lfe, \iota, 0, \lambda) - \hat{\bar{\mathcal{R}}}_V(lfe, \iota', 0, \lambda')\big] \\ &\quad \longrightarrow \quad \bar{\mathcal{R}}(\hat{y}_{T+h}(\iota, \lambda)) - \bar{\mathcal{R}}(\hat{y}_{T+h}(\iota', \lambda')). \end{aligned} \tag{24}$$

Thus, the PC differential provides an asymptotically unbiased estimate of the risk differential for the predictors $\hat{y}_{T+h}(\iota, \lambda)$ and $\hat{y}_{T+h}(\iota', \lambda')$.

Remark. Theorems 1 and 2 suggest that an alternative asymptotically unbiased estimate of the prediction risk differential is

$$T\left\|\bar{\Psi}_T(\iota, \lambda) - \bar{\Psi}_T(lfe, 0)\right\|^2_{W \otimes \Gamma_{yy,0}} + 2\operatorname{tr}\left\{(W \otimes \Gamma_{yy,0})V(lfe, \iota, 0, \lambda)\right\}. \tag{25}$$

In unreported simulations we find that both UREs perform equally. This alternative URE is closely connected to the URE for the IRF risk proposed in Section 4.

3.2 Comparison to Oracle Hyperparameter Selection

So far we have shown that our PC criterion in Definition 1 provides an unbiased estimate for the asymptotic risk. In order to put the PC-based hyperparameter selection into context, we now define an oracle objective. The oracle has an informational advantage relative to the forecaster and is therefore potentially able to choose a more favorable hyperparameter. In the literature on compound decisions, building on Stein (1981), a key result is that the hyperparameter selection based on an unbiased risk estimate is asymptotically as good as the oracle's hyperparameter determination in the sense that the difference between the associated prediction risks vanishes as the sample size tends to infinity. While we will show that this is not the case in the forecasting setup considered in this paper, the oracle will nonetheless provide an important benchmark.

Recall that we are distinguishing between two independent processes, $\{y_t\}$ and $\{\tilde{y}_t\}$, that have the same stochastic properties. The former is used for estimation and the latter for forecasting. The loss function $L(\tilde{y}_{T+h}, \hat{y}_{T+h})$ is subject to three sources of randomness: (i) the value of the process at the forecast origin, $\tilde{y}_T$; shocks that determine the target of the forecast, $\tilde{y}_{T+h}$; and the realization of the estimation sample $\{y_t\}_{t=1}^T$. By assumption $\bar{\Psi}_T(\iota, \lambda)$ is independent of $(\tilde{y}_T, \tilde{y}_{T+h})$ and $PC_T(\iota, \lambda)$ can never capture the variation in $(\tilde{y}_T, \tilde{y}_{T+h})$. Thus, to construct an oracle objective function we define the (scaled) expected loss conditional on $\{y_t\}_{t=1}^T$ denoted by

$$\mathcal{L}\big(\tilde{y}_{T+h}, \hat{y}_{T+h}\big) = \mathbb{E}\big[L(\tilde{y}_{T+h}, \hat{y}_{T+h}) \,\big|\, \{y_t\}_{t=1}^T\big]. \tag{26}$$

Exploiting the properties of the quadratic loss function and matching the structure of the $PC_T(\iota, \lambda)$ objective function in Definition 1, we define the oracle objective function

$$\begin{aligned} Q_T(\iota, \lambda) \;\; = \;\; & T\big[\mathcal{L}\big(\mathbb{E}_T[\tilde{y}_{T+h}], \hat{y}_{T+h}(\iota, \lambda)\big) - \mathcal{L}\big(\mathbb{E}_T[\tilde{y}_{T+h}], \hat{y}_{T+h}(lfe, 0)\big)\big] \\ & +2\left(tr\left\{W\zeta_T(lfe, 0)\Gamma_{yy,0}\zeta_T(lfe, 0)'\right\} - \hat{\mathcal{R}}_V(lfe, 0)\right). \end{aligned}$$

Notice that the term in the second line does not depend on (ι, λ) and therefore does not affect the shape of the objective function. It is simply a level correction so that $Q_T(\iota, \lambda)$ is aligned with the $PC_T(\iota, \lambda)$ differential. The informational advantage of the oracle over the forecaster is that it knows $\mathbb{E}_T[\tilde{y}_{T+h}]$ in addition to the estimation sample $\{y_t\}_{t=1}^T$. Thus, it can choose (ι, λ) to target the conditional expectation directly. We define the oracle estimator of (λ, ι) as

$$(\dot{\lambda}, i) = \underset{\lambda \geq 0,\, \iota \in \{mle, lfe\}}{\operatorname{argmin}} \; Q_T(\iota, \lambda). \tag{27}$$

The following theorem characterizes the difference between the $PC_T(\iota, \lambda)$ and the oracle objective function.

Theorem 4 *Suppose that Assumption 1 is satisfied. Then:*

$$\begin{aligned} PC_T(\iota,\lambda) - PC_T(lfe,0) &= Q_T(\iota,\lambda) + o_p(1) \\ &\quad -2\left(tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} - \hat{\mathcal{R}}_V(lfe,\iota,0,\lambda)\right) \\ &\quad -2tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\left[\delta(\iota,\lambda) + \alpha(\mu(\iota,\lambda) - \mu(pov))\right]'\right\}. \end{aligned}$$

The terms in the second and third line on the right-hand side of Theorem 4 create a wedge between the two objective functions that does not vanish asymptotically. While zero on average, for any given sample $\{y_t\}_{t=1}^T$ they are non-zero. Thus, the $PC_T(\iota, \lambda)$ selection in our setting is not optimal in the usual sense of getting close to the oracle. The term in the second line captures the difference between the weighted outer product of $\zeta_T(lfe, 0)$ and $\zeta_T(\iota, \lambda)$ and its expected value $\hat{\mathcal{R}}_V(lfe, \iota, 0, \lambda)$, which is the MSE adjustment term in Definition 1. The term in the third line arises from the expansion of the MSE terms in the $PC_T(\iota, \lambda)$ criterion and is not present in the the oracle objective function.[2] In the Monte Carlo experiments in Section 5 we will use the prediction associated with the oracle selection of the hyperparameters as a benchmark.

The wedge between PC objective function and the oracle objective function in Theorem 4 should not be interpreted as a shortcoming of the PC criterion. It is a by-product of the local misspecification framework that cannot be overcome in our time series setting, also by other hyperparameter selection criteria.

3.3 MDD Based Hyperparameter Selection

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

$$Y = X\Psi' + V. \tag{28}$$

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

[2] This problem also arises if the prediction criterion is replaced by a direct estimate of the parameter estimation risk.

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}:

$$\Sigma_{vv} \sim IW(\underline{\nu}, \underline{S}). \tag{29}$$

Define

$$\bar{S} = \underline{S} + (\lambda T)\underline{\Psi}_T \underline{P}_\Psi \underline{\Psi}'_T + Y'Y - \bar{\Psi}_T \bar{P}_\Psi \bar{\Psi}'_T, \quad \bar{\nu} = \underline{\nu} + T. \tag{30}$$

It can be shown that the MDD takes the form

$$p(Y|\iota,\lambda) = \int\int p(Y|\Psi,\Sigma_{vv})p(\Psi,\Sigma_{vv})d\Psi d\Sigma_{vv} = (2\pi)^{-nT/2}\frac{|\lambda T\underline{P}_\Psi|^{n/2}}{|\bar{P}_\Psi|^{n/2}}\frac{\underline{C}_{IW}}{\bar{C}_{IW}}, \tag{31}$$

where

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

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

$$\begin{aligned} MDD_T(\iota,\lambda) &= 2\big[\ln p(Y|\iota,\infty) - \ln p(Y|\iota,\lambda)\big] \\ &= \bar{\nu}\big\{\ln|\bar{S}_T(\iota,\lambda)| - \ln|\bar{S}_T(\iota,\infty)|\big\} + n\big\{\ln|\lambda\underline{P}_\Psi + X'X/T| - \ln|\lambda\underline{P}_\Psi|\big\}, \end{aligned} \tag{32}$$

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 (32) 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. 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 Implications for IRF Estimation

The goal of many VAR applications is to estimate impulse functions (IRFs) for structural shocks, instead of forecasting. IRF estimates can be obtained in two ways. First, one could estimate a VAR using a likelihood-based (shrinkage) estimator and iterate the estimated VAR forward to trace out the effect of a shock. Alternatively, one could conduct a so-called local projection (LP) as proposed by Jorda (2005), which regresses y_t on y_{t-h} and possibly higher-order lags as controls. In our notation, the VAR-based IRF estimate corresponds to $\bar{\Psi}_T(mle, \lambda)$, and $\bar{\Psi}(lfe, \lambda)$ is the local projection IRF estimate.

It was originally claimed, and is now widely accepted in the literature, that LPs are inherently more robust to misspecification than VAR-based IRFs. However, theoretical evidence in support of that statement is scarce. Plagborg-Moller and Wolf (2021) prove that LPs and VARs estimate the same IRFs in population when controlling for the infinite past. If only a fixed number p of lags are included, then the two IRF estimands approximately agree out to horizon p, but not further. Li, Plagborg-Moller, and Wolf (2022) conduct a large simulation experiment using several variations of these two estimands and conclude that the preferred estimator largely depends on how much one trades off variance and bias. These results provide limited practical guidance.

In this section, we show that LPs do not always dominate VAR-based IRF estimates in an MSE sense. In particular, LPs are not always more robust to misspecification than

VARs when the forecaster uses a quadratic loss function. We also demonstrate that the PC criterion proposed in Definition 1 is not valid when interest lies in IRFs instead of forecasting. We propose a valid URE for the IRF risk to determine the type of predictor and degree of shrinkage.

4.1 Non-dominance of LPs Over VARs

Recall we can write the DGP in (13) as an MA(∞) process of the form

$$y_t = \sum_{s=0}^{\infty} F^s \epsilon_{t-s} + \frac{\alpha}{\sqrt{T}} \left(\sum_{s=0}^{\infty} F^s L^s \right) \left(\sum_{j=1}^{\infty} A_j L^j \right) \epsilon_t. \tag{33}$$

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

$$\frac{\partial y_t}{\partial \epsilon_{t-h}'} = F^h + \frac{\alpha}{\sqrt{T}} \sum_{j=0}^{h-1} F^j A_{h-j} = F^h + \frac{\alpha}{\sqrt{T}} \mu(irf, h), \quad h \geq 1, \tag{34}$$

and is the sum of the first-order term F^h and a $O(1/\sqrt{T})$ misspecification term.

IRFs typically aim to trace out the effect of a structural shock. Linearized dynamic stochastic general equilibrium (DSGE) models imply that the one-step-ahead forecast errors ϵ_t are a linear combination of orthonormal structural shocks η_t, e.g.,

$$\epsilon_t = \Sigma_{\epsilon\epsilon}^{tr} \Omega \eta_t, \tag{35}$$

where $\Sigma_{\epsilon\epsilon}^{tr}$ is the lower triangular Cholesky factor of the reduced-form covariance matrix $\Sigma_{\epsilon\epsilon}$ and Ω is an orthonormal matrix. For concreteness, suppose we are interested in the response with respect to the ith shock $\eta_{i,t}$, for some $i \in \{1, \ldots, n\}$, and let q be the ith column in $\Sigma_{\epsilon\epsilon}^{tr}\Omega$. We define the asymptotic estimation risk for the IRF with an impact vector q associated with the estimator $\bar{\Psi}_T(\iota, \lambda)$ as

$$\bar{\mathcal{R}}_{irf}\big(\bar{\Psi}_T(\iota, \lambda); h, q\big) = \lim_{T \longrightarrow \infty} T\mathbb{E}\Big[\big\| \big(F^h + \alpha T^{-1/2} \mu(irf, h) - \bar{\Psi}_T(\iota, \lambda)\big) q \big\|_W^2 \Big]. \tag{36}$$

Using the calculations underlying the proof of Theorem 2, one can show that

$$\begin{aligned} &\bar{\mathcal{R}}_{irf}\big(\bar{\Psi}_T(\iota, \lambda); h, q\big) \\ &= \big\| \delta(\iota, \lambda) - \alpha(\mu(irf, h) - \mu(\iota, \lambda)) \big\|_{W \otimes qq'}^2 + \operatorname{tr}\Big\{ (W \otimes qq') V(\iota, \lambda, \lambda) \Big\}. \end{aligned} \tag{37}$$

Let us focus for simplicity on the case of no shrinkage ($\lambda = 0$). In that case, the asymptotic risk formula simplifies to

$$\bar{\mathcal{R}}_{irf}\big(\bar{\Psi}_T(\iota,0);h,q\big) = \alpha^2\big\|\mu(irf,h)-\mu(\iota,0)\big\|^2_{W\otimes qq'} + \operatorname{tr}\Big\{(W\otimes qq')V(\iota,0,0)\Big\}. \tag{38}$$

In regard to a comparison of MLE and LFE, we know from our previous analysis that $V(mle,0,0) < V(lfe,0,0)$. Thus, a necessary condition for LPs to be preferable to VAR-based IRFs is

$$\big\|\mu(irf,h)-\mu(lfe,0)\big\|^2_{W\otimes qq'} < \big\|\mu(irf,h)-\mu(mle,0)\big\|^2_{W\otimes qq'}. \tag{39}$$

It can be shown that this inequality does not hold in general—one can construct misspecification MA polynomials $A(L)$ such that $\mu(mle,0)$ is closer to $\mu(irf,h)$ than $\mu(lfe,0)$. For these DGPs, the LP IRF estimates are inferior to VAR estimates for any scale of misspecification α, in an MSE sense.

4.2 Valid URE for the IRF Risk

Given the non-dominance of LPs over VARs, it is natural to employ a method like the PC criterion in Definition 1 to find the preferred IRF estimand, akin to what has been proposed above for point prediction. Observe, however, that the bias term in the IRF risk in (37) differs from the bias term in the forecasting risk in Theorem 2 in the centering. In particular, as a consequence of the local misspecification, the former involves $\mu(irf)$, while the latter involves $\mu(pov)$. Theorem 3 exploits the fact that the unshrunk LFE estimator is centered at $\mu(pov)$ to propose a URE for the prediction risk. Since in general $\mu(pov) \neq \mu(irf)$, the PC criterion in Definition 1 cannot be a valid URE for the IRF risk. More generally, this warns against using methods tailored for forecasting in IRF applications when there are concerns of misspecification.

Limit Distribution of Lag-augmented Estimator. A natural analogue of the forecasting URE proposed above requires finding an estimator centered at $\mu(irf)$. To achieve that, we employ the lag-augmentation proposed by Montiel Olea and Plagborg-Moller (2021) and used in Montiel Olea, Plagborg-Moller, Qian, and Wolf (2024). Recall that the LFE predictor is based on the multi-step regression in (8). To construct a lag-augmented LFE, let $x_t = (y_t, y_{t-1})'$ and define

$$\Big(\hat{\Psi}_T(lalfe) \quad \hat{\gamma}\Big) = \left(\sum_{t=1}^{T-h} y_{t+h}x_t'\right)\left(\sum_{t=1}^{T-h} x_t x_t'\right)^{-1}, \tag{40}$$

where $\hat{\Psi}_T(lalfe)$ is an estimate of the impulse response coefficients of interest, and $\hat{\gamma}$ is a nuisance coefficient due to lag-augmentation. Using the same prior for the subvector of interest as in (10), by the Frisch-Waugh-Lovell theorem we obtain the quasi-posterior

$$\bar{\Psi}_T(lalfe,\tilde{\lambda}) = \left[\tilde{\lambda}\underline{\Psi}_T\underline{P}_\Psi + \hat{\Psi}_T(lalfe)\tilde{S}_{T,hh}\right]\bar{P}_\Psi^{-1}, \quad \bar{P}_\Psi = \tilde{\lambda}\underline{P}_\Psi + \tilde{S}_{T,hh}, \tag{41}$$

where

$$\tilde{S}_{T,hh} = \sum_{t=1}^{T-h}\hat{u}_t\hat{u}_t', \quad \hat{u}_t = y_t - \hat{\Phi}_T(mle)y_{t-1}.$$

Analogous to Theorem 1, we obtain the following result:

Theorem 5 *Under Assumption 1,*

$$\bar{\Psi}_T(lalfe,\lambda) = F^h + T^{-1/2}[\delta(lalfe,\lambda) + \alpha\mu(lalfe,\lambda) + \zeta_T(lalfe,\lambda)] + o_p(T^{-1/2}) \tag{42}$$

for any $\lambda \geq 0$, where

$$\begin{aligned}\delta(lalfe,\lambda) &= \lambda\underline{\psi}\underline{P}_\Psi\left(\lambda\underline{P}_\Psi + \Sigma_{\epsilon\epsilon}\right)^{-1}\\ \mu(lalfe,\lambda) &= \left(\sum_{j=0}^{h-1}F^jA_{h-j}\Sigma_{\epsilon\epsilon}\right)\left(\lambda\underline{P}_\Psi + \Sigma_{\epsilon\epsilon}\right)^{-1}.\end{aligned}$$

Moreover, for $\iota \in \{lfe, mle, lalfe\}, \lambda \geq 0$, $\zeta_T(\iota,\lambda)$ converges weakly to a centered Gaussian process $\zeta(\iota,\lambda)$ with covariance function $\mathcal{V}(\iota,\iota',\lambda,\lambda')$.

A formula for the asymptotic variance $\mathcal{V}(\iota,\iota',\lambda,\lambda')$ is provided in the Online Appendix. A key takeaway from Theorem 5 is that in the absence of shrinkage ($\lambda = 0$) the limit distribution is correctly centered at $\mu(irf)$.

Unbiased Risk Estimation. As in Section 3, we employ Theorem 5 to construct a URE for the asymptotic IRF risk that serves as a criterion to determine the degree of shrinkage and type of IRF estimand. According to Theorem 5, for $\iota \in \{mle, lalfe\}, \lambda \geq 0$,

$$\begin{aligned}&\sqrt{T}\left(\bar{\Psi}_T(\iota,\lambda) - \bar{\Psi}_T(lalfe,0)\right) \\ &= \sqrt{T}\left(\bar{\Psi}_T(\iota,\lambda) - F^h\right) - \sqrt{T}\left(\bar{\Psi}_T(lalfe,0) - F^h\right)\\ &\Longrightarrow N(\delta(\iota,\lambda) + \alpha(\mu(\iota,\lambda) - \mu(irf)), \mathcal{V}(\iota,\lambda) + \mathcal{V}(lalfe,0) - 2\mathcal{V}(\iota,lalfe,\lambda,0)).\end{aligned} \tag{43}$$

We deduce that the asymptotic IRF estimation risk is given by

$$\begin{aligned}&T\mathbb{E}\left[\left\|\bar{\Psi}_T(\iota,\lambda) - \bar{\Psi}_T(lalfe,0)\right\|^2\right] \\ &= \left\|\delta(\iota,\lambda) + \alpha(\mu(\iota,\lambda) - \mu(irf))\right\|^2 + \mathrm{tr}\left[\mathcal{V}(\iota,\lambda) + \mathcal{V}(lalfe,0) - 2\mathcal{V}(\iota,lalfe,\lambda,0)\right] + o(1).\end{aligned} \tag{44}$$

The previous discussion, together with (37), suggests that the following criterion is URE for the asymptotic IRF estimation risk (up to a term that does not depend on (ι, λ)).

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

$$\widetilde{PC}_T(\iota, \lambda) = T \left\| \bar{\Psi}_T(\iota, \lambda) - \bar{\Psi}_T(lalfe, 0) \right\|^2_{W \otimes qq'} + 2\hat{\mathcal{R}}_{\mathcal{V}}(\iota, lalfe, \lambda, 0),$$

where $\hat{\mathcal{R}}_{\mathcal{V}}(\iota, lalfe, \lambda, 0)$ *is a consistent estimate of* $\operatorname{tr}\{(W \otimes qq')\mathcal{V}(\iota, lalfe, \lambda, 0)\}$.

To summarize, the key modification of the multi-step forecasting problem is the lag-augmentation in the construction of the LFE. Without lag-augmentation, for $\lambda = 0$ the LFE is centered at $\mu(pov)$, which due to the local misspecification is not the correct value for the IRF. In contrast, the lag-augmented LFE is centered at $\mu(irf)$ for $\lambda = 0$, and hence allows us to construct an unbiased criterion for hyperparameter determination.[3]

5 Monte Carlo Experiment

We now conduct a Monte Carlo experiment to assess the finite-sample performance of the MLE and LFE shrinkage predictors. Importantly, we will compare PC-based hyperparameter selection to MDD-based hyperparameter selection. The Monte Carlo design is described in Section 5.1. In Section 5.2 we compare the finite sample risk differentials to the expected value of PC_T. We examine prediction losses obtained with PC- versus MDD-based hyperparameter selection in Section 5.3. Finally, we consider the joint PC-based selection of estimator and hyperparameter in Section 5.4.

5.1 Monte Carlo Design

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

[3]The lag-augmentation technique that is central in this point estimation exercise has proven to be crucial for inference as well; see Montiel Olea and Plagborg-Moller (2021) and Montiel Olea, Plagborg-Moller, Qian, and Wolf (2024).

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 will be 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 (14) in regard 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:

$$\underline{\psi} = \varphi\mu(pov). \tag{45}$$

If $\varphi = 1$ then the prior is centered at the pov. Using a first-order Taylor expansion of $\Phi^h - F^h$, we obtain

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

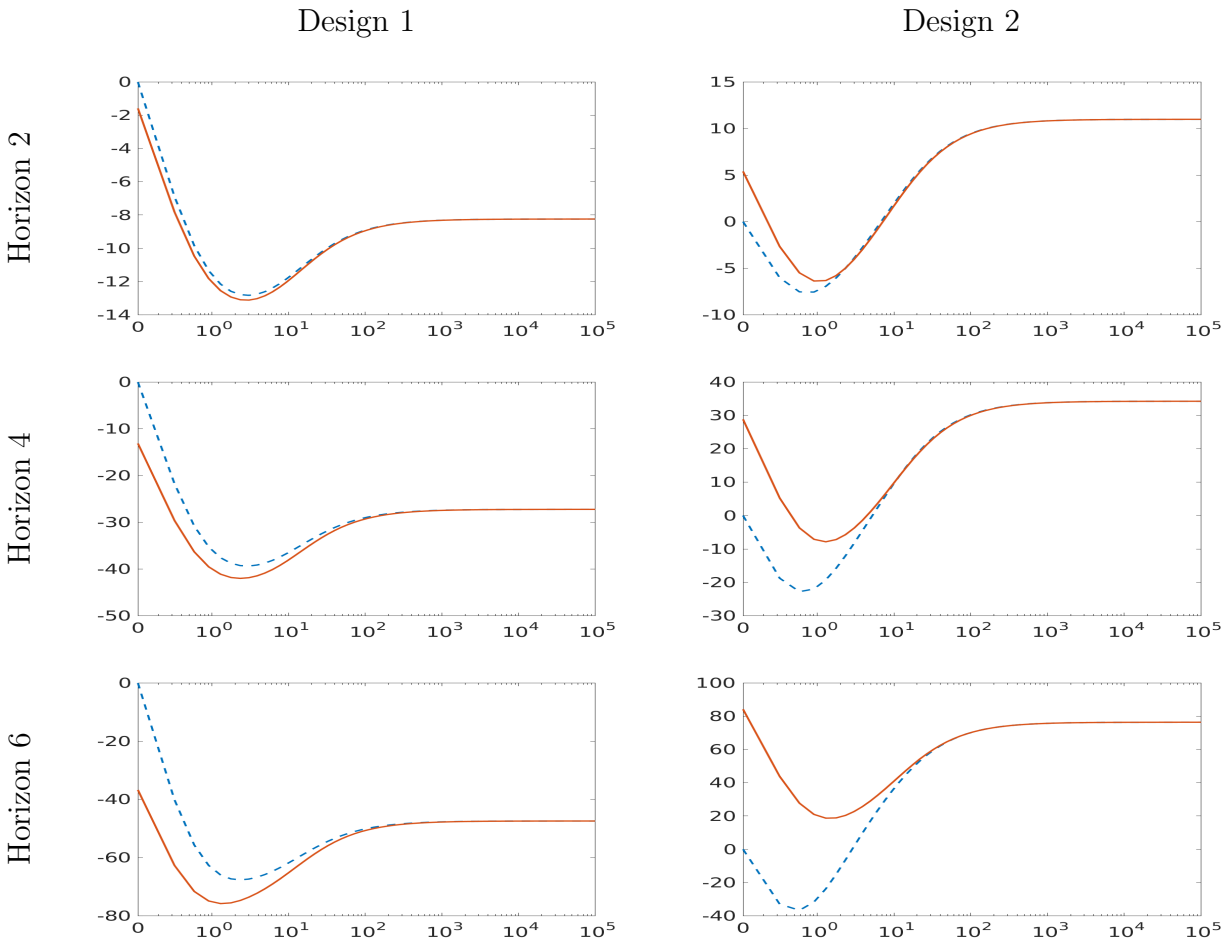
In turn, we choose the (local) prior mean for the MLE such that it satisfies

$$\underline{\psi} = \sum_{j=0}^{h-1} F^j \underline{\phi} F^{h-1-j}. \tag{46}$$

Monte Carlo Designs. We consider two different Monte Carlo designs. Figure 1 depicts the asymptotic risk differentials $\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota,\lambda)) - \bar{\mathcal{R}}(\hat{y}_{T+h}(lfe,0))$ for $\iota \in \{mle, lfe\}$ as a function of λ for the two designs. Under Design 1 there is no misspecification as $\alpha = 0$. By setting $\varphi = 1$ we ensure that the prior is not centered at the "true" value, which in this case would correspond to $\varphi = 0$. The optimal value of λ that minimizes the asymptotic risk $\bar{\mathcal{R}}(\hat{y}_{T+h}(lfe,\lambda))$ is approximately equal to 3 for $h = 2$. Because there is no misspecification, the MLE dominates the LFE. As the horizon increases, the benefit of using the MLE increases and the optimal λ is closer to 2. For $\lambda = \infty$, MLE and LFE are equal to the prior mean values. Because of (46), the resulting predictors are equivalent up to first order and have identical risks, which is clearly visible in Figure 1.

Under Design 2 the VAR is misspecified ($\alpha = 2$) and we center the prior at $\varphi = 0.5$ to keep it away from the pov. In this design the LFE visibly dominates the MLE at all horizons for values for values of λ less than 2. As the precision λ increases further the parameter estimates are dominated by the prior and the risk differential vanishes.

Figure 1: Asymptotic Risk Differentials for MC Designs

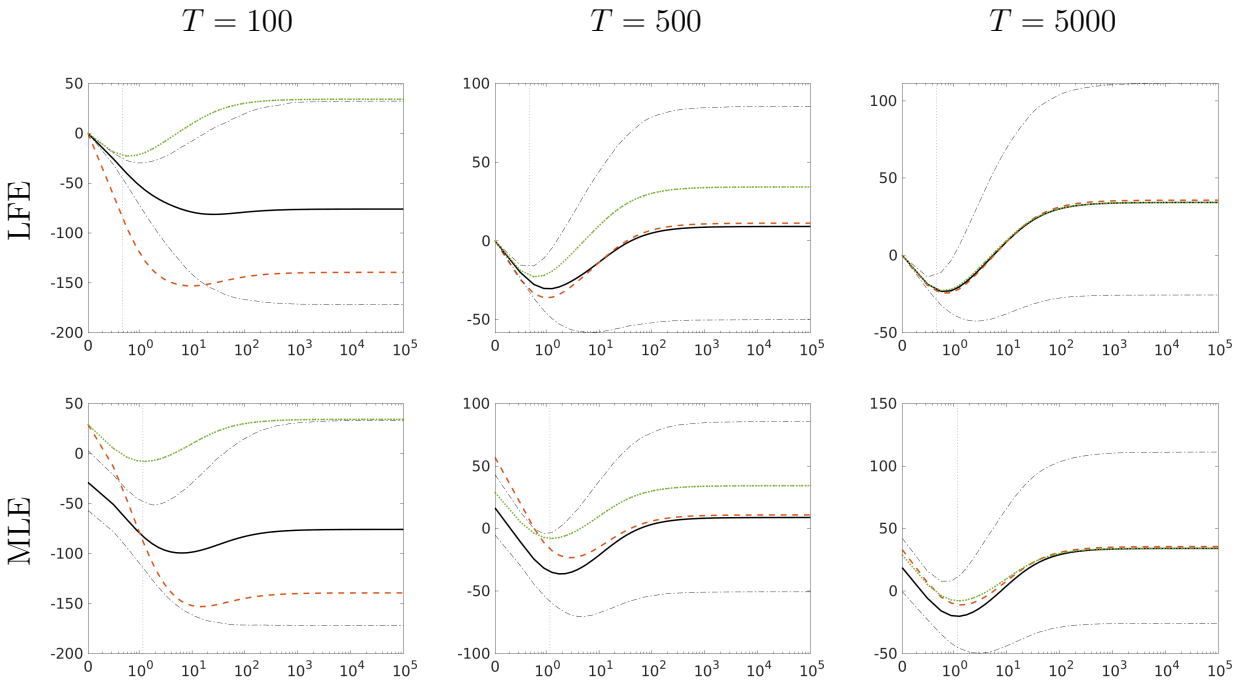


Notes: The x-axis is the hyperparameter λ on a logarithmic scale with zero as the left endpoint. On the y-axis we plot $\bar{\mathcal{R}}(\hat{y}_{T+h}(\iota,\lambda)) - \bar{\mathcal{R}}(\hat{y}_{T+h}(lfe,0))$. The dashed blue line is the LFE and solid orange is the MLE.

5.2 Simulated Risk Differentials and Expected PC

Figure 2 depicts simulated risk differentials and the expected value of PC for Designs 2 (misspecification). The dotted green line is the asymptotic risk. The solid black line is $\mathbb{E}[PC_T(\iota,\lambda)]$. The dashed orange line is the MC risk and the dashed black lines are 90% coverage intervals for the finite sample losses. The vertical line indicates the value of λ that minimizes the asymptotic risk. As the sample size T increases the Monte Carlo risk and $\mathbb{E}[PC_T(\iota,\lambda)]$ converge to the asymptotic risk as predicted by the large-sample result in (24). This illustrates that in large samples PC provides an unbiased estimate of the asymptotic risk.

Figure 2: PC versus Finite Sample Risk, $\alpha = 2$, Horizon $h = 4$



Notes: The x-axis is the hyperparameter λ on a logarithmic scale with zero as the left endpoint. The dotted green line is the asymptotic risk. The solid black line is $\mathbb{E}[PC_T(\iota, \lambda)]$. The dashed orange line is the Monte Carlo risk and the dashed black lines are 90% coverage intervals for the finite sample losses. The vertical line indicates the value of λ that minimizes the asymptotic risk.

5.3 PC versus MDD-Based Hyperparameter Selection

We now examine the Monte Carlo risk of LFE and MLE predictors based on data-driven hyperparameter choices. We consider three different hyperparameter choices: oracle, PC, and MDD based.

Finite-Sample Risk Differentials. Finite sample risk differentials relative to the predictor $\hat{y}_{T+h}(lfe, 0)$ for Design 1 (no misspecification) are reported in Table 1. The entries in the Table can be compared to the asymptotic values plotted in the first column of Figure 1. Large negative numbers indicate substantial improvements over the benchmark predictor $\hat{y}_{T+h}(lfe, 0)$.

For the LFE the PC-based hyperparameter selection leads to lower risk than the MDD-based hyperparameter selection for all sample sizes and forecast horizons considered. This is to be expected because the (quasi) MDD criterion is derived under the incorrect assumption that the multi-step forecast errors are uncorrelated. For the MLE, on the other hand, we find

Table 1: Finite Sample Risk Differentials for $\hat{y}_{T+h}(\iota, \hat{\lambda})$, $\alpha = 0$

	LFE			MLE		
h	Oracle	PC	MDD	Oracle	PC	MDD
Sample Size $T = 100$						
2	-34	-29	-25	-34	-29	-32
4	-91	-74	-61	-93	-76	-82
6	-146	-116	-99	-157	-126	-129
Sample Size $T = 500$						
2	-17	-14	-11	-17	-14	-14
4	-51	-42	-29	-54	-45	-45
6	-88	-64	-51	-99	-78	-75
Sample Size $T = 5,000$						
2	-14	-12	-9	-15	-12	-12
4	-41	-32	-22	-44	-36	-36
6	-76	-52	-41	-84	-65	-63

Notes: The finite sample risk differentials are computed relative to $\hat{y}_{T+h}(lfe, 0)$.

that the MDD-based hyperparameter determination gives the lower prediction risk. Again, this is unsurprising because the forecasting model is correctly specified and the likelihood function generates more efficient estimates. We also tabulate the oracle risk. The oracle risk associated with the MLE is weakly smaller than the LFE. As we discussed before, there is no sense in which the feasible procedures can asymptotically attain the oracle risk, which is why PC and MDD based hyperselection always leads to a greater risk than the oracle procedure.

Table 2 contains results for Design 2 in which the forecast model is misspecified. The major difference relative to Table 1 is that under misspecification the PC-based hyperparameter selection also beats the MDD-based selection for the LFE-based predictor. The risk differentials between PC and MDD λ selection are generally increasing with forecast horizon.

Inspecting the Hyperparameter Selection. We now take a closer look at the PC and MDD objective functions that are used for the hyperparameter determination. In rows 1 and 2 of Figure 3 we plot hairlines of the $PC_T(\iota, \lambda)$ and $MDD(\iota, \lambda)$ objective functions. We normalize the objective functions such that they are equal to zero for $\lambda = \infty$. Each hairline

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

	LFE			MLE		
h	Oracle	PC	MDD	Oracle	PC	MDD
Sample Size $T = 100$						
2	-68	-55	-45	-66	-54	-54
4	-153	-135	-73	-154	-127	-124
6	-297	-284	-114	-308	-256	-226
Sample Size $T = 500$						
2	-15	-13	-10	-12	-10	-7
4	-38	-28	-25	-25	-17	-3
6	-66	-44	-43	-38	-19	24
Sample Size $T = 5,000$						
2	-8	-7	-6	-5	-4	-2
4	-26	-21	-19	-13	-8	-2
6	-42	-25	-33	11	25	45

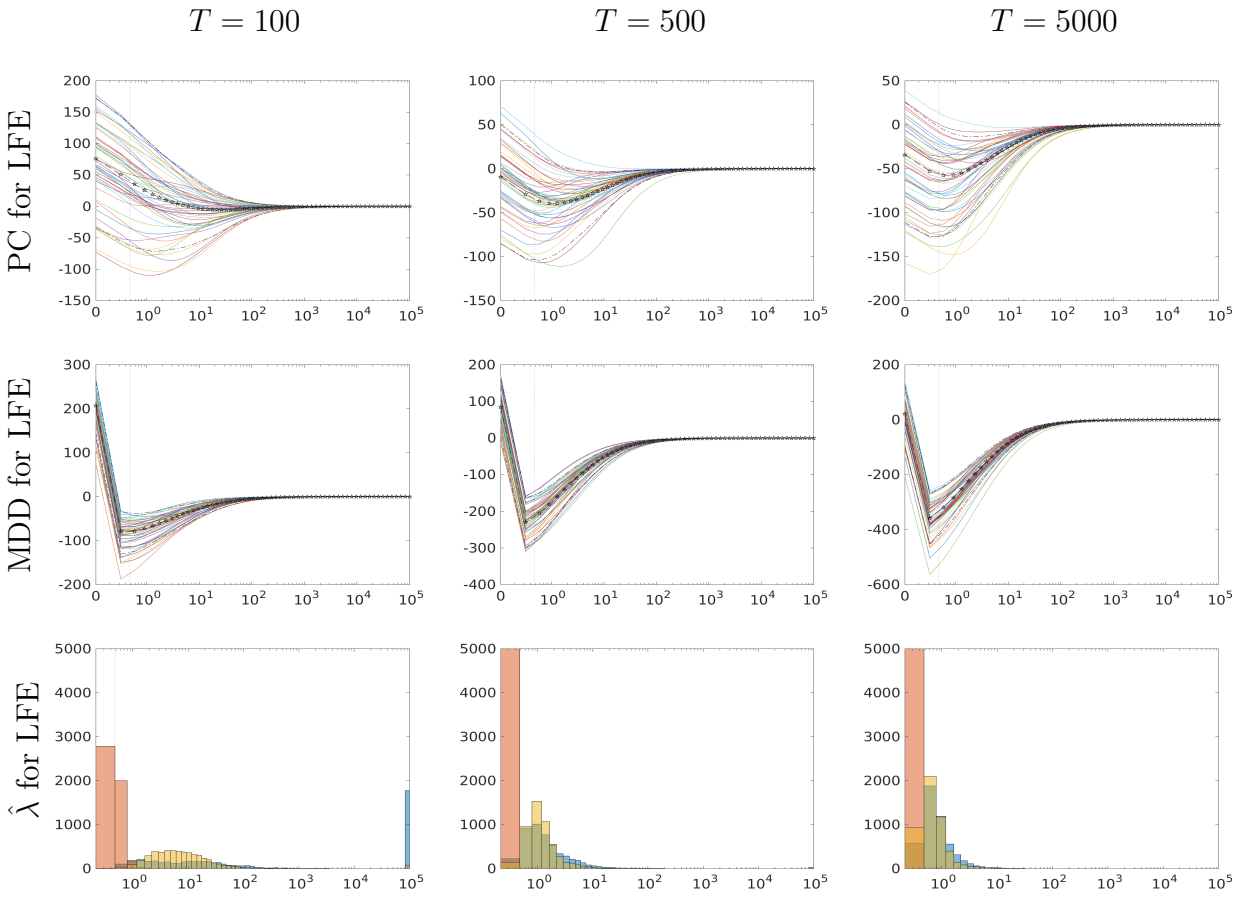
Notes: The finite sample risk differentials are computed relative to $\hat{y}_{T+h}(lfe, 0)$.

corresponds to a Monte Carlo repetition for misspecification $\alpha = 2$ and horizon $h = 4$. The line with the stars is the pointwise expected value of the objective function, obtained by averaging across the Monte Carlo repetitions, and the dashed vertical line indicates the asymptotically optimal value for λ.

For the interpretation of Figure 3 it is instructive to inspect the $\alpha = 2$ (Design 2) and $h = 4$ panel of Figure 1. 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. Returning to Figure 3, for $T = 500$ and $T = 1,000$ most of the hairlines in row 1 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 decreasing over the entire domain of λ. Overall, the hairline pattern is broadly consistent with the asymptotic risk differential function.

The hairlines in row 2 of Figure 3 depict the MDD selection criterion. There is less variation across Monte Carlo repetitions in terms of the overall shape of the hairlines and the minima are more concentrated in a fairly narrow interval ranging from 0.1 to 1. In

Figure 3: PC versus MDD Objective Function, $\alpha = 2$ and $h = 4$

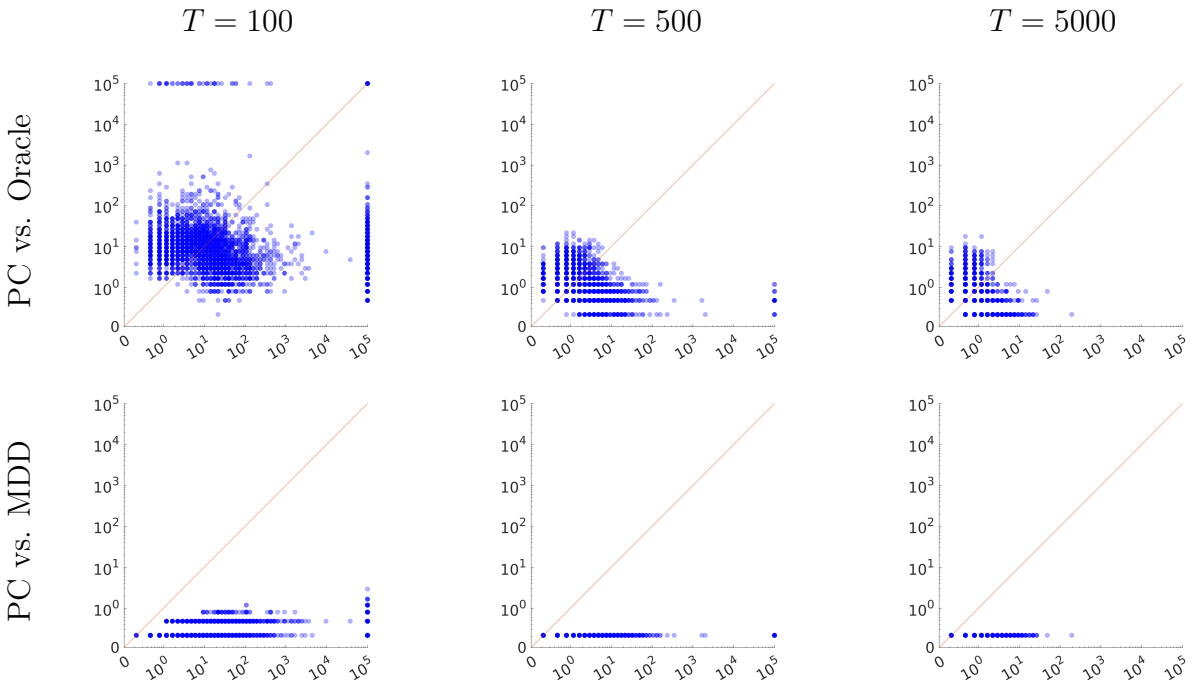


Notes: Rows 1, and 2: hairlines of PC and MDD criteria as a function of λ. Each hairline corresponds to a Monte Carlo repetition. The line with the stars is the pointwise expected value of the objective function and the dashed vertical line indicates the asymptotically optimal value for λ. Row 3: distribution of the optimally selected hyperparameter λ across Monte Carlo repetitions. Histogram colors: yellow is oracle, blue is PC, and orange is MDD.

general, the MDD minimum is to the left of the PC minimum, meaning that there is less shrinkage toward the prior mean.

In the last row of Figure 3 we plot histograms of the $\hat{\lambda}$ distribution across Monte Carlo repetitions for the PC, MDD, and the oracle objective functions, respectively. For PC most of the mass concentrates near the argmin of the asymptotic risk function, but the distribution is skewed to the right with a small pointmass at $\lambda = \infty$ which vanishes as the sample size T increases. The flatness of the asymptotic risk function between $\lambda = 0.5$ to 5 translates into a fairly diffuse distribution of $\hat{\lambda}$ over this range. There is essentially no mass for intermediate values of $\hat{\lambda}$ and a small point mass at the maximum value of the λ grid. This point mass

Figure 4: LFE Hyperparameter Selection, $\alpha = 2$, Horizon $h = 4$



Notes: We use log scales for the hyperparameter λ. The solid line is the 45 degree line. Each dot corresponds to a Monte Carlo repetition. PC is always on the x-axis, and the predictor to be compared on the y-axis.

corresponds to Monte Carlo repetitions for which the objective function is monotonically decreasing. The $\hat{\lambda}$ histograms obtained from the oracle looks similar to that of the PC. For the MDD criterion, on the other hand, there is a large mass near zero.

The panels of Figure 4 show scatter plots of $\hat{\lambda}$ based on the three objective functions: PC, MDD, and Oracle. Each dot corresponds to a Monte Carlo repetition. The first row shows correlation between the PC and Oracle $\hat{\lambda}$. Because the PC objective function remains random in the limit and does not converge to the oracle objective function there is hardly any correlation between the two estimates of λ, even though the marginal distributions of $\hat{\lambda}$ are quite similar. The second row compares PC and MDD $\hat{\lambda}$s. The key result here is that the MDD criterion generates less shrinkage than the PC criterion.

5.4 Joint Selection of Estimator and Hyperparameter

One can also use PC to determine the estimator type and the shrinkage parameter simultaneously by minimizing $PC_T(\iota, \lambda)$ jointly with respect to (ι, λ). The results are summarized

Table 3: Finite Sample Risk Differentials for $\hat{y}_{T+h}(\iota, \hat{\lambda})$, Joint (ι, λ) Selection

	$\alpha = 0$			$\alpha = 2$		
h	LFE	MLE	Joint	LFE	MLE	Joint
Sample Size $T = 100$						
2	-29	-29	-29	-55	-54	-54
4	-74	-76	-75	-135	-127	-127
6	-116	-126	-126	-284	-256	-255
Sample Size $T = 500$						
2	-14	-14	-14	-13	-10	-11
4	-42	-45	-45	-28	-17	-20
6	-64	-78	-78	-44	-19	-28
Sample Size $T = 5,000$						
2	-12	-12	-12	-7	-4	-5
4	-32	-36	-36	-21	-8	-14
6	-52	-65	-65	-25	25	-9

Notes: The finite sample risk differentials are computed relative to $\hat{y}_{T+h}(lfe, 0)$.

in Table 3. In the LFE and MLE columns we reproduce the numbers from Tables 1 and 2. The columns labeled "Joint" contain the additional risk numbers obtained by using PC to determine ι and λ jointly. For $\alpha = 0$ the risk differentials between the LFE and MLE predictors are fairly small, albeit increasing in forecast horizon h. Here the joint selection generally attains the performance of the better among the two predictors, which tends to be the MLE predictor in the absence of misspecification. For the $\alpha = 2$ case and $T = 100$ or $T = 5,000$ the performance of the joint selection is typically somewhere in between the performance of the LFE and MLE predictor.

6 Multiple Lags and Lag Length Selection

Companion Form. So far, we considered a VAR in (1) with a single lag. It turns out, that the formulas we derived also cover the case of multiple lags, because they can be interpreted

as companion form notation. Consider a VAR with q lags:

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_q y_{t-q} + u_t, \quad u_t \sim N(0, \Sigma_{uu}). \tag{47}$$

Define

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

where M is a selection matrix such that $y_t = M'Y_t$. Thus we can express (47) in companion form as

$$Y_t = \Phi Y_{t-1} + U_t, \quad \Sigma_{UU} = M\Sigma_{uu}M', \tag{48}$$

which looks identical to (1), except that we replaced lower case by upper case variables. In addition, we define

$$\underbrace{\phi}_{n\times nq} = [\phi_1, \ldots, \phi_q], \quad \underbrace{\Upsilon}_{n(q-1)\times nq} = \begin{bmatrix} I_n & \cdots & 0_n & 0_n \\ \vdots & \ddots & \vdots & \vdots \\ 0_n & \cdots & I_n & 0_n \end{bmatrix}, \quad \underbrace{M_\Upsilon}_{nq\times(n-1)q} = \begin{bmatrix} 0_n & \cdots & 0_n \\ I_n & \cdots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \cdots & I_n \end{bmatrix} \tag{49}$$

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

$$M'\Phi = \phi, \quad M_\Upsilon'\Phi = \Upsilon. \tag{50}$$

We assume that the prior mean used to construct the shrinkage estimator shares the companion form structure and can be written as

$$\underline{\Phi}_T = \begin{bmatrix} \underline{\phi}_T \\ \Upsilon \end{bmatrix} \tag{51}$$

such that by construction the properties in (50) are satisfied. Now define

$$\bar{S}_{T,01} = \sum_{t=1}^{T} Y_t Y_{t-1}' + \lambda \underline{\Phi}_T \underline{P}_\phi, \quad \bar{S}_{T,11} = \sum_{t=1}^{T} Y_{t-1} Y_{t-1}' + \lambda \underline{P}_\phi \tag{52}$$

such that the posterior mean in companion form can be expressed as

$$\bar{\Phi}_T = \bar{S}_{T,01}\bar{S}_{T,11}^{-1}, \tag{53}$$

which is identical to the formula provided in (5). Without companion form notation, a direct calculation of the posterior mean of ϕ would yield

$$\bar{\phi}_T = \left(\sum_{t=1}^{T} y_t Y_{t-1}' + \lambda \underline{\phi}_T \underline{P}_\phi\right)\left(\sum_{t=1}^{T} Y_{t-1} Y_{t-1}' + \lambda \underline{P}_\phi\right)^{-1}. \tag{54}$$

It is straightforward to verify that $M'\bar{\Phi}_T = \bar{\phi}_T$ and $M_\Upsilon'\bar{\Phi}_T = \Upsilon_T$. Thus, the companion form posterior mean also satisfies (50). We conclude that the formulas derived in Sections 2 and (3) also apply to the VAR(q).

Lag Length Selection. S2005 also considered the problem of lag length selection. To develop a concise notation, the VAR is written in q-companion form, and estimated subject to the restriction that only coefficients for lags 1 through $p \leq q$ are non-zero. The "true" asymptotic lag order of the DGP is assumed to be $p_* \leq q$. While we refer the reader for a detailed discussion of lag length selection to S2005, we briefly show how the shrinkage estimator for a VAR(p) could be expressed in q companion form. Φ has the restricted form

$$\underbrace{\Phi}_{nq\times nq} = \begin{bmatrix} \phi_1 & \cdots & \phi_p & 0_n & \cdots & 0_n & 0_n \\ I_n & \cdots & 0_n & 0_n & \cdots & 0_n & 0_n \\ \vdots & & & & & & \vdots \\ 0_n & \cdots & 0_n & 0_n & \cdots & I_n & 0_n \end{bmatrix}. \tag{55}$$

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

$$\underbrace{R_p}_{nq\times n(q-p)} = \begin{bmatrix} 0_n & \cdots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \cdots & 0_n \\ I_n & \cdots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \cdots & I_n \end{bmatrix}, \quad \underbrace{R_{p\perp}}_{nq\times np} = \begin{bmatrix} I_n & \cdots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \cdots & I_n \\ 0_n & \cdots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \cdots & 0_n \end{bmatrix}. \tag{56}$$

This allows us to express the zero restriction on coefficients associated with lags greater than p as $M'\Phi R_p = 0$. We assume that the prior mean $\bar{\Phi}_T$ has the companion form structure in (51) and satisfies the restriction:

$$M'\bar{\Phi}_T R_p = 0. \tag{57}$$

Using the definitions in (52), one can express

$$\bar{\Phi}_T = \bar{S}_{T,01}\bar{S}_{T,11}^{-1}\left[I_{nq} - R_p(R_p'\bar{S}_{T,11}^{-1}R_p)^{-1}R_p'\bar{S}_{T,11}^{-1}\right]. \tag{58}$$

It is straightforward to verify that $\bar{\Phi}_T R_p = 0$, meaning that the last $n(q-p)$ columns are equal to zero. One can also show that

$$\begin{aligned} &\bar{S}_{T,11}^{-1}\left[I_{nq} - R_p(R_p'\bar{S}_{T,11}^{-1}R_p)^{-1}R_p'\bar{S}_{T,11}^{-1}\right] \\ &= \begin{bmatrix} \left(\sum_{t=1}^T R_{p\perp}'Y_{t-1}Y_{t-1}'R_{p\perp} + \lambda R_{p\perp}'\underline{P}_\phi R_{p\perp}\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \end{aligned} \tag{59}$$

where the selection matrix $R_{p\perp}$ eliminates lags $p+1,\ldots,q$ from the companion form regressor vector Y_{t-1}. Thus, the first np columns of $M'\bar{\Phi}_T$ are equivalent to the posterior mean that would be obtained using a p-companion form and the remaining columns are zero. Rows $n+1$ to np of $\bar{\Phi}_T$ correspond to the first $(n-1)p$ rows of Υ and the remaining rows of the matrix are irrelevant to our prediction problems. Following the arguments in S2005, one can generalize the formulas derived in Sections 2, 3 and 4 to express predictors derived from VARs of different lag orders in a companion form of fixed order q.

7 Empirical Application

We apply the proposed methodology to the FRED-QD database; see McCracken and Ng (2020). We filter each series using the procedure proposed by Hamilton (2018) to induce stationarity, and drop the series which contain missing values. Many users of the database apply a transformation originally proposed in work by James Stock and Mark Watson, which, for instance, would temporally difference a random walk series and turn it into a serially uncorrelated time series. Thus, the autocorrelation drops from one to zero. The Hamilton filter transforms a random walk into a stationary series, but preserves a lot of the persistence. We prefer to use stationary yet persistent series in our subsequent application, so that they are predictable over multiple horizons.

We follow Marcellino, Stock, and Watson (2006) in that we are creating a large number of data sets by randomly selecting groups of time series from the FRED-QD database. We do so by selecting uniformly at random 200 different six-tuples of series. For each of the data sets we estimate a VAR(1) of dimension $n = 6$ and focus on the selection of the estimator $\iota \in \{mle, lfe\}$ and the degree of shrinkage λ. Because the series are chosen at random, this implies that for some data sets there is more, for others less misspecification.

We demean and standardize each series, and thus center our prior at zero. The estimation sample size is set to $T = 100$, and we consider forecast horizons $h = 2, 4, 6$. The goal is to

generate a mean square errors (MSE) analysis akin to the Monte Carlo section. We do so as follows. Suppose we have raw data $y_1, \ldots, y_{T_*}$, where T_* is the total number of observations for the vector y_t in the FRED-QD database. We use rolling samples, denoted by $\tau = 1, \ldots, \tau_*$, to keep the relative weight on prior and likelihood constant. The recursive samples are

$$y_\tau, \ldots, y_{(\tau-1)+T}, \ldots, y_{(\tau-1)+T+h},$$

starting from $\tau = 1$ and ending at $\tau = T_* - T - h + 1$. Let

$$\hat{y}_{(\tau-1+T)+h}$$

be the forecast for the observation in period $(\tau - 1 + T) + h$ given the information at the forecast origin $(\tau - 1) + T$. Recall that the MLE and LFE-based shrinkage predictors were defined in (7) and (12), respectively. We compute

$$\widehat{MSE}(\hat{y}; T, h) = \frac{1}{T_* - T - h + 1} \sum_{\tau=1}^{T_*-T-h+1} L\big(y_{(\tau-1+T)+h}, \hat{y}_{(\tau-1+T)+h}\big), \tag{60}$$

where the quadratic loss $L(\cdot)$ was defined in (2). We set the weight matrix W in the loss function equal to the identity matrix because the time series have been standardized. Below we report percentage changes of $\widehat{MSE}(\hat{y}; T, h)$ relative to the MSE associated with the predictor that uses PC to jointly select the estimator ι and the degree of shrinkage λ (baseline predictor).

Figure 5 reports the distribution of the relative MSEs (top) and optimally selected shrinkage levels (bottom) across the randomly selected data sets. The plots in the left column are based on the LFE-based shrinkage predictor and the panels in the right column are generated from the MLE-based shrinkage predictor. According to the normalization of $\widehat{MSE}(\hat{y}; T, h)$, positive values indicate that the performance of the predictor under consideration is worse than the baseline predictor.

The top left panel of the figure implies that always using the LFE-based shrinkage estimator and using PC to select λ, tends to lead to MSE improvements (negative values), relative to the baseline predictor. If λ is selected by the MDD criterion, then much of the mass in the histogram shifts to the left of zero, meaning MDD selection of λ leads to worse performance than the baseline predictor.

For the MLE in the top right panel the situation is reversed. However, the mass to the left of zero seems to be always concentrated close to zero, meaning that the joint selection is very close to the best shrinkage selection when fixing the estimator. Regarding the optimal

Figure 5: Distribution of Relative MSEs and Optimal PC Shrinkage

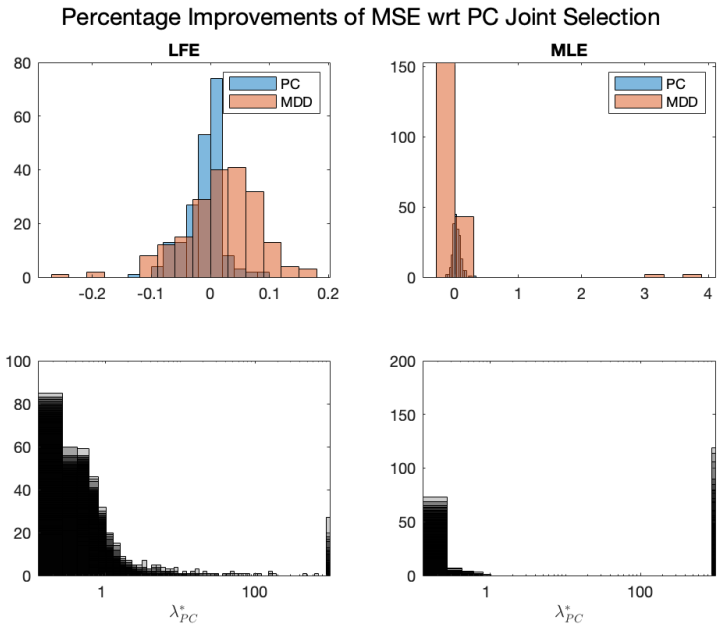
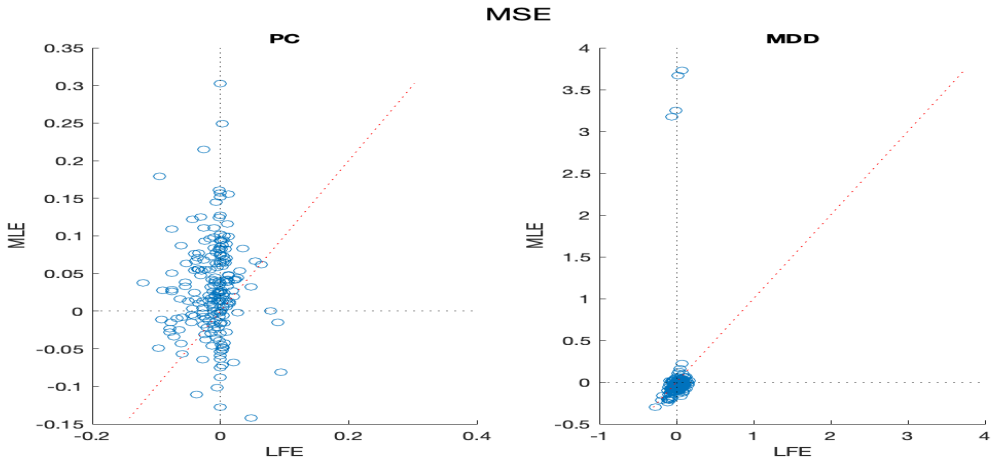


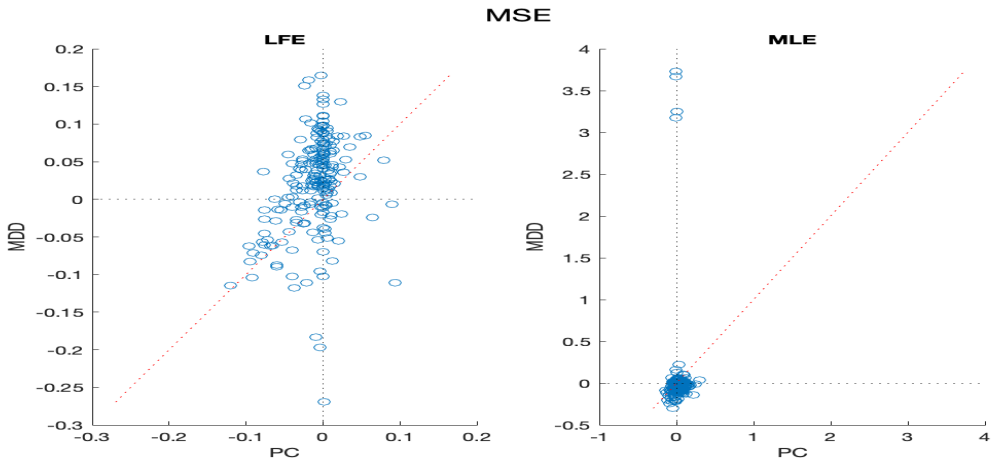
Figure 6: Scatter of MSEs for a fixed selection mechanism



shrinkage level, we see that in the LFE case the shrinkage level tends to be interior, while it falls on the extremes for the MLE. All this is in line with the MC results.

Figure 6 depicts MSE scatter plots. The left panel is generated by selecting λ using PC and the right panel contains results for MDD hyperparameter selection. Each dot in the

Figure 7: Scatter of MSEs for a fixed estimator



scatter plot represents a data set and corresponds to a pair of MSEs, associated with the MLE and the LFE-based shrinkage predictors. A dot falling into the Northeast quadrant means that MLE and LFE with shrinkage chosen by the corresponding mechanism are worse than joint PC selection. The Northwest quadrant contains data sets for which the MLE is worse than joint PC selection, but LFE is better (hence LFE with PC is better than MLE with PC). The Southwest quadrant means that both MLE and LFE are better than joint PC selection. Finally, the 4th quadrant contains data sets for which the LFE is worse than joint PC selection, but MLE is better. Overall, it is hard to see big patterns for the PC by eyeballing, it seems evenly distributed around zero in all directions. For the MDD, clearly the MLE has much less dispersion, although both distributions seem to be centered around zero (maybe MLE slightly below). The 45 degree line helps to identify the relative performance between LFE and MLE.

Figure 7 is similar to Figure 6 but now fixing the estimator instead of the selection mechanism – hence we are basically comparing the performance of the selection mechanism here. The Northeast quadrant means that PC and MDD selection mechanisms for the corresponding estimator are worse than joint PC selection. The Northwest quadrant means that MDD is worse than joint PC selection, but PC is better (hence PC is better than MDD). The Southwest quadrant means that both PC and MDD are better than joint PC selection. The Southeast quadrant means that PC is worse than joint PC selection, but MDD is better. Again, same observation about the patterns as before. The 45 degree line helps to identify the relative performance between PC and MDD. For the LFE, the majority

of points is above the 45 degree line, meaning that PC hyperparameter selection leads to more accurate predictions than MDD-based selection. For the MLE, the results seems to be reversed.

8 Conclusion

We consider a framework in which a researcher uses a VAR that is dynamically misspecified. However, the misspecification is assumed to be fairly small, to that it cannot be easily detected. We capture this notion in an asymptotic framework in which the misspecification vanishes at the same rate at which model parameters can be estimated. We consider two applications for the VAR: multi-step forecasting and impulse response function estimation and develop criteria that provided asymptotically unbiased, but inconsistent estimates of the prediction risk or the impulse response estimation risk. We show how these criteria can be used for the hyperparameter determination in a quasi Bayesian setting that shrinks the a MLE or a LFE toward a prior mean.

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.

Jorda, O. (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-Moller, 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.

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.

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

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

Plagborg-Moller, 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: VAR Hyperparameter Determination Under Misspecification

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 Auxiliary Results

Theorem A-1 *Under Assumption 1, the process* $\sup_{\lambda\geq 0}\|\zeta_t(\iota,\lambda)\|^2$, $\iota\in\{lfe,mle\}$, *is uniformly integrable.*

Proof of Theorem A-1. Throughout, we let $\langle f\rangle_l=\left(\mathbb{E}[|f|^l]\right)^{1/l}$. The proof consists of two parts. We first bound the norms $\langle\sup_{\lambda\geq 0}\|\zeta_T(\iota,\lambda)\|\rangle_{2+\delta}$ for $\iota\in\{lfe,mle\}$, with moments of

$$\frac{1}{\sqrt{T}}\sum\epsilon_{t-j}y'_{t-h},\quad \left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1},\quad \frac{1}{T}\sum z_{t-j}y'_{t-h}-\Gamma_{zy,h-j} \tag{A.1}$$

Next, we apply Theorems 4 and 5 in S2005.

Loss-function-based Estimator. By definition

$$\begin{aligned}\zeta_T(lfe,\lambda)=&\sum_{j=0}^{h-1}F^j\left(\frac{1}{\sqrt{T}}\sum\epsilon_{t-j}y'_{t-h}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\\&+\quad\alpha\left[\sum_{j=0}^{h-1}F^j\left(\frac{1}{T}\sum z_{t-j}y'_{t-h}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}-\mu(lfe,\lambda)\right].\end{aligned}$$

By Minkowski's inequality,

$$\begin{aligned}\left\langle\sup_{\lambda\geq 0}\|\zeta_T(lfe,\lambda)\|\right\rangle_{2+\delta}\leq&\sum_{j=0}^{h-1}\|F^j\|\left\langle\sup_{\lambda\geq 0}\left\|\left(\frac{1}{\sqrt{T}}\sum\epsilon_{t-j}y'_{t-h}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{2+\delta}\\&+\alpha\sum_{j=0}^{h-1}\|F^j\|\left\langle\sup_{\lambda\geq 0}\left\|\left(\frac{1}{T}\sum z_{t-j}y'_{t-h}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right.\right.\\&\left.\left.-\Gamma_{zy,h-j}(\lambda\underline{P}_\Psi+\Gamma_{yy,0})^{-1}\right\|\right\rangle_{2+\delta},\end{aligned}$$

Let us first bound the first term in the previous expression. It is easy to see that

$$\begin{aligned}&\left\langle\sup_{\lambda\geq 0}\left\|\left(\frac{1}{\sqrt{T}}\sum\epsilon_{t-j}y'_{t-h}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{2+\delta}\\&\quad\leq\left\langle\left\|\left(\frac{1}{\sqrt{T}}\sum\epsilon_{t-j}y'_{t-h}\right)\left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{2+\delta}\\&\quad\leq\left[\left\langle\left\|\frac{1}{\sqrt{T}}\sum\epsilon_{t-j}y'_{t-h}\right\|\right\rangle_{q_1(2+\delta)}\left\langle\left\|\left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{\frac{q_1(2+\delta)}{q_1-1}}\right]^{\frac{q_1-1}{q_1}}\end{aligned}$$

for $q_1 > 1$, where the last inequality follows by Hölder's inequality.

The second term can be bounded as follows:

$$\begin{aligned}
&\left\langle \sup_{\lambda\geq 0}\left\| \left(\frac{1}{T}\sum z_{t-j}y'_{t-h}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}-\Gamma_{zy,h-j}(\lambda\underline{P}_\Psi+\Gamma_{yy,0})^{-1}\right\|\right\rangle_{2+\delta}\\
&\leq \left\langle \sup_{\lambda\geq 0}\left\|\left(\frac{1}{T}\sum z_{t-j}y'_{t-h}-\Gamma_{zy,h-j}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{2+\delta}\\
&\quad \left\langle \sup_{\lambda\geq 0}\left\|\Gamma_{zy,h-j}\left[\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}-(\lambda\underline{P}_\Psi+\Gamma_{yy,0})^{-1}\right]\right\|\right\rangle_{2+\delta}\\
&\leq \left\langle \left\|\left(\frac{1}{T}\sum z_{t-j}y'_{t-h}-\Gamma_{zy,h-j}\right)\left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{2+\delta}\\
&\quad \|\Gamma_{zy,h-j}\|\left(\left\langle \left\|\left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{2+\delta}+\|\Gamma^{-1}_{yy,0}\|\right)\\
&\leq \left[\left\langle \left\|\left(\frac{1}{T}\sum z_{t-j}y'_{t-h}-\Gamma_{zy,h-j}\right)\right\|\right\rangle_{q_2(2+\delta)}\left\langle \left\|\left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{\frac{q_2(2+\delta)}{q_2-1}}\right]^{\frac{q_2-1}{q_2}}\\
&\quad \|\Gamma_{zy,h-j}\|\left(\left\langle \left\|\left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{2+\delta}+\|\Gamma^{-1}_{yy,0}\|\right)
\end{aligned}$$

for some $q_2 > 1$. Thus, altogether, a sufficient condition the uniform integrability of $\sup_{\lambda\geq 0}\|\zeta_T(lfe,\lambda)\|^2$ is

$$\begin{aligned}
&\sup_{t\geq T^*(h)}\left\langle \left\|\frac{1}{\sqrt{T}}\sum \epsilon_{t-j}y'_{t-h}\right\|\right\rangle_{q_1(2+\delta)}<\infty\\
&\sup_{t\geq T^*(h)}\left\langle \left\|\left(\frac{1}{T}\sum y_{t-h}y'_{t-h}\right)^{-1}\right\|\right\rangle_{(2+\delta)\min\left\{\frac{q_1}{q_1-1},\frac{q_2}{q_2-1}\right\}}<\infty\\
&\sup_{t\geq T^*(h)}\left\langle \left\|\frac{1}{T}\sum z_{t-j}y'_{t-h}-\Gamma_{zy,h-j}\right\|\right\rangle_{q_2(2+\delta)}<\infty
\end{aligned}$$

for some $q_1, q_2 > 1$ and $T^*(h)$.

Maximum Likelihood Estimator. By definition

$$\begin{aligned}
\zeta_T(mle,\lambda) &= \sum_{j=0}^{h-1}F^j\left(\frac{1}{\sqrt{T}}\sum \epsilon_t y'_{t-1}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-1}y'_{t-1}\right)^{-1}F^{h-1-j}\\
&\quad +\alpha\left[\sum_{j=0}^{h-1}F^j\left(\frac{1}{T}\sum z_t y'_{t-1}\right)\left(\lambda\underline{P}_\Psi+\frac{1}{T}\sum y_{t-1}y'_{t-1}\right)^{-1}F^{h-1-j}-\mu(lfe,\lambda)\right]\\
&\quad +\sqrt{T}\mathcal{R}(\bar{\Phi}_T(mle,\lambda)-F).
\end{aligned}$$

By Minkowski's inequality,

$$\begin{aligned}\left\langle \sup_{\lambda\geq 0}\|\zeta_T(mle,\lambda)\|\right\rangle_{2+\delta} \leq{}& (h-1)\|F^{h-1}\|\left\langle \sup_{\lambda\geq 0}\left\|\left(\frac{1}{\sqrt{T}}\sum \epsilon_t y_{t-1}'\right)\left(\lambda \underline{P}_\Psi + \frac{1}{T}\sum y_{t-1}y_{t-1}'\right)^{-1}\right\|\right\rangle_{2+\delta}\\ &+\alpha(h-1)\|F^{h-1}\|\Bigg\langle\Bigg\|\left(\frac{1}{T}\sum z_t y_{t-1}'\right)\left(\lambda \underline{P}_\Psi + \frac{1}{T}\sum y_{t-1}y_{t-1}'\right)^{-1}\\ &-\Gamma_{zy,1}(\lambda \underline{P}_\Psi + \Gamma_{yy,0})^{-1}\Bigg\|\Bigg\rangle_{2+\delta}\\ &+\sqrt{T}\left\langle \sup_{\lambda\geq 0}\left\|\mathcal{R}(\bar{\Phi}_T(mle,\lambda)-F)\right\|\right\rangle_{2+\delta}.\end{aligned}$$

Since the first two terms are equivalent to the terms that arise in an $h=1$-step ahead LFE predictor, we now focus on the remainder term. From an h-order Taylor series expansion of Φ^h around F, deduce

$$\|\mathcal{R}(\Phi - F)\| \leq C_h(F)\sum_{j=2}^{h}\|\Phi - F\|^j$$

for some constant $C_h(F)$ that depends on the forecast horizon h and the autoregressive matrix F. Thus,

$$\sqrt{T}\left\langle \sup_{\lambda\geq 0}\left\|\mathcal{R}(\bar{\Phi}_T(mle,\lambda)-F)\right\|\right\rangle_{2+\delta} \leq |C_h(F)|\sum_{j=2}^{h}T^{-(j+1)/2}\left\langle \sup_{\lambda\geq 0}\left\|\sqrt{T}(\bar{\Phi}_T(mle,\lambda)-F)\right\|\right\rangle_{2+\delta}$$

Therefore, a sufficient condition for the uniform integrability of $\sup_{\lambda\geq 0}\|\zeta_T(mle,\lambda)\|^2$ is

$$\begin{aligned}&\sup_{t\geq T^*(h)}\left\langle\left\|\frac{1}{\sqrt{T}}\sum \epsilon_{t-j}y_{t-h}'\right\|\right\rangle_{q_3 h(2+\delta)} < \infty\\ &\sup_{t\geq T^*(h)}\left\langle\left\|\left(\frac{1}{T}\sum y_{t-h}y_{t-h}'\right)^{-1}\right\|\right\rangle_{(2+\delta)h\min\left\{\frac{q_3}{q_3-1},\frac{q_4}{q_4-1}\right\}} < \infty\\ &\sup_{t\geq T^*(h)}\left\langle\left\|\frac{1}{T}\sum z_{t-j}y_{t-h}' - \Gamma_{zy,h-j}\right\|\right\rangle_{q_4 h(2+\delta)} < \infty\end{aligned}$$

for some $q_3, q_4 > 1$ and $T^*(h)$.

Finally, note that the sufficient conditions are the same as those derived in Theorem 6 in S2005, hence per the same arguments as in his proof we can invoke his Theorems 4 and 5 and the result follows. ■

A.2 Proofs for Section 2

Proof of Theorem 1. The asymptotic covariance matrix takes the form

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

with the elements defined as

$$V(lfe,\lambda,\lambda') = \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i\Sigma_{\epsilon\epsilon}F^{j'}) \otimes ((\lambda\underline{P}'_\Psi + \Gamma_{yy,0})^{-1}\Gamma_{yy,j-i}(\lambda'\underline{P}_\Psi + \Gamma_{yy,0})^{-1})$$

$$V(mle,\lambda,\lambda') = \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}(F^i\Sigma_{\epsilon\epsilon}F^{j'}) \otimes (F^{h-1-i'}(\lambda\underline{P}'_\Phi + \Gamma_{yy,0})^{-1}\Gamma_{yy,0}(\lambda'\underline{P}_\Phi + \Gamma_{yy,0})^{-1}F^{h-1-j})$$

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

Because $\Gamma_{yy,h-1-j} = F^{h-1-j}\Gamma_{yy,0}$, we obtain that

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

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:

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

Since

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

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

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

with the understanding that on the right-hand side of the equation $\zeta_T(\iota,\lambda)$ is vectorized. Under the conditions in Schorfheide (2005), the sequence $\|\zeta_T(\iota,p)\|^2$ is uniformly integrable. Hence, we can deduce that (see Theorem 3.5 of Billingsley (1968))

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

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.2) 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.2) are $o(1)$, too. The above simplifications allow us to rewrite the normalized prediction risk as

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

Hence, the desired result follows. ■

A.3 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

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

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

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

Working out the expected values according to Theorem 1 yields

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

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

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

Proof of Theorem 4. According to Theorem 1 we can use

$$\bar{\Psi}_T(\iota,\lambda) = F^h + T^{-1/2}\big[\delta(\iota,\lambda) + \alpha\mu(\iota,\lambda) + \zeta_T(\iota,\lambda)\big] + o_p(T^{-1/2}) \tag{A.3}$$

to replace $\hat{y}_{T+h}(\iota,\lambda)$ by $\bar{\Psi}_T(\iota,\lambda)\tilde{y}_T$. Using the expression of $\mathbb{E}_T[\tilde{y}_{T+h}]$ from the proof of Theorem 2 we obtain

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

In turn

$$\begin{aligned}
&T\mathcal{L}\big(\mathbb{E}_T[\tilde{y}_{T+h}], \hat{y}_{T+h}(\iota,\lambda)\big) \\
&=_{(1)} \alpha^2 tr\left\{W\left(\sum_{j=0}^{h-1} F^j \mathbb{E}_T[\tilde{z}_{T+h-j}] - \mu(pov)\tilde{y}_T\right)\left(\sum_{j=0}^{h-1} F^j \mathbb{E}_T[\tilde{z}_{T+h-j}] - \mu(pov)\tilde{y}_T\right)'\right\} \\
&\quad_{(2)} +tr\left\{W\big[\delta(\iota,\lambda) + \alpha\big(\mu(\iota,\lambda) - \mu(pov)\big)\big]\Gamma_{yy,0}\big[\delta(\iota,\lambda) + \alpha\big(\mu(\iota,\lambda) - \mu(pov)\big)\big]'\right\} \\
&\quad_{(3)} +tr\left\{W\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad_{(4)} -2\alpha tr\Bigg\{W\left(\sum_{j=0}^{h-1} F^j \mathbb{E}\big[\mathbb{E}_T[\tilde{z}_{T+h-j}]\tilde{y}_T' - \mu(pov)\tilde{y}_T\tilde{y}_T' \,\big|\, \{y_t\}_{t=1}^T\big]\right) \\
&\qquad \times\big[\delta(\iota,\lambda) + \alpha\big(\mu(\iota,\lambda) - \mu(pov)\big)\big]'\Bigg\} \\
&\quad_{(5)} -2\alpha tr\left\{W\left(\sum_{j=0}^{h-1} F^j \mathbb{E}\big[\mathbb{E}_T[\tilde{z}_{T+h-j}]\tilde{y}_T' - \mu(pov)\tilde{y}_T\tilde{y}_T' \,\big|\, \{y_t\}_{t=1}^T\big]\right)\zeta_T(\iota,\lambda)'\right\} \\
&\quad_{(6)} +2tr\left\{W\big[\delta(\iota,\lambda) + \alpha\big(\mu(\iota,\lambda) - \mu(pov)\big)\big]\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} + o_p(1).
\end{aligned}$$

Notice that (1) does not depend on (ι,λ) and drops out in the calculation of loss differences. Moreover,

$$\sum_{j=0}^{h-1} F^j \mathbb{E}\big[\mathbb{E}_T[\tilde{z}_{T+h-j}]\tilde{y}_T'\big] = \mu(pov)\Gamma_{yy,0},$$

which implies that terms (4) and (5) are equal to zero. Using that $\delta(\iota,0) = 0$ and $\mu(lfe,0) =$

$\mu(pov)$ we obtain that

$$\begin{aligned}
Q_T(\iota,\lambda) &= tr\left\{W\big[\delta(\iota,\lambda)+\alpha\big(\mu(\iota,\lambda)-\mu(pov)\big)\big]\Gamma_{yy,0}\big[\delta(\iota,\lambda)+\alpha\big(\mu(\iota,\lambda)-\mu(pov)\big)\big]'\right\} \\
&\quad +tr\left\{W\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad +2tr\left\{\big[\delta(\iota,\lambda)+\alpha\big(\mu(\iota,\lambda)-\mu(pov)\big)\big]\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad -tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(lfe,0)'\right\} \\
&\quad +2\left(tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(lfe,0)'\right\}-\hat{\mathcal{R}}_V(lfe,0)\right)+o_p(1) \\
&= tr\left\{W\big[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\big]\Gamma_{yy,0}\big[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\big]'\right\} \\
&\quad +\alpha^2 tr\left\{W\mu(pov)\Gamma_{yy,0}\mu(pov)'\right\} \\
&\quad -2\alpha tr\left\{W\big[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\big]\Gamma_{yy,0}\mu(pov)'\right\} \\
&\quad +tr\left\{W\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad +2tr\left\{\big[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\big]\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad -2\alpha tr\left\{\mu(pov)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad +tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(lfe,0)'\right\}-2\hat{\mathcal{R}}_V(lfe,0)+o_p(1).
\end{aligned}$$

Recall from the proof of Theorem 3 that the MSE differential that determines $PC_T(\iota,\lambda)$

can be written as

$$
\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)\\
&= -2tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\}-2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\}\\
&\quad -2tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\}-2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\}\\
&\quad +tr\left\{W\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\}\\
&\quad +tr\left\{W\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T'(\iota,\lambda)\right\}\\
&\quad +2tr\left\{W\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\}\\
&\quad +tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(lfe,0)'\right\}+\alpha^2tr\left\{W\mu(pov)\Gamma_{yy,0}\mu(pov)'\right\}\\
&\quad +2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\zeta_T(lfe,0)'\right\}+o_p(1)\\
&= tr\left\{W\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\}\\
&\quad +\alpha^2tr\left\{W\mu(pov)\Gamma_{yy,0}\mu(pov)'\right\}\\
&\quad -2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\}\\
&\quad +tr\left\{W\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T'(\iota,\lambda)\right\}\\
&\quad +2tr\left\{W\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\}\\
&\quad -2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\}\\
&\quad +tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(lfe,0)'\right\}\\
&\quad -2tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha(\mu(\iota,\lambda)-\mu(pov))\right]'\right\}\\
&\quad -2tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\}+o_p(1).
\end{aligned}
$$

Therefore,

$$\begin{aligned}
&PC_T(\iota,\lambda) - PC_T(lfe,0) \\
&= tr\left\{W\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\} \\
&\quad +\alpha^2 tr\left\{W\mu(pov)\Gamma_{yy,0}\mu(pov)'\right\} \\
&\quad -2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]'\right\} \\
&\quad +tr\left\{W\zeta_T(\iota,\lambda)\Gamma_{yy,0}\zeta_T'(\iota,\lambda)\right\} \\
&\quad +2tr\left\{W\left[\delta(\iota,\lambda)+\alpha\mu(\iota,\lambda)\right]\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad -2\alpha tr\left\{W\mu(pov)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} \\
&\quad +tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(lfe,0)'\right\} - 2\hat{\mathcal{R}}_V(lfe,0) \\
&\quad -2\left(tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} - \hat{\mathcal{R}}_V(lfe,\iota,0,\lambda)\right) \\
&\quad -2tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha(\mu(\iota,\lambda)-\mu(pov))\right]'\right\} + o_p(1).
\end{aligned}$$

Thus, we obtain the statement of the theorem:

$$\begin{aligned}
&PC_T(\iota,\lambda) - PC_T(lfe,0) \\
&= Q_T(\iota,\lambda) - 2\left(tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\zeta_T(\iota,\lambda)'\right\} - \hat{\mathcal{R}}_V(lfe,\iota,0,\lambda)\right) \\
&\quad -2tr\left\{W\zeta_T(lfe,0)\Gamma_{yy,0}\left[\delta(\iota,\lambda)+\alpha(\mu(\iota,\lambda)-\mu(pov))\right]'\right\} + o_p(1). \quad \blacksquare
\end{aligned}$$

A.4 Proofs for Section 4

Proof of Theorem 5. First, note that

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

Moreover, letting $\hat{u}_t = y_t - \hat{\Phi}_T(mle)y_{t-1}$, the Frisch-Waugh-Lovell theorem implies that the lag-augmenged LFE can be written as

$$\hat{\Psi}_T(lalfe) = \left(\sum_{t=1}^{T-h} y_{t+h}\hat{u}_t'\right)\left(\sum_{t=1}^{T-h}\hat{u}_t\hat{u}_t'\right)^{-1}.$$

Recall that, letting $u_t = y_t - Fy_{t-1}$,

$$\begin{aligned} y_{t+h} &= F^h y_t + \sum_{j=0}^{h-1} F^j \epsilon_{t+h-j} + \alpha T^{-1/2} \sum_{j=0}^{h-1} F^j z_{t+h-j} \\ &= F^h (y_t - F y_{t-1}) + F^{h+1} y_{t-1} + \sum_{j=0}^{h-1} F^j \epsilon_{t+h-j} + \alpha T^{-1/2} \sum_{j=0}^{h-1} F^j z_{t+h-j} \\ &= F^h u_t + F^{h+1} y_{t-1} + \sum_{j=0}^{h-1} F^j \epsilon_{t+h-j} + \alpha T^{-1/2} \sum_{j=0}^{h-1} F^j z_{t+h-j}. \end{aligned}$$

The lag-augmented LFE can then be rewritten as

$$\begin{aligned} \hat{\Psi}_T(lalfe) &= \left(\sum_{t=1}^{T-h} \left[F^h u_t + F^{h+1} y_{t-1} + \sum_{j=0}^{h-1} F^j \epsilon_{t+h-j} + \alpha T^{-1/2} \sum_{j=0}^{h-1} F^j z_{t+h-j} \right] \hat{u}_t' \right) \left(\sum_{t=1}^{T-h} \hat{u}_t \hat{u}_t' \right)^{-1} \\ &= \left(\sum_{t=1}^{T-h} \left[F^h u_t + \sum_{j=0}^{h-1} F^j \epsilon_{t+h-j} + \alpha T^{-1/2} \sum_{j=0}^{h-1} F^j z_{t+h-j} \right] \hat{u}_t' \right) \left(\sum_{t=1}^{T-h} \hat{u}_t \hat{u}_t' \right)^{-1} \\ &= F^h + \left(\sum_{t=1}^{T-h} \left[F^h (u_t - \hat{u}_t) + \sum_{j=0}^{h-1} F^j \epsilon_{t+h-j} + \alpha T^{-1/2} \sum_{j=0}^{h-1} F^j z_{t+h-j} \right] \hat{u}_t' \right) \left(\sum_{t=1}^{T-h} \hat{u}_t \hat{u}_t' \right)^{-1} \\ &= F^h + \left(\sum_{j=0}^{h-1} \sum_{t=1}^{T-h} F^j \epsilon_{t+h-j} \hat{u}_t' \right) \left(\sum_{t=1}^{T-h} \hat{u}_t \hat{u}_t' \right)^{-1} + \alpha T^{-1/2} \left(\sum_{j=0}^{h-1} \sum_{t=1}^{T-h} F^j z_{t+h-j} \hat{u}_t' \right) \left(\sum_{t=1}^{T-h} \hat{u}_t \hat{u}_t' \right)^{-1}. \end{aligned}$$

The first line plugs in the expression for y_{t+h} above. The second line uses the fact that $\sum_{t=1}^{T-h} y_{t-1} \hat{u}_t' = 0$ by construction. The third line adds and subtracts $F^h \hat{u}_t$. The last line uses $u_t - \hat{u}_t = (F - \hat{\Phi}_T(mle)) y_{t-1}$, and again $\sum_{t=1}^{T-h} y_{t-1} \hat{u}_t' = 0$ by definition of $\hat{u}_t$.

Therefore,

$$\begin{aligned} \sqrt{T}(\bar{\Psi}_T(lfe, \lambda) - F^h) = \sqrt{T}(\underline{\Psi}_T - F^h) \tilde{\lambda} \underline{P}_\Psi \bar{P}_\Psi^{-1} &+ \alpha \left(\sum_{j=0}^{h-1} \sum_{t=1}^{T-h} F^j z_{t+h-j} \hat{u}_t' \right) \bar{P}_\Psi^{-1} \\ &+ \sqrt{T} \left(\sum_{j=0}^{h-1} \sum_{t=1}^{T-h} F^j \epsilon_{t+h-j} \hat{u}_t' \right) \bar{P}_\Psi^{-1}. \end{aligned} \tag{A.4}$$

For the first term on the RHS of (A.4), using the drifting sequence of priors we have

$$\begin{aligned} \sqrt{T}(\underline{\Psi}_T - F^h) \tilde{\lambda} \underline{P}_\Psi \bar{P}_\Psi^{-1} &= \underline{\psi} \lambda \underline{P}_\Psi \left(\lambda \underline{P}_\Psi + T^{-1} \sum_{t=1}^{T-h} \hat{u}_t \hat{u}_t' \right)^{-1} \\ &= \underline{\psi} \lambda \underline{P}_\Psi \left(\lambda \underline{P}_\Psi + \Sigma_{\epsilon\epsilon} \right)^{-1} + o_p(1). \end{aligned}$$

In particular, in the last step we use $\Sigma_{uu} = \Sigma_{\epsilon\epsilon} + O(1/T)$.

For the second term on the RHS of (A.4), note that

$$\frac{1}{T}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j z_{t+h-j}\hat{u}_t' = \frac{1}{T}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j z_{t+h-j}u_t' + \frac{1}{T}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j z_{t+h-j}(\hat{u}_t - u_t)'.$$

The ergodic theorem implies that

$$\begin{aligned}\frac{1}{T}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j z_{t+h-j}u_t' &= \sum_{j=0}^{h-1} F^j \mathbb{E}[z_{t+h-j}u_t'] + o_p(1)\\ &= \sum_{j=0}^{h-1} F^j A_{h-j}\Sigma_{\epsilon\epsilon} + o_p(1).\end{aligned}$$

At the same time, consistency of $\hat{\Phi}_T(mle)$ implies that

$$\frac{1}{T}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j z_{t+h-j}(\hat{u}_t - u_t)' = \frac{1}{T}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j z_{t+h-j}y_{t-1}'(\hat{\Phi}_T(mle) - F)' = o_p(1).$$

The second term on the RHS of (A.4) is then

$$\alpha\left(\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j z_{t+h-j}\hat{u}_t'\right)\bar{P}_\Psi^{-1} = \alpha\left(\sum_{j=0}^{h-1} F^j A_{h-j}\Sigma_{\epsilon\epsilon}\right)\left(\lambda\underline{P}_\Psi + \Sigma_{\epsilon\epsilon}\right)^{-1} + o_p(1).$$

For the third term on the RHS of (A.4), the central limit theorem implies that

$$\begin{aligned}T^{-1/2}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j\epsilon_{t-j}\hat{u}_t' &= T^{-1/2}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j\epsilon_{t-j}u_t' + T^{-1/2}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j\epsilon_{t-j}(\hat{u}_t - u_t)'\\ &= T^{-1/2}\sum_{j=0}^{h-1}\sum_{t=1}^{T-h} F^j\epsilon_{t-j}u_t' + o_p(1)\\ &= \nu + o_p(1),\end{aligned}$$

for a centered Gaussian vector $vecr(\nu)$.

The covariance matrix $\mathcal{V}$ follows from the same arguments as in Theorem 1. In particular,

$$vecr\left(\zeta_T(lalfe,\lambda)\right) = \sum_{j=0}^{h-1}\left(F^j \otimes \left(\lambda\underline{P}_\Psi + \Sigma_{\epsilon\epsilon}\right)^{-1}\right) vec\left(T^{-1/2}\sum_{t=1}^{T-h} u_t\epsilon_{t+h-j}'\right),$$

which implies that

$$\mathcal{V}(lalfe,\lambda,\lambda') = \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}\left(F^i\Sigma_{\epsilon\epsilon}F^{j'}\right)\otimes\left((\lambda\underline{P}'_{\Psi}+\Sigma_{\epsilon\epsilon})^{-1}\Sigma_{\epsilon\epsilon}(\lambda'\underline{P}_{\Psi}+\Sigma_{\epsilon\epsilon})^{-1}\right)$$

$$\mathcal{V}(mle,lalfe,\lambda,\lambda') = \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}\left(F^i\Sigma_{\epsilon\epsilon}F^{j'}\right)\otimes\left(F^{h-1-j'}(\lambda\underline{P}'_{\Phi}+\Gamma_{yy,0})^{-1}\Sigma_{\epsilon\epsilon}(\lambda'\underline{P}_{\Psi}+\Sigma_{\epsilon\epsilon})^{-1}\right)$$

$$\mathcal{V}(lfe,lalfe,\lambda,\lambda') = \sum_{i=0}^{h-1}\sum_{j=0}^{h-1}\left(F^i\Sigma_{\epsilon\epsilon}F^{j'}\right)\otimes\left((\lambda\underline{P}'_{\Psi}+\Gamma_{yy,0})^{-1}\Sigma_{\epsilon\epsilon}(\lambda'\underline{P}_{\Psi}+\Sigma_{\epsilon\epsilon})^{-1}\right).$$

For $\iota\in\{mle,lfe\},\lambda\geq 0$, $\mathcal{V}$ equals V in Theorem 1. ■

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 (46) for $\underline{\phi}$ as a function of $\underline{\psi}$. Note that

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

Thus,

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

In turn,

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

Table A-1: Relative Risk, Horizon $h = 2$, Prior 2

		$\lambda = 0$		$\lambda = 5$		$\lambda = 15$		$\lambda = 50$		$\lambda = 200$	
T		MLE	LFE	MLE	LFE	MLE	LFE	MLE	LFE	MLE	LFE
Misspecification $\alpha = 0$											
100	Risk	-3.1	0	-11.1	-9.5	-13.8	-11.5	-13.2	-10.1	-13.4	-10
	$\mathbb{E}[PC]$	-1.8	0	-7.7	-6.6	-9.1	-7.2	-9.5	-6.8	-8.8	-5.5
	$\sigma[PC]$	2.6	0	7.3	6	9.8	9.4	12.2	12.2	14.4	14.8
500	Risk	-2.5	0	-7.8	-6.7	-9.4	-8.1	-8.5	-7	-7.5	-5.8
	$\mathbb{E}[PC]$	-1.7	0	-7.2	-6.5	-8.2	-7.1	-8.1	-6.6	-7	-5.4
	$\sigma[PC]$	2.6	0	6.8	5.6	9.3	8.7	11.8	11.7	14.3	14.4
5000	Risk	-2.6	0	-7.6	-6.8	-7.9	-7.2	-8	-7.2	-5.3	-4.7
	$\mathbb{E}[PC]$	-1.6	0	-6.8	-6.4	-7.9	-7.3	-6.9	-6.3	-6.3	-5.7
	$\sigma[PC]$	2.7	0	7	5.7	9	8.4	12.1	11.8	14	14
∞	Risk	-2.6	0	-6.9	-6.3	-7.6	-7.1	-6.9	-6.6	-5.4	-5.3
Misspecification $\alpha = 2$											
100	Risk	2.4	0	-4.2	-6	-6	-5.8	-5.9	-3.5	-4	0.2
	$\mathbb{E}[PC]$	3.7	0	-0.6	-3.4	-1.6	-1.8	-1.3	1.1	-0.5	3.7
	$\sigma[PC]$	4.9	0	9.4	6.9	11.8	11	14.5	14.8	16.4	17.1
500	Risk	1.9	0	-1.7	-4.5	-2.3	-3.6	-1.6	-1.4	-1.1	0.4
	$\mathbb{E}[PC]$	3.6	0	0.4	-3.2	0.3	-1.5	1.2	1.3	2.7	4.3
	$\sigma[PC]$	4.2	0	8.7	6.4	11.2	10.2	13.9	13.6	16	16.1
5000	Risk	1.2	0	-1.8	-4.9	-2.3	-4.5	-0.5	-1.5	1	1.2
	$\mathbb{E}[PC]$	3.1	0	1	-3	1.2	-1.6	2.3	1.1	4.6	4.7
	$\sigma[PC]$	3.9	0	8.5	6.4	10.7	9.7	13.8	13.3	16.1	16
∞	Risk	2.1	0	0.9	-3	1.5	-1.5	3.1	1.4	5	4.4
Misspecification $\alpha = 5$											
100	Risk	29.3	0	18.3	0.4	15	6.8	13.4	14.4	13.8	20
	$\mathbb{E}[PC]$	27.3	0	20.4	0.6	18	8	17.7	17.2	18.6	24
	$\sigma[PC]$	12.9	0	16.6	7.8	19.1	14.9	22.1	21.5	24.9	26
500	Risk	31.4	0	20.2	2.5	17.1	6.9	21.4	17.7	17.1	18.1
	$\mathbb{E}[PC]$	34.7	0	24	2.8	21.9	9.3	22.4	17.7	23.3	24
	$\sigma[PC]$	10.1	0	14.4	8.6	16.8	14	19.2	18.2	21.9	21.9
5000	Risk	25.1	0	16.1	0.2	15.3	4.2	14.1	8.8	16.8	15.5
	$\mathbb{E}[PC]$	31	0	24.4	3.8	23.8	10	24.7	17.9	26.2	24.7
	$\sigma[PC]$	8.4	0	12.4	8.1	14.4	12.1	17	16	19.6	19.3
∞	Risk	26.8	0	24	4.1	24.5	10.4	26	18.3	27.7	25.1

Notes: The table reports standardized prediction risk differentials $T[\mathcal{R}(\hat{y}_{T+h}(\iota, \lambda)) - \mathcal{R}(\hat{y}_{T+h}(lfe, 0))]$. Negative entries correspond to improvements (risk reductions) relative to the benchmark. $\mathbb{E}[PC]$ and $\sigma[PC]$ refer to expected value and standard deviation of the PC criterion.

Table A-2: Relative Risk, Horizon $h = 4$, Prior 2

		$\lambda = 0$		$\lambda = 5$		$\lambda = 15$		$\lambda = 50$		$\lambda = 200$	
T		MLE	LFE	MLE	LFE	MLE	LFE	MLE	LFE	MLE	LFE
Misspecification $\alpha = 0$											
100	Risk	-17.9	0	-31.9	-26.9	-37.7	-33.3	-36.4	-31.1	-39.3	-33.3
	$\mathbb{E}[PC]$	-9.3	0	-21.7	-18.1	-27.5	-22.7	-31.2	-25.5	-31.5	-25.7
	$\sigma[PC]$	13.5	0	19.8	13.7	24.8	22.4	31.9	32.4	37.2	38.9
500	Risk	-14.3	0	-24.7	-21.4	-32.9	-29.1	-32.5	-28.9	-32.6	-28.9
	$\mathbb{E}[PC]$	-9.6	0	-22	-18.8	-26.4	-23.3	-30.4	-26.7	-29.7	-26.1
	$\sigma[PC]$	12.8	0	18.6	12.2	24.5	20.9	30.9	30.1	38.6	39
5000	Risk	-12.2	0	-23.8	-19.9	-27.2	-24.4	-32.5	-30.2	-28.9	-27.2
	$\mathbb{E}[PC]$	-9.5	0	-21.1	-18.7	-26.7	-24.4	-28.6	-26.5	-27.6	-26
	$\sigma[PC]$	12.7	0	19.1	12.4	22.9	19.1	31.3	29.6	37.3	37
∞	Risk	-12.9	0	-22.3	-18.9	-26.8	-24.5	-28.4	-27.1	-26.9	-26.4
Misspecification $\alpha = 2$											
100	Risk	-17.8	0	-28.5	-22.9	-36.2	-30	-34	-26	-35.8	-26.3
	$\mathbb{E}[PC]$	-5.3	0	-16.2	-14.2	-19.5	-14.8	-23.2	-16.1	-21.7	-13.2
	$\sigma[PC]$	18.1	0	23.9	15.2	30.6	27	38.3	38.4	46.2	48.1
500	Risk	-9.1	0	-20.7	-18.2	-24.3	-21.1	-23.9	-19.8	-20.7	-15.6
	$\mathbb{E}[PC]$	-5.2	0	-15.9	-15.3	-19.3	-17.5	-19.9	-16.4	-19.5	-14.9
	$\sigma[PC]$	16.8	0	23.4	14.7	28.9	24.2	37.3	35.9	44.3	44.9
5000	Risk	-5.8	0	-14.5	-13.7	-17.3	-16.3	-18.1	-17	-16.1	-14.4
	$\mathbb{E}[PC]$	-5	0	-14.9	-15.5	-18.5	-18.5	-18.6	-17.8	-16.8	-15.3
	$\sigma[PC]$	15.8	0	23.4	14.6	27.2	22.2	35	32.6	44	43.5
∞	Risk	-8.4	0	-16	-15.6	-18.8	-18.9	-18.5	-18.8	-16	-16.1
Misspecification $\alpha = 5$											
100	Risk	-14.4	0	-24.3	-23.9	-25.6	-21.1	-23.2	-12.5	-17.7	-2
	$\mathbb{E}[PC]$	11.6	0	3.9	-5.7	1.2	0.8	4.2	13.4	6	20.3
	$\sigma[PC]$	30	0	35.5	16.2	42.3	32.6	52.7	51.4	61.3	64
500	Risk	9.9	0	-0.6	-9	-5.6	-9.1	0.3	2.4	4.5	10.9
	$\mathbb{E}[PC]$	17.2	0	5.6	-6.9	3.3	-2.3	4.7	5.7	7.9	13.5
	$\sigma[PC]$	28.8	0	34.6	18.8	40.5	32	48.7	45.9	56.1	56.3
5000	Risk	21.4	0	13.1	-2	8.2	-0.7	12.5	9.2	15.9	16.7
	$\mathbb{E}[PC]$	19.1	0	6.2	-8.2	4	-5.1	4.7	1	8	8.5
	$\sigma[PC]$	27.6	0	31.7	18.2	36.3	29	43.6	39.9	53	52.1
∞	Risk	15.1	0	4.8	-8.7	2.9	-7	4.6	-1.1	7.9	5.8

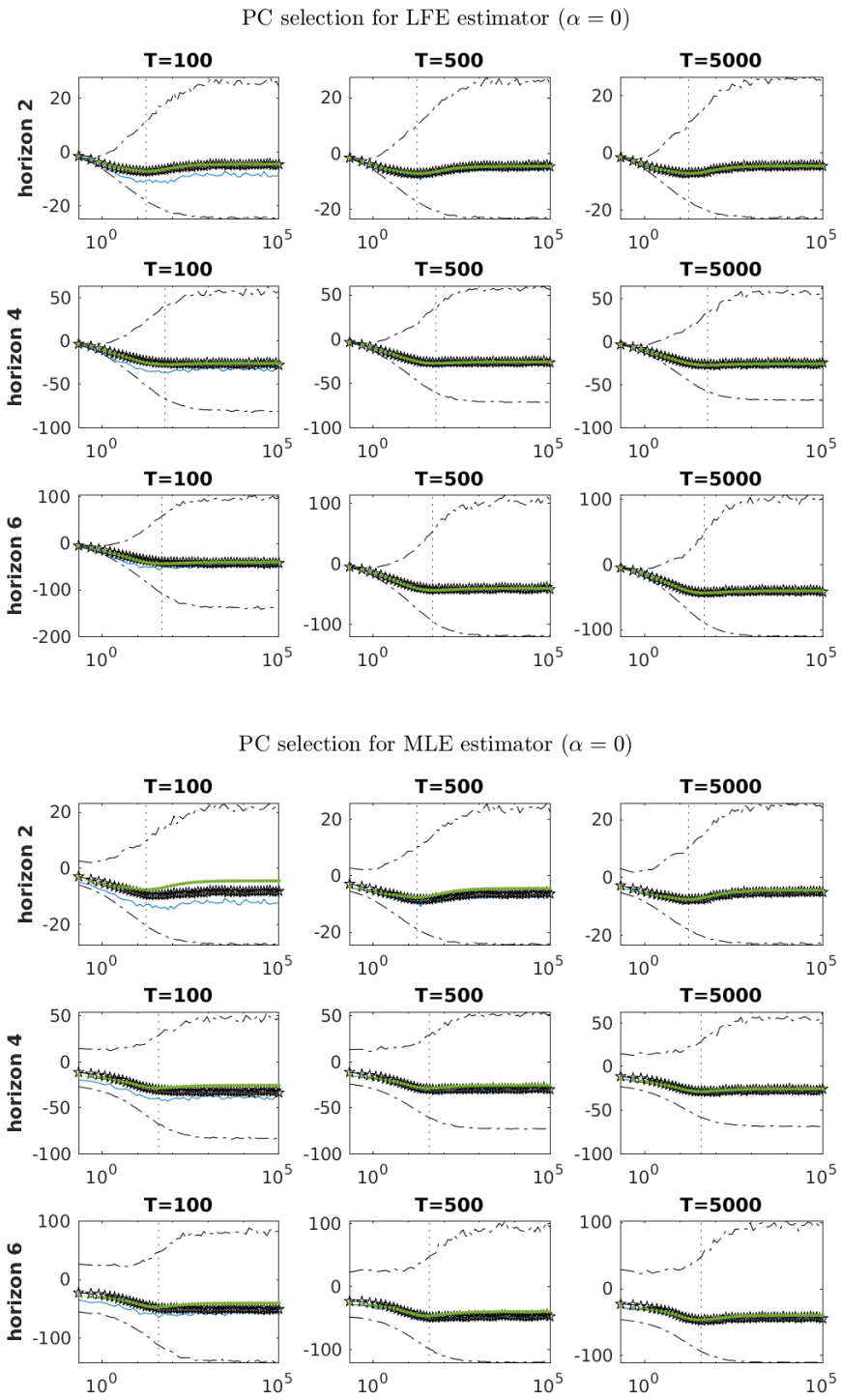
Notes: The table reports standardized prediction risk differentials $T[\mathcal{R}(\hat{y}_{T+h}(\iota, \lambda)) - \mathcal{R}(\hat{y}_{T+h}(lfe, 0))]$. Negative entries correspond to improvements (risk reductions) relative to the benchmark. $\mathbb{E}[PC]$ and $\sigma[PC]$ refer to expected value and standard deviation of the PC criterion.

Table A-3: Relative Risk, Horizon $h = 6$, Prior 2

		$\lambda = 0$		$\lambda = 5$		$\lambda = 15$		$\lambda = 50$		$\lambda = 200$	
T		MLE	LFE	MLE	LFE	MLE	LFE	MLE	LFE	MLE	LFE
Misspecification $\alpha = 0$											
100	Risk	-31.2	0	-50	-41.6	-57.4	-50.7	-62.3	-53.9	-54.4	-45.8
	$\mathbb{E}[PC]$	-19.9	0	-35.5	-27	-43.3	-34.7	-49.2	-40.5	-50	-41.5
	$\sigma[PC]$	25.5	0	32	19.2	40	33.9	52.1	51.5	64.3	66.6
500	Risk	-26.6	0	-39.5	-31.3	-49.7	-42.8	-51.2	-44.1	-51.7	-45
	$\mathbb{E}[PC]$	-21.9	0	-35.9	-28.7	-44.4	-37.4	-49.1	-41.9	-48.4	-41.7
	$\sigma[PC]$	23.5	0	32.2	18.1	38.4	30.6	51.3	49.3	62.8	63.8
5000	Risk	-25.1	0	-38	-31.4	-45.7	-39.9	-50.5	-45.5	-46.8	-43.2
	$\mathbb{E}[PC]$	-20.2	0	-35.3	-29.5	-43.5	-38.8	-46.7	-42.4	-46	-42.4
	$\sigma[PC]$	25.7	0	31.8	17.1	38	28.7	51.4	46.9	62.6	62.1
∞	Risk	-25.8	0	-36.5	-29.5	-44	-39.3	-46.4	-43.3	-43	-41.8
Misspecification $\alpha = 2$											
100	Risk	-37	0	-58.3	-45.7	-62.7	-53.1	-57.7	-47.5	-56.3	-44.4
	$\mathbb{E}[PC]$	-17.6	0	-30.2	-23.3	-35	-26.5	-37.3	-26.5	-37.6	-25.9
	$\sigma[PC]$	32.6	0	40.5	21.8	49.1	39.4	64.1	61.6	76	78.7
500	Risk	-27.1	0	-41	-30.5	-48.5	-41.2	-48.1	-40.8	-43.1	-34.9
	$\mathbb{E}[PC]$	-17.3	0	-32	-26.4	-36.9	-30.7	-36.1	-28.7	-36.1	-27.5
	$\sigma[PC]$	29.2	0	37	19.9	45.9	35.8	62.1	58.2	73.1	74.2
5000	Risk	-22.1	0	-36.2	-29.2	-39	-34.3	-40.6	-36.8	-38.3	-34.4
	$\mathbb{E}[PC]$	-16.8	0	-29	-25.6	-35.8	-32.5	-34.4	-30.9	-28.8	-24.9
	$\sigma[PC]$	28	0	37.6	20.4	43.6	32.1	59.3	53.4	76.2	74.9
∞	Risk	-21.5	0	-31.2	-26	-35	-32.4	-32.9	-31.9	-26.9	-26.6
Misspecification $\alpha = 5$											
100	Risk	-43.7	0	-62.9	-46.1	-63.4	-52.2	-54.7	-39.5	-48.8	-29.9
	$\mathbb{E}[PC]$	-7	0	-16.3	-15.9	-17.8	-11.4	-14.4	-0.5	-5.2	13.1
	$\sigma[PC]$	44.8	0	50	20	63	44.3	83	75.9	103	104.6
500	Risk	-25.3	0	-41.9	-35.5	-39.1	-35.9	-33.1	-26.4	-23.1	-12.6
	$\mathbb{E}[PC]$	-2.6	0	-14.3	-18.6	-15.3	-16.1	-10.8	-4.9	-1.3	9.7
	$\sigma[PC]$	41.4	0	50.1	23.5	61	42.7	78.3	70.6	98.4	98
5000	Risk	-11.9	0	-24.8	-25.7	-25.4	-27.9	-22.4	-22.9	-17.1	-14.1
	$\mathbb{E}[PC]$	3.3	0	-10.7	-19.1	-10.5	-17	-5.3	-7.1	5.6	8.6
	$\sigma[PC]$	41	0	46.4	23.3	56.5	40	74.3	64.8	92.7	90
∞	Risk	1	0	-9.3	-18.4	-8.7	-17.9	-1	-7.4	8.2	5.8

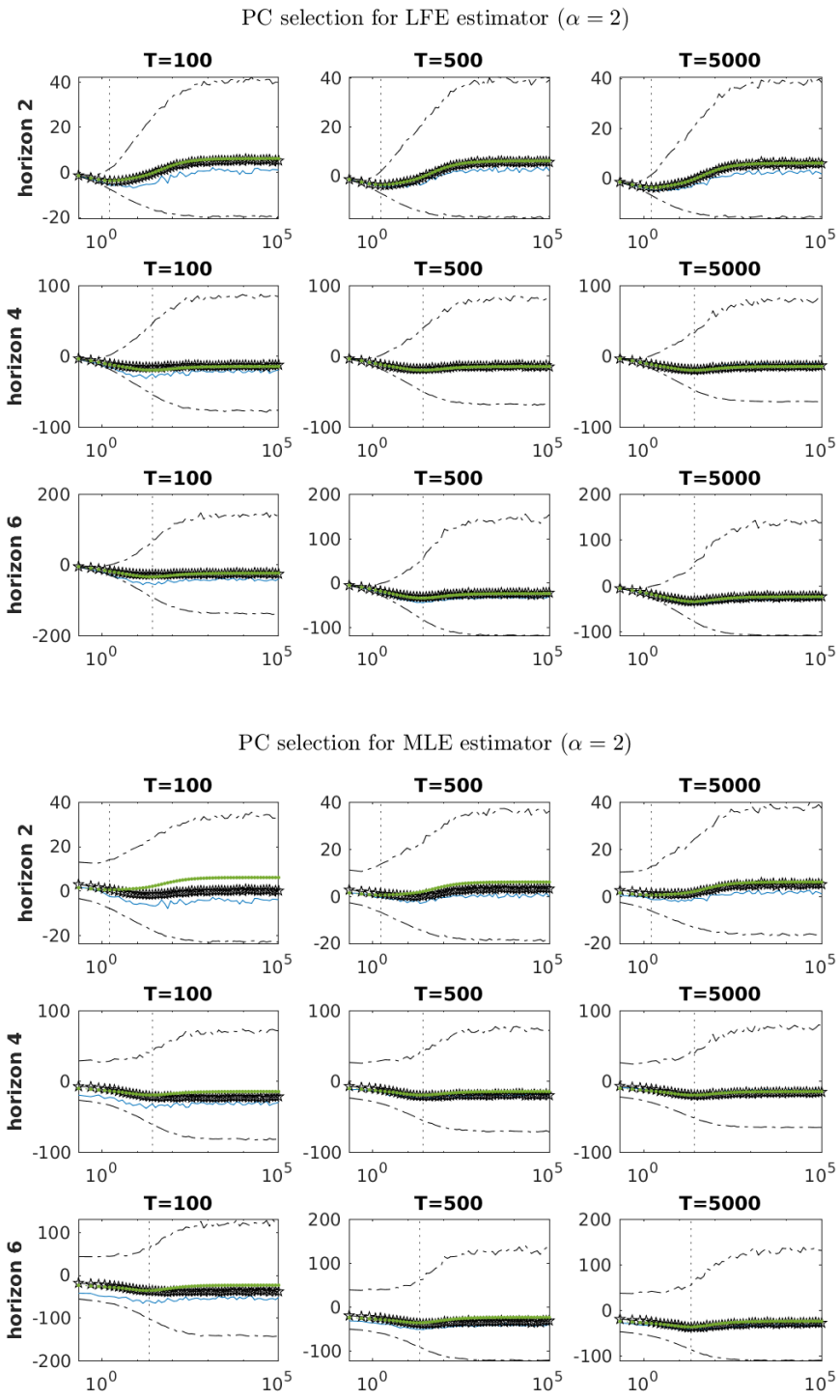
Notes: The table reports standardized prediction risk differentials $T[\mathcal{R}(\hat{y}_{T+h}(\iota, \lambda)) - \mathcal{R}(\hat{y}_{T+h}(lfe, 0))]$. Negative entries correspond to improvements (risk reductions) relative to the benchmark. $\mathbb{E}[PC]$ and $\sigma[PC]$ refer to expected value and standard deviation of the PC criterion.

Figure A-1: PC versus Finite Sample Risk: Design 1



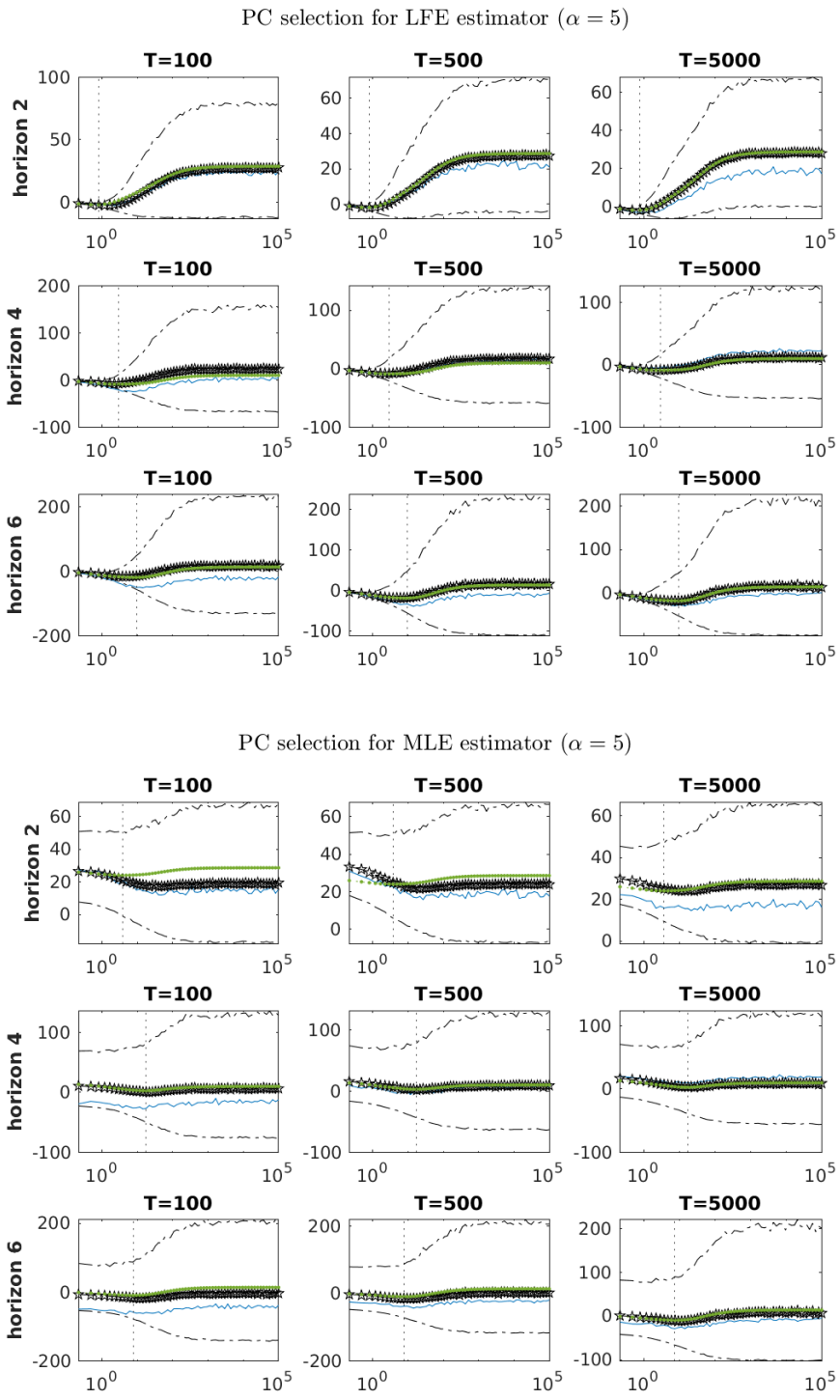
Notes: The dotted green line is the asymptotic risk. The starred black line is $\mathbb{E}[PC]$. The solid blue line is the MC risk and the dashed black lines are 90% coverage intervals for the finite sample losses. The vertical line indicates the value of λ that minimizes the asymptotic risk.

Figure A-2: PC versus Finite Sample Risk: Design 2



Notes: The dotted green line is the asymptotic risk. The starred black line is $\mathbb{E}[PC]$. The solid blue line is the MC risk and the dashed black lines are 90% coverage intervals for the finite sample losses. The vertical line indicates the value of λ that minimizes the asymptotic risk.

Figure A-3: PC versus Finite Sample Risk: Design 3



Notes: The dotted green line is the asymptotic risk. The starred black line is $\mathbb{E}[PC]$. The solid blue line is the MC risk and the dashed black lines are 90% coverage intervals for the finite sample losses. The vertical line indicates the value of λ that minimizes the asymptotic risk.

Figure A-4: PC versus MDD Objective Function

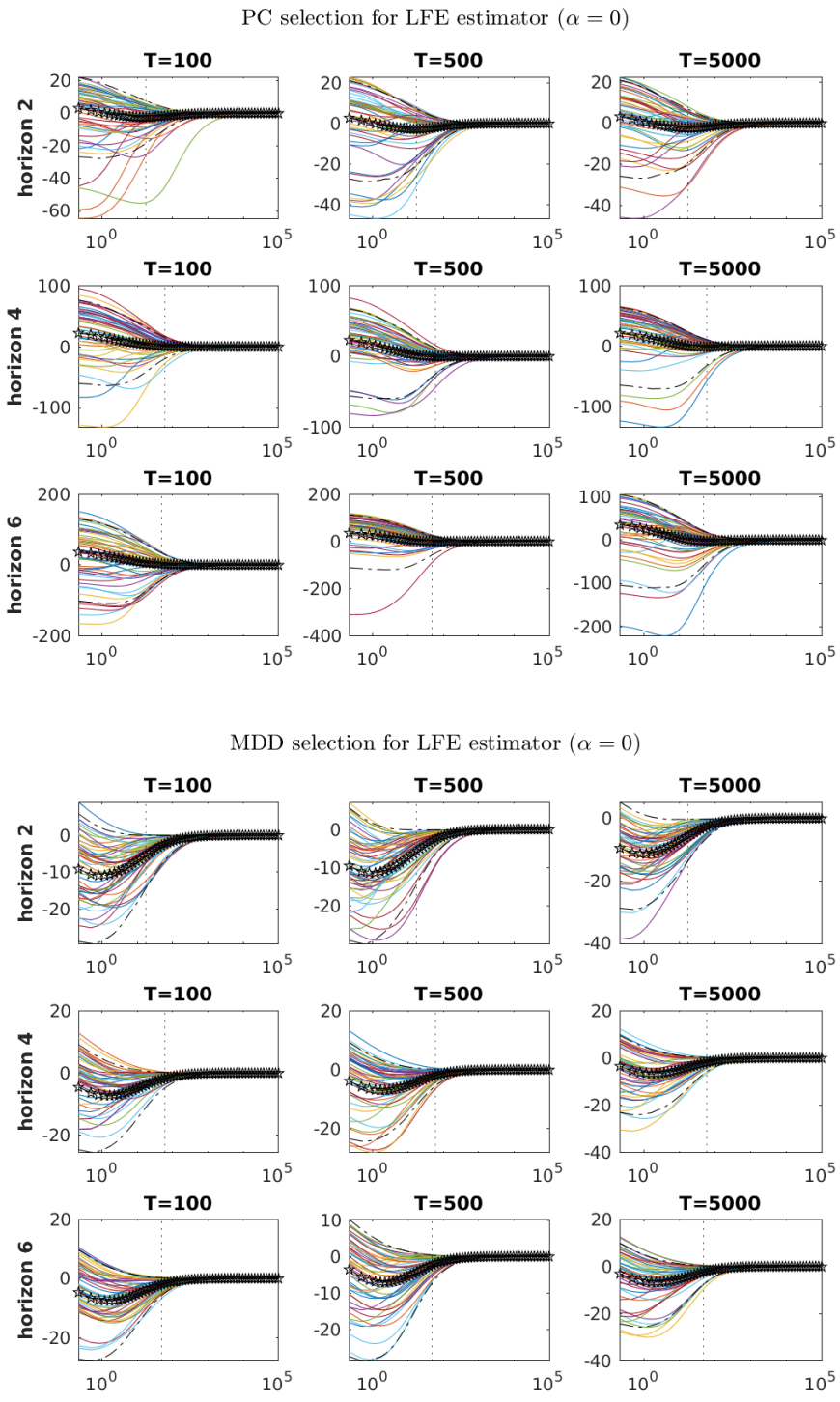


Figure A-5: PC versus MDD Objective Function

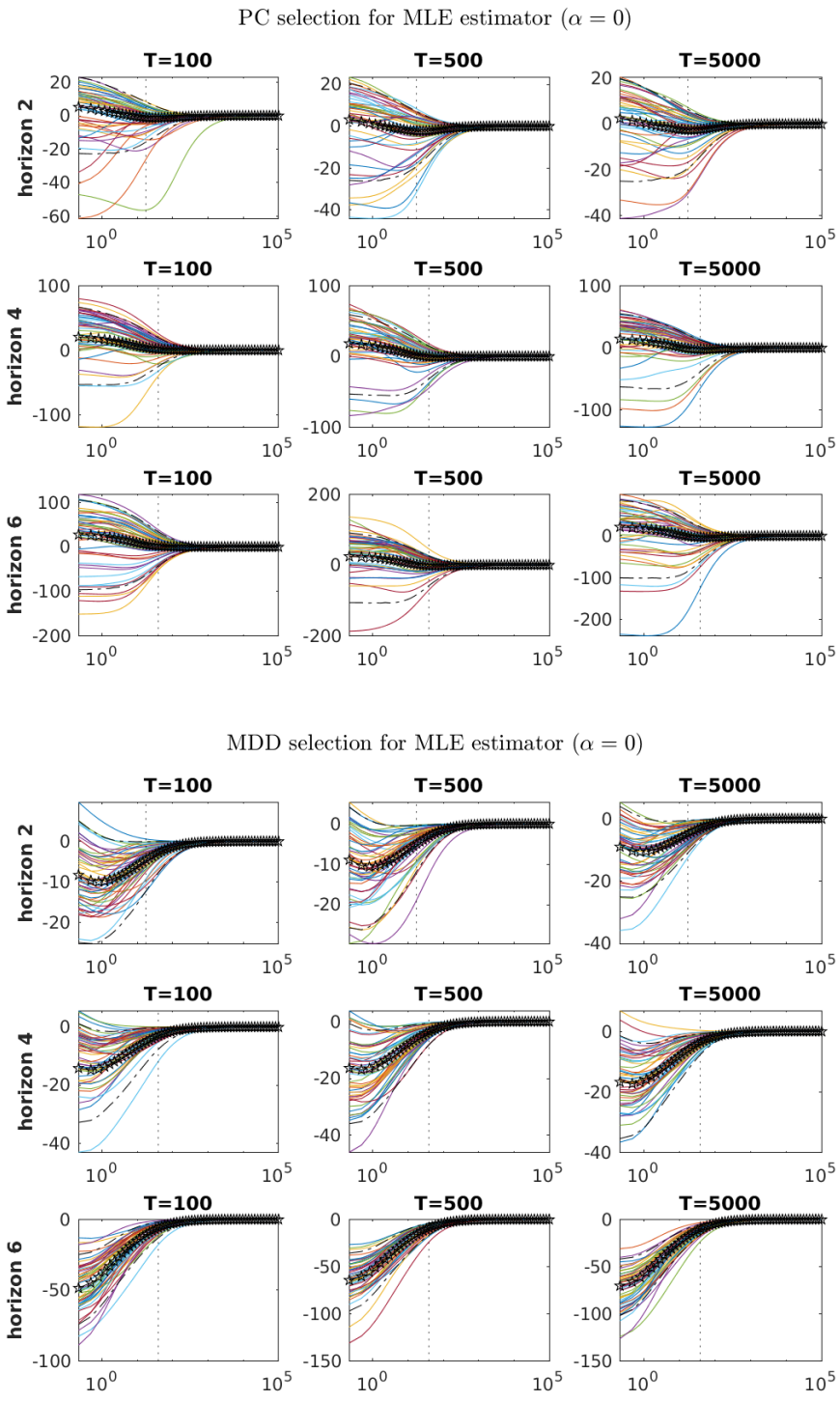


Figure A-6: PC versus MDD Objective Function

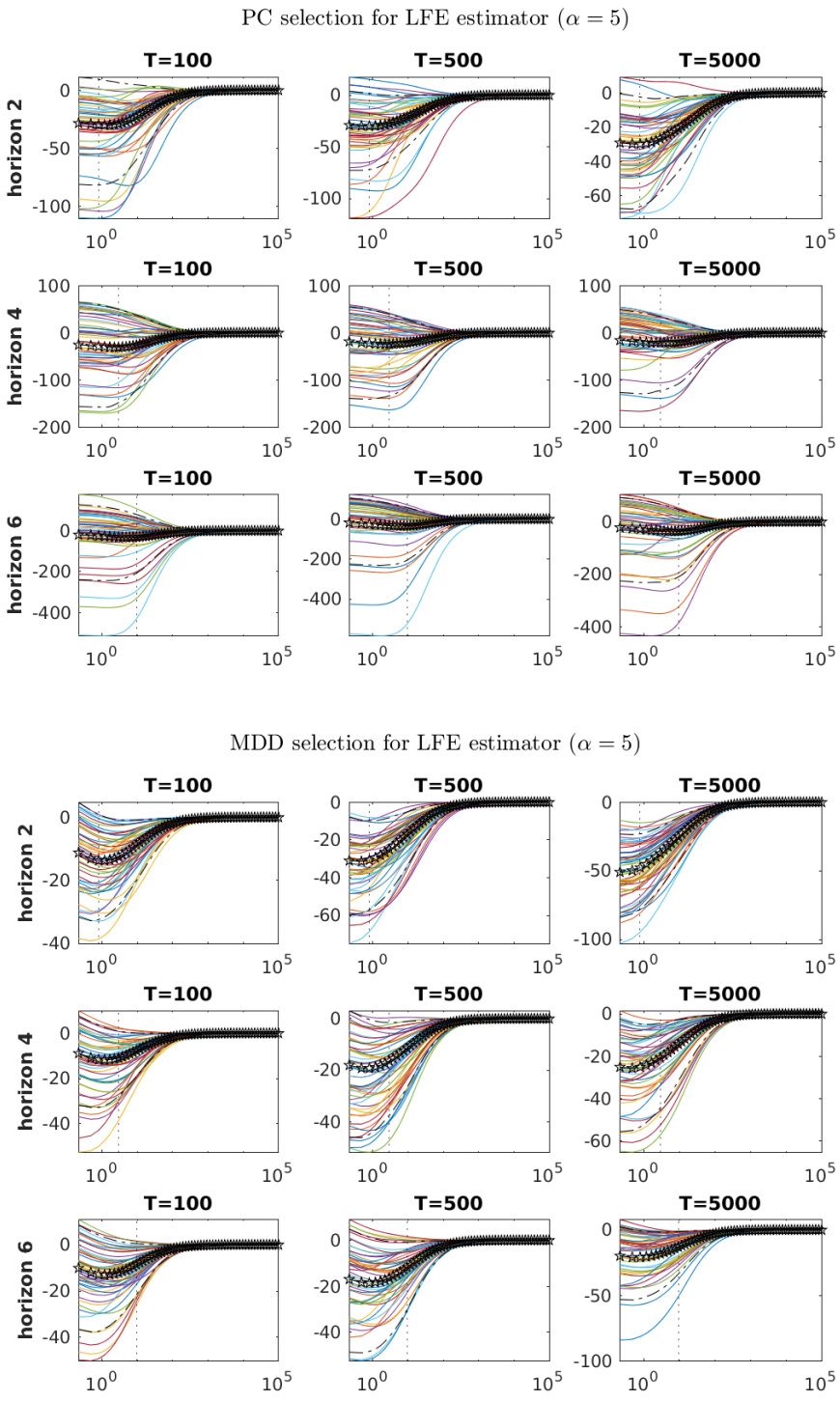


Figure A-7: PC versus MDD Objective Function

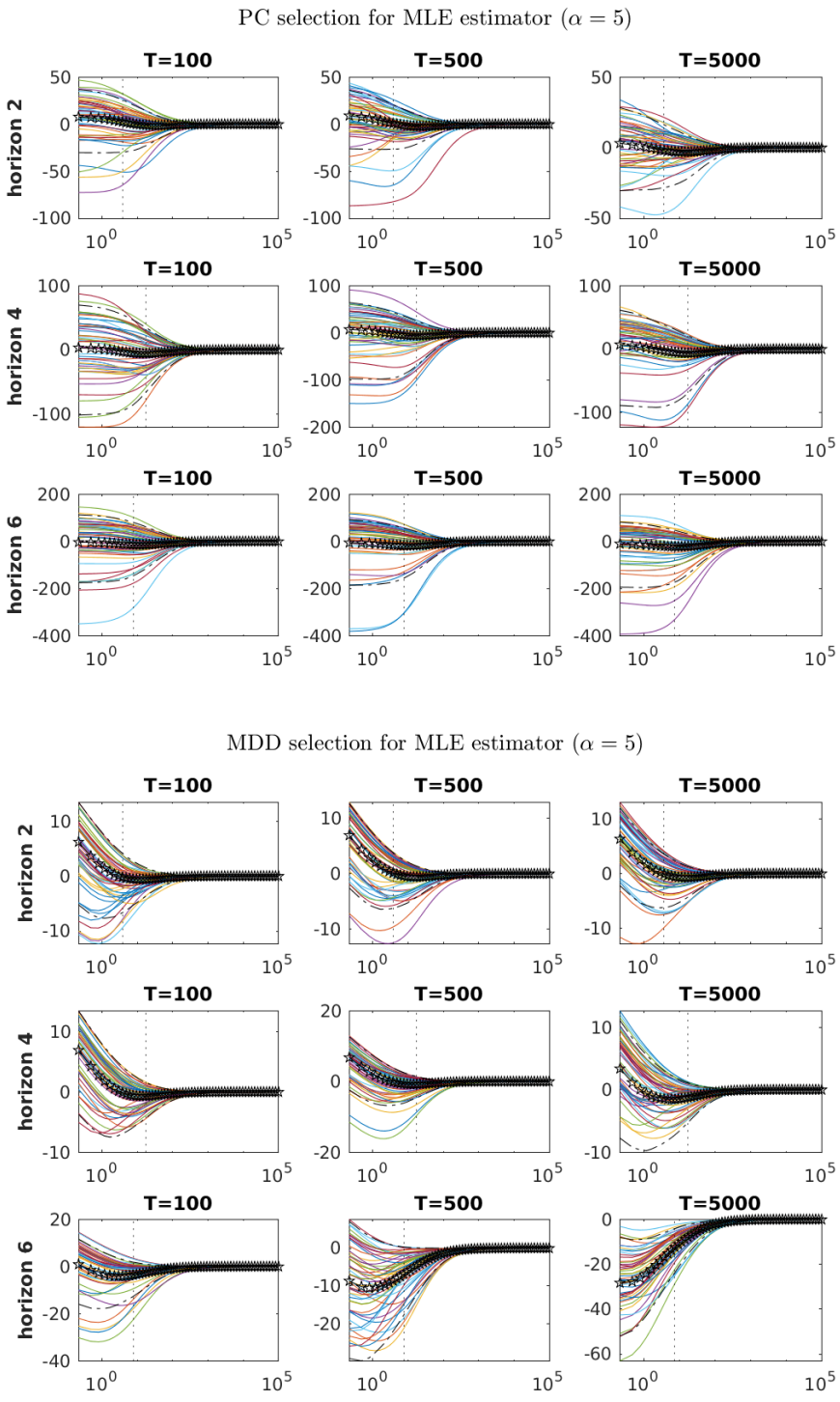


Figure A-8: Distribution of Optimal Shrinkage Hyperparameter without Misspecification

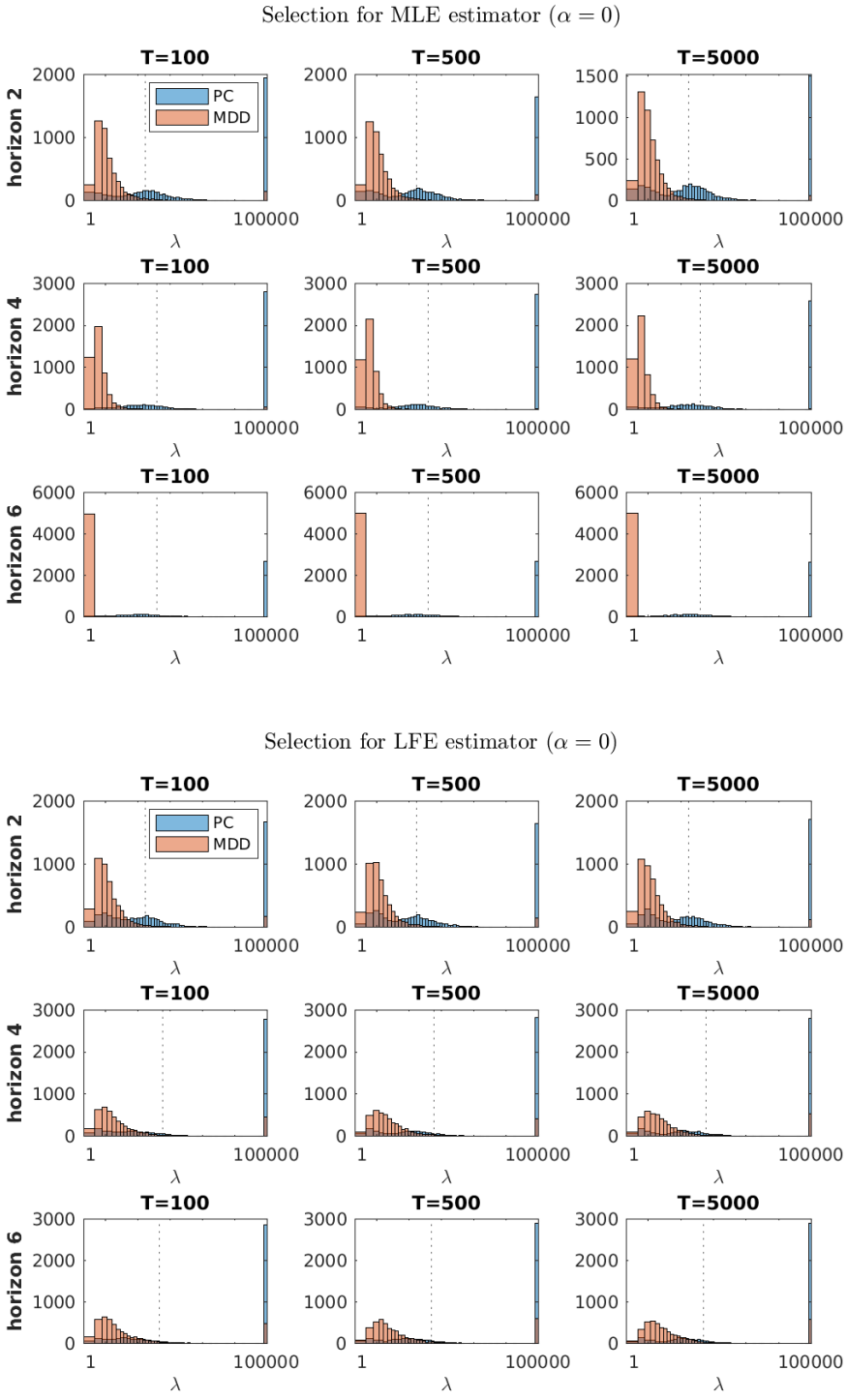
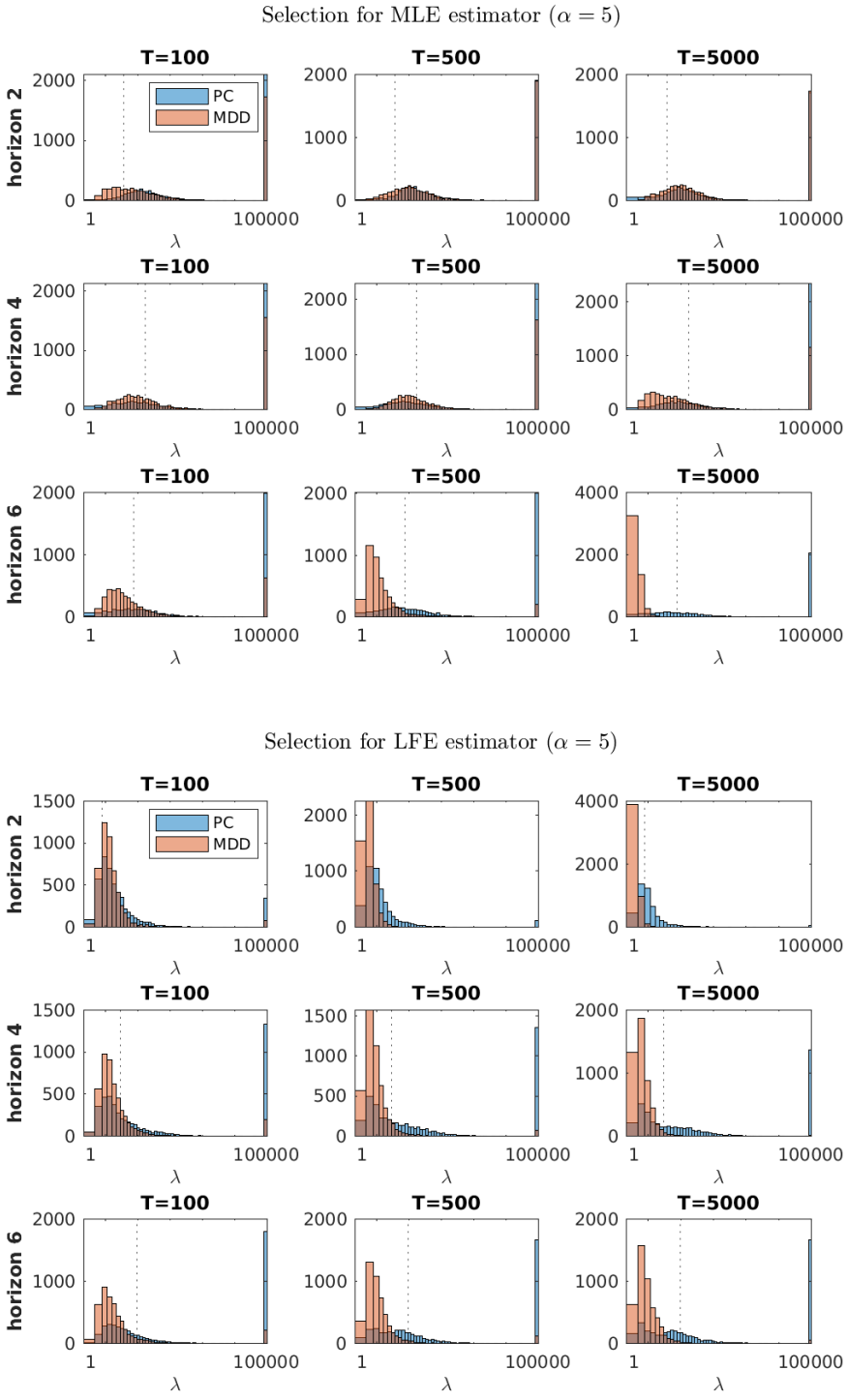


Figure A-9: Distribution of Optimal Shrinkage Hyperparameter with Misspecification



C Further Details on the Empirical Analysis