Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19103-6297

pier@ssc.upenn.edu
http://www.econ.upenn.edu/pier

PIER Working Paper 99-006

"Loss Function vs. Likelihood Estimation of Forecasting Models: A Pre-test Procedure and a Bayesian Interpretation"

by

Frank Schorfheide

Loss Function vs. Likelihood Estimation of Forecasting Models: A Pre-test Procedure and a Bayesian Interpretation

Frank Schorfheide*

June, 1999

*Correspondence: c/o University of Pennsylvania, Department of Economics, Philadelphia, PA 19104-6297. Email: *schorf@ssc.upenn.edu.* URL: *http://www.econ.upenn.edu/~schorf.* This paper is based on Chapter 3 of my Ph.D. dissertation. An earlier version circulated under the title "Prediction with Misspecified Models - A Bayesian Perspective". I wish to thank my advisors Peter Phillips and Chris Sims for valuable comments, suggestions, and support. Financial Assistance from the German Academic Exchange Service through "Doktorandenstipendium aus Mitteln des zweiten Hochschulsonderprogramms" and the Alfred P. Sloan Foundation are gratefully acknowledged.

Abstract

The paper considers the problem of using a vector autoregression (VAR) to forecast a stationary process several periods into the future. If the VAR is misspecified, it might be best to use the loss function under which the forecasts are evaluated also for parameter estimation. It is a plausible and straightforward procedure to conduct a model check of the VAR before adopting a loss function estimator. If the VAR is discredited by the data then a loss function estimator is used, otherwise the parameters are estimated by a likelihood based technique. We calculate the asymptotic prediction risk for such a pre-test procedure under the assumptions that the data are generated from a linear process that drifts toward the VAR as the sample size tends to infinity. The pre-test can avoid picking the inferior estimator when the stakes are high. This is confirmed by a small Monte Carlo study. A Bayesian interpretation of loss function estimation and the pre-test procedure is provided. A forecaster places non-zero prior probability on a reference model but finds it too onerous to calculate its posterior predictive distribution. Instead he chooses a prediction procedure based on the VAR that has a small integrated prediction risk.

Key Words: Bayesian Analysis, Forecasting, Loss Function Estimation, Pre-testing

JEL Classification: C11, C32, C53

1 Introduction

Forecasts of future observations y_{T+h} are often based on parametric probability models. Suppose the distribution of the future observation is given by a density $p(y_{T+h}|\theta, \Upsilon_T)$, where Υ_T is a vector of past observations that are available at time T, and θ is a vector of parameters. The forecast $\hat{y}_{T+h|T}$, made at time T, is evaluated according to a loss function $L(y_{T+h}, \hat{y}_{T+h|T})$. If the parameter vector θ is known, then the optimal forecast as a function of θ is given by[1]

$$\hat{y}_{T+h|T}(\theta) = \text{argmin}_{y^* \in \boldsymbol{R}^n} \int_{y_{T+h}} L(y_{T+h}, y^*) p(y_{T+h}|\theta, \Upsilon_T) dy_{T+h} \qquad (1)$$

However, before this predictor can be used to generate a forecast, a suitable parameter value has to be determined. An interesting question is whether one should employ the loss function that is used to evaluate the forecasts, to obtain a parameter value for θ. A loss function estimator of θ can be obtained by minimizing in-sample forecast losses:

$$\hat{\theta}_{T,l} = \text{argmin}_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T-h} L(y_{t+h}, \hat{y}_{t+h|t}(\theta)) \qquad (2)$$

Future observations can be forecasted with the resulting plug-in predictor $\hat{y}_{T+h|T}(\hat{\theta}_{T,l})$. This predictor will be called loss function predictor.[2]

From a Bayesian perspective the loss function predictor is sub-optimal. Standard analysis suggests to use the predictor that minimizes the posterior expected loss.

$$\hat{y}^*_{T+h|T} = \text{argmin}_{y^* \in \boldsymbol{R}^n} \int_{\theta \in \Theta} \left[\int_{y_{T+h}} L(y_{T+h}, y^*) p(y_{T+h}|\theta, \Upsilon_T) dy_{T+h} \right] p(\theta|\Upsilon_T) d\theta \qquad (3)$$

where $p(\theta|\Upsilon_T)$ is the posterior density of the parameter θ. Throughout the paper, we will restrict our attention to quadratic forecast error loss functions. If the loss function is quadratic, then the optimal predictor is a weighted average of parametric conditional expectations

$$\hat{y}^*_{T+h|T} = \int_{\theta \in \Theta} \left[\int_{y_{T+h}} y_{T+h} p(y_{T+h}|\theta, \Upsilon_T) dy_{T+h} \right] p(\theta|\Upsilon_T) d\theta \qquad (4)$$

[1] In slight abuse of notation, integrals with respect to various probability distributions are expressed through densities: $\int g(x) dP_\theta = \int g(x) p(x|\theta) dx$, where x can be a vector.

[2] The term is not quite precise. It is shorthand for loss function estimation based plug-in predictor.

The posterior distribution $p(\theta|\Upsilon_T)$ provides the optimal parameter weights. The posterior depends on the likelihood function and the prior distribution of the parameters θ, but is unrelated to the loss function $L(y_{T+h}, \hat{y}_{T+h|T})$. In large samples, the posterior distribution is concentrated in a neighborhood around the maximum likelihood estimator. Thus, Bayes predictor and maximum likelihood plug-in predictor are approximately equivalent. Despite a seemingly lack of optimality, loss function estimators are frequently used in practice and received considerable attention in the recent econometrics literature, e.g., Christoffersen and Diebold (1996, 1997), Weiss (1996), Tsay *et al.* (1993, 1994, 1996).

In the context of h-step ahead forecasting the loss function estimators are also called multi-step or dynamic estimators. The properties of such estimators have been examined, for instance, by Findlay (1983), Weiss and Andersen (1984), and Weiss (1991). These authors consider cases where the expectation of a random variable y_{t+h}, conditional on information available up to time t, is misspecified and conclude that, if the misspecification is substantial or the sample size is small, predictors based on multi-step estimation may be preferable. Weiss (1991) provides conditions under which multi-step estimators are consistent and asymptotically normal. More general convergence results are reported in Findlay *et al.* (1998). Weiss (1991) conducts a series of Monte Carlo experiments in which he compares the efficiency of estimators based on single- and multi-step estimation for various kinds of misspecification. The Monte Carlo analysis has been extended by Clemens and Hendry (1998), who are generally skeptical about the benefits of loss function estimation. Tsay (1993), Tiao and Tsay (1994), and Lin and Tsay (1996) refer to loss function estimation as adaptive forecasting and document the performance of such procedures in various applications. Their assessment is more favorable to loss function estimation.

The existing literature demonstrates that the benefits of a loss function estimation approach hinge on the potential misspecification of the forecasting model, in particular, the conditional expectation of y_{t+h} given past data Υ_T. The derivation of the Bayes predictor (3) does not take such a misspecification into account. To capture potential misspecification it is important to distinguish between a candidate model, that is used to compute the forecasts, and a reference model, from which the data are assumed to be generated.

This paper considers the problem of multi-step forecasting with a vector autoregression (VAR) under quadratic prediction error loss. A stationary moving average process of infinite order will serve as a reference model. It is a plausible strategy to base the choice between loss function and Bayes predictor upon the assessment of the candidate model's adequacy. A negative outcome of the model check indicates misspecification and suggests to use the loss function estimator. In Section 2 of the paper, we propose a model check based on the difference between the maximum likelihood and the loss function estimator of the VAR parameters in the spirit of Hausman (1978). The linear process theory of Phillips and Solo (1992) is used to obtain an asymptotic approximation to the expected prediction loss of this pre-test procedure. The performance of the pre-test predictor is compared to the loss function and the Bayes predictor. If the loss differential between the two predictors is large, then the pre-test avoids choosing the inferior predictor. If both predictors perform about equally well, then the forecaster is worse off by pre-testing. The asymptotic results are confirmed in a small Monte Carlo study in Section 3. The experiments closely resemble the ones conducted by Weiss (1991).

In Section 4 we propose a Bayesian interpretation of loss the function predictor and the pre-test procedure. One of the conceptual difficulties associated with the frequentist analysis of prediction losses under misspecification is that the ranking of predictors depends on the specific parametrization of the reference model. These parameters are unknown to the forecaster, yet he has to choose a predictor at time T. One popular solution to this problem, reflected in the papers by Tsay *et al.* and a statement by Granger (1993), is the minimax solution. If the misspecification of the candidate model is believed to be potentially severe, the loss function predictor is preferable to the maximum likelihood plug-in or the Bayes predictor under worst case assumptions for the reference model. Hence, the forecaster should always choose the loss function predictor.

A Bayesian framework offers an alternative solution. Consider a forecaster who places non-zero prior probability on the candidate model as well as the reference model. However, the forecaster finds it too onerous to evaluate posterior predictive distributions based on the latter. In practice, forecasters often choose simple candidate models, such as linear vector autoregressions, instead of more complicated specifications, such as vector autoregressive moving average (VARMA) models. The former can be easily analyzed with standard econometric software packages. Moreover, even if the candidate model is very sophisticated the forecaster might still be concerned about misspecification and tacitly consider a more general specification as reference.

A forecaster who faces the above constraint, is not able to calculate the predictor that minimizes the overall expected loss conditional on the observed data. However, the prior distribution enables averaging over the different parametrizations of the reference model. Thus, the forecaster is able to evaluate and rank prediction procedures *a priori* based on their integrated risk. The pre-test is called a model check in the Bayesian literature, see Gelman *et al.* (1995). It is illustrated how the optimal calibration of the model check depends on the prior distribution placed on candidate and reference model. Vice versa, the willingness to accept a certain rejection level for the test can be interpreted as an indicator for an implicit prior distribution. Section 5 concludes and the Appendix provides derivations of the formulae that appear in the main text.

2 Multi-Step Forecasting of a Linear Process

2.1 Notation and Setup

The forecaster uses a time series $\{y_t\}_{t=1}^{T}$ to compute parameter estimates for the candidate models. We will assume that the process to be predicted, $\{\tilde{y}_t\}_{t=T+h}^{T}$, is independent of $\{y_t\}$, but otherwise has exactly the same probabilistic structure. This assumption has often been made for mathematical convenience in studies of the present nature. Throughout this section, the random quantity to be predicted by a point forecast is $\varphi = \tilde{y}_{T+h}$. The loss function is assumed to be quadratic. Let $tr[\cdot]$ denote the trace operator.

Assumption 1 (Loss Function) *The forecasts are evaluated under the quadratic prediction error loss function*

$$L(\varphi, \hat{\varphi}) = tr[W(\varphi - \hat{\varphi})(\varphi - \hat{\varphi})']$$

W is a symmetric and positive definite weight matrix. □.

The candidate model is comprised of p'th order Gaussian VARs of the form

$$\mathcal{M}: \quad y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma) \tag{5}$$

where y_t is an $n \times 1$ vector and $\phi_1, \ldots, \phi_p$ are $n \times n$ matrices. The VAR coefficients are collected in the matrix $\phi = [\phi_1, \ldots, \phi_p]$. The presence of correctly modeled deterministic components does not affect our conclusions in any substantive way. Hence, we will proceed as if they are absent just to keep derivations as simple as possible. Let O_n denote an $n \times n$ matrix of zeros, and I_n an $n \times n$ identity matrix. To express the VAR in companion form we introduce the following additional notation:

$$Y_t = \begin{bmatrix} y_t \\ \vdots \\ y_{t-p+1} \end{bmatrix}, \quad \Phi = \begin{bmatrix} \phi_1 & \cdots & \phi_{p-1} & \phi_p \\ I_n & \cdots & 0_n & 0_n \\ \vdots & \ddots & \vdots & \vdots \\ 0_n & \cdots & I_n & 0_n \end{bmatrix}, \quad M_n = \begin{bmatrix} I_n \\ 0_n \\ \vdots \\ 0_n \end{bmatrix}, \quad E_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus,

$$Y_t = \Phi Y_{t-1} + E_t \tag{6}$$

where $y_t = M_n' Y_t$. The VAR parameters ϕ are related to the companion form parameters Φ according to $\phi = M_n' \Phi$. Let $\theta \in \Theta$ be the generic parameter vector for the candidate model that consists of the non-redundant elements of Φ and Σ. The conditional likelihood function is of the form

$$p(\Upsilon_T|\theta, Y_0) \propto |\Sigma|^{-T/2} \exp\left\{ -\frac{1}{2} tr \left[\sum_{t=1}^{T} \Sigma^{-1}(M_n' Y_t - \phi Y_{t-1})(M_n' Y_t - \phi Y_{t-1})' \right] \right\} \tag{7}$$

Υ_t denotes the sample $y_1, \ldots, y_T$, and Y_0 contains the initialization for the lags $y_{1-p}, \ldots, y_0$. The likelihood function is maximized at

$$\begin{aligned} \hat{\phi}_{T,b} &= M_n' \hat{\Phi}_{T,b} = M_n' \left(\sum Y_t Y_{t-1}' \right) \left(\sum Y_{t-1} Y_{t-1}' \right)^{-1} && (8) \\ \hat{\Sigma}_T &= \frac{1}{T} \sum M_n' (Y_t - \hat{\Phi}_{T,b} Y_{t-1})(Y_t - \hat{\Phi}_{T,b} Y_{t-1})' M_n && (9) \end{aligned}$$

For completeness, we will also specify a prior distribution for the parameters.

Assumption 2 (Candidate Model) *The candidate model from which all forecasts are derived is a Gaussian vector autoregression of order p, specified in Equation (5). The prior distribution of θ concentrates on a compact subset Θ and has non-zero density $p(\theta)$ in the stationary regions of the parameter space for which the largest eigenvalue of Φ is less than one in absolute value.* □

Under the quadratic prediction error loss function the Bayes predictor, derived from the candidate model is equal to the posterior predictive mean of $\tilde{y}_{T+h}$

$$\varphi_b = \int_{\Theta} M_n' \Phi^h \tilde{Y}_T p(\Phi, \Sigma | \Upsilon_T, \mathcal{M}) d\Phi d\Sigma \tag{10}$$

where $p(\Phi, \Sigma|\Upsilon_T, \mathcal{M})$ denotes the posterior density of the VAR parameters conditional on the data. In large samples, the posterior distribution concentrates around the maximum of the likelihood function and the Bayes predictor is approximately equal to the maximum likelihood plug-in predictor. Throughout the paper we will treat the two predictor as asymptotically equivalent. This can be formally expressed as follows.

Lemma 1 *Suppose a time series $\{y_t\}_{t=1}^T$ is generated from the VAR specified in Equation (5) or the reference model specified in Equation (11) below, then the Bayes predictor φ_b can be approximated as follows*

$$\sqrt{T}|\hat{\varphi}_b - M_n'\hat{\Phi}_T^h\tilde{Y}_T| \xrightarrow{p} 0 \quad \square$$

Proof: The Lemma can be proved by Laplace approximation of the posterior mean of Φ^h, see for instance, Crowder (1988). Details are omitted. □

Under the reference model $\mathcal{M}_*$, y_t is the sum of a $VAR(p)$ process x_t and a disturbance process z_t. For any fixed sample size T, it is simply a $VMA(\infty)$ process. Wold's theorem implies that the class of $VMA(\infty)$ processes encompasses a large collection of stationary processes and hence comprises a reasonable reference class. Let $||A|| = (tr[A'A])^{1/2}$.

Assumption 3 (Reference Model) *The reference model is a drifting vector moving average process of the form*

$$\mathcal{M}_*: \quad y_t = x_t + \alpha T^{-1} z_t \tag{11}$$

where

$$x_t = \sum_{j=0}^{\infty} M_n' F^j M_n \epsilon_t, \quad z_t = \sum_{j=0}^{\infty} M_n A_j M_n u_t$$

The matrices F and A_j are $np \times np$. The matrix F is in companion form and has the same structure as Φ. The largest eigenvalue of F is less than one in absolute value. The sequence $\{A_j\}_{j=0}^{\infty}$ is summable in the sense $\sum_{j=0}^{\infty} j^2 ||A_j|| < \infty$. The innovations ϵ_t and u_t are independent across time and jointly distributed with variances $\Sigma_{\epsilon\epsilon}$ and Σ_{uu}, covariance matrix $\Sigma_{\epsilon u}$, and finite forth moments: $||\mathbb{E}_*[(\epsilon_t\epsilon_t')(\epsilon_t\epsilon_t')']|| < \infty$, $||\mathbb{E}_*[(u_t u_t')(u_t u_t')']|| < \infty$. □

Conditional on a disturbance process x_t, the parameter α controls the size of the misspecification. The drift term $T^{-1/2}$ ensures that the trade-off between bias and efficiency of the different estimation procedures does not vanish as the sample size tends to infinity. The assumption that the linear process innovations are identical distributed and independent across time is stronger than necessary for the subsequent analysis. For alternative assumptions see Phillips and Solo (1992). Let ψ denote the generic parameter vector of the reference model that contains the non-redundant elements of the matrices $F, \Sigma_{uu}, \Sigma_{\epsilon\epsilon}, \Sigma_{\epsilon u}, A_0, A_1, \ldots$. The discussion of a prior distribution for the parameters of the reference model is deferred to Section 4.

The following additional notation is introduced. The companion form version of x_t and z_t is denoted by $X_t = \sum_{j=0}^{\infty} F^j E_{t-j}$, $Z_t = \sum_{j=0}^{\infty} A_j U_{t-j}$, where $U_t = [u_t', 0, \ldots, 0]'$. Generally, upper case letters refer to the companion form representation. For instance, autocovariances are denoted as

$$\begin{aligned}
\Gamma_{xx}(h) &= M_n' \Gamma_{XX}(h) M_n = M_n' \left(\sum_{j=0}^{\infty} F^{j+h} \Sigma_{EU} F_j' \right) M_n \\
\Gamma_{xz}(h) &= M_n' \Gamma_{XZ}(h) M_n = M_n' \left(\sum_{j=0}^{\infty} F^{j+h} \Sigma_{EU} A_j' \right) M_n
\end{aligned}$$

We will now calculate expected quadratic prediction error losses conditional on the parametrization of the reference model. This expected loss is the frequentist prediction risk.

2.2 Risk Calculations for Bayes and Loss Function Predictor

2.2.1 Pseudo-true Values and Loss Function Estimation

At first we will define pseudo-true values for the vector autoregression based on the h-step ahead forecasting problem. The concept of pseudo-true values has been widely used in the econometrics. Our definition is most closely related to one used in the indirect inference literature, e.g., Gourieroux *et al.*, (1993). If the time series is generated from a linear process with parametrization ψ, then the optimal predictor is $\hat{\varphi}_\psi = \mathbb{E}^*_T[M_n'\tilde{Y}_{T+h}]$, where $\mathbb{E}^*_T$ denotes the expectation under $\mathcal{M}_*$ conditional on time T and infinitely many past observations. The expected loss of $\hat{\varphi}_\psi$ provides a lower bound for the frequentist risk of estimators.

We normalize the prediction risk $R(\hat{\varphi}|\psi, \mathcal{M}_*)$ of a predictor $\hat{\varphi}$ as follows

$$\begin{aligned} R(\hat{\varphi}|\psi, \mathcal{M}_*) &= \mathbb{E}^*\Big[tr[W(\varphi - \hat{\varphi})(\varphi - \hat{\varphi})']\Big] - \mathbb{E}^*\Big[tr[W(\varphi - \hat{\varphi}_\psi)(\varphi - \hat{\varphi}_\psi)']\Big] \\ &= \mathbb{E}^*\Big[tr[W(\varphi_\psi - \hat{\varphi})(\varphi_\psi - \hat{\varphi})']\Big] \geq 0 \qquad (12) \end{aligned}$$

The relative ranking of predictors is not affected by this normalization.

Definition 1 *The pseudo-true parameter vector θ_T of the VAR(p) for the problem of predicting a process generated from the reference model $\mathcal{M}_*$ h periods into the future is given by the solution of the minimization problem*

$$\min_{\theta \in \Theta} R(M_n'\Phi_T^h(\theta)\tilde{Y}_T|\psi, \mathcal{M}_*) \quad \square$$

Using tedious but straightforward algebra it can be shown that the pseudo-true autoregressive parameters have the companion form representation[3]

$$\Phi_{T,l}^h = F^h + T^{-1/2}\alpha\mu_l + o(T^{-1/2}) \qquad (13)$$

[3] Since the predictor $M_n'\Phi^h Y_T$ does not depend on the variance parameters Σ, the pseudo-true value is, strictly speaking, not uniquely defined. However, this non-uniqueness is irrelevant for the prediction problem.

where

$$\mu_l = [\Gamma_{ZX}(h) - F^h \Gamma_{ZX}(0)]\Gamma_{XX}(0)^{-1} \tag{14}$$

The matrix norm of the remainder term converges to zero faster than $T^{-1/2}$.

As pointed out in the previous Section, the loss function or multi-step estimator is designed to obtain an estimate of the pseudo-true value $\Phi^h_{T,l}$. The loss function estimator of Φ^h is of the form

$$\hat{\Phi}^h_{T,l} = M_n' \Big(\sum Y_t Y_{t-h}' \Big) \Big(\sum Y_{t-h} Y_{t-h}' \Big)^{-1} \tag{15}$$

and does not depend on the weight matrix W. The corresponding h-step plug-in predictor is $\hat{\varphi}_l = M_n' \hat{\Phi}^h_{T,l} \tilde{Y}_T$.

Loss function estimation procedures are based on the idea that in a large sample the observed frequencies of hypothetical prediction losses at times $t < T$ are a reliable indicator for the frequentist risk associated with different predictors. Granger (1993), for instance, proposes that if we believe that a particular criterion should be used to evaluate forecasts, then it should also be used at the estimation stage of the modeling process. More formally, loss function estimators are designed to converge under mild regularity conditions to the "true" parameter value θ if the data are generated from the candidate model, and otherwise to the pseudo-true parameter value θ_0, defined above.[4]

2.2.2 Limit Distribution of the Estimators

The asymptotic theory for linear processes developed by Phillips and Solo (1992) can be used to derive the limit distribution of the approximation to the Bayes estimator $M_n' \hat{\Phi}^h_{T,b}$ and the loss function estimator $M_n' \hat{\Phi}^h_{T,l}$. The limit distribution is used to calculate the contribution of the parameter uncertainty to the prediction risk. Define

$$\Phi^h_{T,b} = F^h + \alpha T^{-1/2} \mu_b \tag{16}$$

[4] If the sample average of the observed prediction loss does not converge to the frequentist prediction risk, then loss function estimation is difficult to interpret. This paper focuses on stationary models for which the convergence occurs.

where

$$\mu_b = \left[\sum_{k=0}^{h-1} F^{h-k-1}[\Gamma_{XZ}(1) - F\Gamma_{ZX}(1)]\Gamma_{XX}(0)^{-1}F^k\right] \tag{17}$$

Proposition 1 *Suppose the time series Y_T is generated from the reference model specified in Equation (11) and Assumption 3. Then*[5]

$$\sqrt{T}\begin{bmatrix} M_n'(\hat{\Phi}^h_{T,b} - F^h) \\ M_n'(\hat{\Phi}^h_{T,l} - \hat{\Phi}^h_{T,b}) \end{bmatrix} \Longrightarrow \mathcal{N}\left(\begin{bmatrix} \alpha M_n'\mu_b \\ \alpha M_n'(\mu_l - \mu_b) \end{bmatrix}, \begin{bmatrix} V_b^0 & 0 \\ 0 & V_l^0 - V_b^0 \end{bmatrix}\right) \tag{18}$$

where

$$\begin{aligned} V_b^0 &= \sum_{k=0}^{h-1}\sum_{l=0}^{h-1} M_n' F^k \Sigma_{EE} F^{l'} M_n \otimes [F^{h-k-1'}\Gamma_{XX}(0)^{-1}F^{h-l-1}] \\ V_l^0 &= \sum_{k=0}^{h-1}\sum_{l=0}^{h-1} M_n' F^k \Sigma_{EE} F^{l'} M_n \otimes [\Gamma_{XX}(0)^{-1}\Gamma_{XX}(k-l)\Gamma_{XX}(0)^{-1}] \quad \Box \end{aligned}$$

The proposition implies that even under the considered misspecification the asymptotic distribution of the inefficient loss function estimator can be expressed as the sum of the limit distribution of the likelihood based estimator and an uncorrelated random variable. It is centered at the pseudo-true value $M_n'\Phi^h_{T,l}$ and has variance $V_l^0 \geq V_b^0$. The sampling distribution of the likelihood based estimator $M_n\hat{\Phi}^h_{T,b}$ is centered at $M_n'\Phi^h_{b,T} \neq M_n'\Phi^h_{l,T}$. Since

$$M_n'\Phi^h_{T,b} - M_n'\Phi^h_{T,l} = \alpha T^{-1/2} M_n'(\mu_b - \mu_l) \tag{19}$$

the discrepancy between the location of the sampling distributions of the two estimators is an increasing function of the parameter α.

2.2.3 Frequentist Prediction Risk

The subscript $\iota = b, l$ will be used to index the Bayes and loss function predictor and the corresponding risk. Equation (12) and the properties of a quadratic loss

[5] The notation is shorthand for: $\sqrt{T}vech[M_n'(\hat{\Phi}^h_{T,b} - F^h)] \Longrightarrow \mathcal{N}(\alpha vech[M_n'\mu_b], V_b^0)$ where $vech$ denotes the operator that stacks the rows of a matrix.

function imply

$$R(\hat{\varphi}_\iota|\psi,\mathcal{M}_*) = \mathbb{E}^*\left[tr[WM_n'(\mathbb{E}_T^*\tilde{Y}_{T+h} - \hat{\Phi}_{T,\iota}^h\tilde{Y}_T)(\mathbb{E}_T^*\tilde{Y}_{T+h} - \hat{\Phi}_{T,\iota}^h\tilde{Y}_T)'M_n]\right] \quad (20)$$

The normalization removes the portion of the forecast error loss that is due to the randomness of the $\tilde{y}_T$ process and does not vanish in large samples. The effects of parameter uncertainty and model misspecification are both of order T^{-1} and determine the asymptotic behavior of the normalized risk. The asymptotic prediction risk can be defined as $\bar{R}(\cdot|\cdot) = \lim_{T\to\infty} T \cdot R(\cdot|\cdot)$.

The risk can be decomposed as follows

$$\begin{aligned} R(\hat{\varphi}_\iota|\psi,\mathcal{M}_*) \quad = \quad & \mathbb{E}^*\left[tr[WM_n'(\mathbb{E}_T^*\tilde{Y}_{T+h} - \Phi_{T,\iota}^h\tilde{Y}_T)(\mathbb{E}_T^*\tilde{Y}_{T+h} - \Phi_{T,\iota}^h\tilde{Y}_T)'M_n]\right] \\ & + \mathbb{E}^*\left[tr[WM_n'(\hat{\Phi}_{T,\iota}^h\tilde{Y}_T - \Phi_{T,\iota}^h\tilde{Y}_T)(\hat{\Phi}_{T,\iota}^h\tilde{Y}_T - \Phi_{T,\iota}^h\tilde{Y}_T)'M_n]\right] \quad (21) \\ & - 2\mathbb{E}^*\left[tr[WM_n'(\mathbb{E}_T^*\tilde{Y}_{T+h} - \Phi_{T,\iota}^h\tilde{Y}_T)(\hat{\Phi}_{T,\iota}^h\tilde{Y}_T - \Phi_{T,\iota}^h\tilde{Y}_T)'M_n]\right] \end{aligned}$$

The first term captures the risk attained by the plug-in predictors $M_n\Phi_{T,\iota}\tilde{Y}_T$. For $\alpha = 1$ its limit will be denoted by R_ι^*

$$R_\iota^* = \lim_{T\to\infty} T \cdot \mathbb{E}^*\left[tr[WM_n'(\mathbb{E}_T^*Y_{T+h} - \Phi_{T,\iota}^hY_T)(\mathbb{E}_T^*Y_{T+h} - \Phi_{T,\iota}^hY_T)'M_n]\right] \quad \alpha = 1$$

Clearly, $R_l^* \leq R_b^*$, since $M_n'\Phi_{T,l}^h$ was defined as the set of VAR(p) coefficients that minimizes the frequentist risk. The second term can be approximated by the variance of the limit distribution of $M_n\hat{\Phi}_{T,\iota}^h$. The third term vanishes asymptotically because the limit distribution of $M_n\hat{\Phi}_{T,\iota}^h$ has expected value $M_n\Phi_{T,\iota}^h$.

Proposition 2 *As the sample size tends to infinity, the frequentist risk $R(\hat{\varphi}_\iota|\psi,\mathcal{M}_*)$ for $\iota = b, l$ converges to:*

$$\bar{R}(\hat{\varphi}_\iota|\psi,\mathcal{M}_*) = \alpha^2 R_\iota^* + tr[(W \otimes \Gamma_{xx}(0))V_\iota^0] \quad (22)$$

where

$$\begin{aligned} R_\iota^* \quad = \quad & tr\left[WM_n'\mu_\iota\Gamma_{XX}(0)\mu_\iota'M_n\right] - 2tr[WM_n'\mu_l\Gamma_{XX}(0)\mu_\iota'M_n] \\ & + tr\left[WM_n'\left(\sum_{j=0}^{\infty}(A_{j+h} - R^hA_j)\Sigma_{UU}(A_{j+h} - R^hA_j)'\right)M_n\right] \quad \square \quad (23) \end{aligned}$$

Since $R_l^* \leq R_b^*$ and $V_l^0 \geq V_b^0$ in the matrix sense, there exists a trade-off between the two predictors. Based on the inequality

$$\alpha^2 \geq tr[(W \otimes \Gamma_{XX}(0))(V_l - V_b)]/(R_b^* - R_l^*) \tag{24}$$

we can distinguish two cases:

(i) The misspecification of the candidate model is small in the sense that Inequality (24) not satisfied for any parameter value in the support of the prior distribution. In this case, the candidate model $\mathcal{M}$ approximates the reference model $\mathcal{M}_*$ well enough in terms of h-step ahead prediction under quadratic loss the Bayes predictor is preferable.

(ii) The misspecification of the candidate model is severe in the sense that Inequality (24) is satisfied. In this case the loss function predictor $\hat{\varphi}_l$ dominates the Bayes predictor.

Since the reference model is approaching the candidate model at rate $T^{-1/2}$, the choice between the loss function estimator and Bayes estimator does not depend on the sample size. The trade-off between the efficiency of the likelihood based estimator and the bias toward the pseudo-true value of the loss function estimator does not vanish. In practice, the sample size is fixed and the choice of predictor depends ultimately on the conjectured misspecification of the candidate model relative to the available amount of data. This is captured by the weight α. The forecaster does not know the parameters of the reference model and cannot determine which predictor has the smallest expected loss. A plausible procedure is to assess the adequacy of the candidate model before choosing between the Bayes and the loss function predictor.

2.3 A Prediction Rule Based on Model Checking

The previous analysis showed that the divergence of $M_n' \hat{\Phi}_{T,b}^h$ and $M_n' \hat{\Phi}_{T,l}^h$ is an indicator for misspecification of the $\mathcal{M}$ model. If the misspecification is severe, it is preferable to use the loss function predictor $\hat{\varphi}_l$. Although this type of selection rule does not take the contribution of the variance terms V_l^0 and V_b^0 to the asymptotic frequentist risk into account, it seems intuitively reasonable because the gain from choosing $\hat{\varphi}_l$ is large if the misspecification is substantial and the Hausman-type test is powerful. Define the difference $\hat{D}_T \equiv \hat{\Phi}^h_{T,b} - \hat{\Phi}^h_{T,l}$. Let $\hat{V}_{T,b}$ and $\hat{V}_{T,l}$ denote estimators of V_b^0 and V_l^0, respectively, that are consistent under the candidate model $\mathcal{M}$. Define the checking function

$$g(\Upsilon_T) = T \cdot \hat{D}_T(\hat{V}_{T,L} - \hat{V}_{T,b})^{-1}\hat{D}_T' \tag{25}$$

One can deduce from Proposition 2 and the Continuous Mapping Theorem, that the checking function is in large samples approximately χ^2 distributed with np^2 degrees of freedom, if data are generated from the candidate model $\mathcal{M}$. Consider the following pre-test prediction procedure.

Procedure 1 *Choose the loss function based predictor $\hat{\varphi}_l$ if $g(\Upsilon_T) > c^2$ and the Bayes predictor $\hat{\varphi}_b$ otherwise. The resulting predictor is denoted by $\hat{\varphi}_c$.*

Define $V_d = V_l^0 - V_b^0$ and $\hat{V}_{T,d} = \hat{V}^0_{T,l} - \hat{V}^0_{T,b}$. Moreover, define the integrals

$$\begin{aligned} I_{(0)}(c^2, m) &= (2\pi)^{-1/2} \int_{(x-m)'(x-m) \geq c^2} e^{-x'x/2} dx \\ I_{(0)}(c^2, m) &= (2\pi)^{-1/2} \int_{(x-m)'(x-m) \geq c^2} xx' e^{-x'x/2} dx \end{aligned}$$

where m is a vector with the same dimension as x. It is easily seen that $I_{(0)}(0, m) = 0$, $I_{(0)}(\infty, m) = 1$, $I_{(2)}(0, m) = 0_{pn^2}$, and $I_{(2)}(\infty, m) = I_{pn^2}$. The following proposition provides the asymptotic risk $\bar{R}(\hat{\varphi}_c|\psi, \mathcal{M}_*)$ and $\bar{R}(\hat{\varphi}_c|\theta, \mathcal{M})$.

Proposition 3 *The asymptotic frequentist risk of predictions according to Procedure 1 is*

$$\begin{aligned} \bar{R}(\hat{\varphi}_c|\psi, \mathcal{M}_*) &= \alpha^2 R_b^* + tr[(W \otimes \Gamma_{xx}(0))V_b^0] \\ &\quad + \alpha^2 (R_l^* - R_b^*) I_{(0)}(c^2, \alpha V_d^{-1/2}(\mu_b - \mu_l)) \\ &\quad + tr[(W \otimes \Gamma_{xx}(0)) V_d^{1/2} I_{(2)}(c^2, \alpha V_d^{-1/2}(\mu_b - \mu_l)) V_d^{1/2'}] \quad \square \end{aligned} \tag{26}$$

Suppose that y_t is a univariate time series and the candidate model is AR(1). For this case the integrals that appear in the formula of Proposition 3 simplify to

$$\begin{aligned} I_{(0)}(c^2, m) &= 1 - F_N(m+c) + F_N(m-c) \\ I_{(2)}(c^2, m) &= I_{(0)}(c^2, m) + (m+c) f_N(m+c) - (m-c) f_N(m-c) \end{aligned}$$

F_N and f_N denote the cumulative and probability density functions of a $\mathcal{N}(0,1)$ random variable, respectively. The term $I_{(0)}(c^2, m)$ corresponds to the power of the Hausman test. The overall risk of the pre-test predictor, setting $W = 1$, can be expressed as

$$\begin{aligned} \bar{R}(\hat{\varphi}_c|\psi, \mathcal{M}_*) &= [1 - I_{(0)}(c^2, m)]\bar{R}(\hat{\varphi}_b|\psi, \mathcal{M}_*) + I_{(0)}(c^2, m)\bar{R}(\hat{\varphi}_l|\psi, \mathcal{M}_*) \\ &\quad +(V_l^0 - V_d^0)[(m+c) f_N(m+c) - (m-c) f_N(m-c)] \qquad (27) \end{aligned}$$

where $m = \alpha V_d^{-1/2}(\mu_b - \mu_l)$. The first part of Equation (27) is a linear combination of the prediction risk of the Bayes and the loss function predictor. The weights for this linear combination are given by the power function of the model check procedure. It can be verified that the second part is non-negative and, thus, is a penalty for the testing pre-testing. While in some regions of the parameter space of the reference model, for instance, α very small, or α very large, the pre-test prediction procedure dominates either the loss function or the Bayes predictor, it is not guaranteed that

$$\min\{\bar{R}(\hat{\varphi}_b|\psi, \mathcal{M}_*), \bar{R}(\hat{\varphi}_l|\psi, \mathcal{M}_*)\} \le \bar{R}(\hat{\varphi}_c|\psi, \mathcal{M}_*) \le \max\{\bar{R}(\hat{\varphi}_b|\psi, \mathcal{M}_*), \bar{R}(\hat{\varphi}_l|\psi, \mathcal{M}_*)\} \qquad (28)$$

This is easily seen by choosing the parameters of the reference model such that $\bar{R}(\hat{\varphi}_b|\psi, \mathcal{M}_*) = \bar{R}(\hat{\varphi}_l|\psi, \mathcal{M}_*)$. Since $V_l^0 \ge V_b^0$ and $(m+c) f_N(m+c) - (m-c) f_N(m-c) \ge 0$, it follows that the pre-test predictor performs worse than both the Bayes and the loss function predictor.

3 Monte Carlo Experiment

A small Monte Carlo study is conducted to examine the finite sample properties of the pre-test predictor and how well the expected prediction losses are approximated by the asymptotic formulae derived in the previous section. The specification and parametrizations of the reference model resemble the ones used by Weiss (1991). Data are generated from the model

$$\begin{aligned} y_t &= \beta_1 y_{t-1} + \beta_2 y_{t-2} + \gamma_1 v_t + \eta_t \quad (29) \\ v_t &= \rho_1 v_{t-1} + \xi_t \quad (30) \end{aligned}$$

The disturbances η_t and ξ_t are *iid* $\mathcal{N}(0,1)$ and independent of each other. Table 1 lists the parameter values and the model identification numbers used by Weiss (1991). Throughout this section, all forecasts are generated from an AR(1) model.

To apply the asymptotic formulae y_t has to be decomposed in the processes x_t and z_t. This decomposition is clearly not unique but it will have an effect on the goodness of the asymptotic approximation. The process y_t is decomposed such that the variance of the misspecification process z_t is minimized. It can be seen in the Appendix that the autocovariances of z_t enter the calculations with a factor T^{-1}, whereas the cross covariances between x_t and z_t enter with a factor $T^{-1/2}$ and dominate asymptotically. The adopted decomposition is not necessarily the most favorable one to the asymptotic approximation. However, it is a reasonable decomposition because the size of higher order autocovariances of z_t is bounded in absolute value by the variance of z_t.

From Table 1 it can be seen that model specifications $M4$ to $M8$ correspond to AR(2) models. The AR(2) model has a moving average representation of the form $y_t = \sum_{j=0}^{\infty} c_j(\beta_1, \beta_2)\eta_{t-j}$. The autoregressive coefficient F of the x_t process is chosen to minimize $\sum_{j=0}^{\infty}(c_j(\beta_1, \beta_2) - F^j)^2$. The MA representation of models $M9$ to $M12$ is of the form $y_t = \sum_{j=0}^{\infty} \beta_1^j \eta_{t-j} + \gamma_1 \sum_{j=0}^{\infty} d_j(\beta_1, \rho_1)\xi_{t-j}$ where η_t and ξ_t are independent and *iid* $\mathcal{N}(0,1)$. The process $x_t = \sum_{j=0}^{\infty} F^j(w_1\eta_{t-j} + w_2\xi_{t-j})$ and F, w_1, and w_2 are to chosen to minimize the variance of the distortion z_t. It turns out that for most specifications of Weiss' (1991) study the variance of the misspecification process z_t is substantially smaller than the variance of the signal x_t.

The sample size is chosen to be $T = 100$, the forecast horizon is $h = 2, 4, 6$ and 8 periods ahead. The threshold for the pre-test is $c = 2$ which leads to approximately five percent of rejections if the data are generated from an autoregressive model of order one. The number of replications used to compute the expected prediction losses is 50,000. During each replication a sample of 300 observations is generated. The first 200 observations are discarded and the remaining 100 observations are used to calculate $I\!E^*_T[Y_{T+h}]$, $\hat{\Phi}^h_{T,b}Y_T$, $\hat{\Phi}^h_{T,l}Y_T$, and $g(\Upsilon_T)$. The results are tabulated in Tables 2 and 3. The columns contain Monte Carlo averages of the normalized prediction losses $(I\!E^*_T[Y_{T+h}] - \hat{\varphi}_\iota)^2$, and the asymptotic prediction risks $T \cdot \bar{R}(\hat{\varphi}_\iota|\psi, \mathcal{M}_*)$ for the Bayes predictor, the loss function predictor, and the pre-test predictor.

For $M1$, $M2$, and $M3$ the candidate AR(1) model is correctly specified. The loss function predictor performs clearly worse than the Bayes predictor. The asymptotic approximations for the loss function predictor are quite precise. Due to the power transformations of the parameter estimate, the discrepancy between asymptotic and finite sample risk is somewhat larger for the Bayes predictor. Due to the choice of the threshold c the candidate model passes the model check in about 95 percent of the iterations. The pre-test predictor clearly dominates the loss function predictor. As indicated by the asymptotic calculations, there is however, a small price to be paid for the pre-testing.

Model specifications $M4$ to $M8$ correspond to the five AR(2) models. Only under $M8$ the loss function predictor is clearly dominated by the Bayes predictor. Table 1 shows that the variance of the misspecification process z_t is less than one percent of the variance of x_t. The model check confirms the adequacy of the forecasting model. In particular, at horizons 6 and 8, the pre-test predictor avoids the large losses of the loss function predictor. Under specification $M6$ both loss function predictor and Bayes predictor perform about equally well. In this case, the variance of the distortion z_t is roughly two percent of the variance of x_t. Overall, the pre-test predictor performs worse than the other two predictors. This is consistent with the asymptotic theory developed in the previous section. Whenever the risks of the Bayes predictor and loss function predictor are approximately equal, the pre-test procedure leads to inferior predictions. At horizon 8, the asymptotic approximation is imprecise and somewhat misleading about the actual ranking of the three forecast rules. Conditional on $M4$, $M5$, and $M7$ the loss function predictor dominates the Bayes predictor. This is consistent with the fact that the relative variance of z_t is much larger than in the previous two cases. The pre-test predictor dominates the Bayes predictor. Under specification $M4$ it performs almost as well as the loss function predictor, under $M5$ and $M7$ the improvements are smaller.

Table 3 contains the results for model specifications $M9$ through $M12$. The difference between $M9$, $M10$, and $M11$ is the weight γ with which the exogenous process v_t enters the determination of y_t. If γ is small ($M9$) the Bayes predictor dominates the loss function predictor. If the weight is large, the ranking is reversed. As before, the pre-test procedure avoids choosing the inferior predictor if the discrepancy in performance between the Bayes and the loss function predictor is large. If the performance of the two predictors is roughly equal then the pre-test predictor performs worse than both of them. The Monte Carlo study confirms the analytical results obtained in the previous section.

Model	Weiss-ID	β_1	β_2	γ_1	ρ_1	F	w_1	w_2	$var[y_t]$	$var[x_t]$	$var[z_t]$
M1	(2)	0.8				0.80			2.78	2.78	
M2	(3)	0.5				0.50			1.33	1.33	
M3	(4)	0.2				0.20			1.04	1.04	
M4	(7)	1.6	-0.64			0.95			35.2	11.3	10.6
M5	(8)	1.3	-0.40			0.90			8.64	5.19	0.92
M6	(9)	1.0	-0.16			0.84			4.00	3.43	0.06
M7	(10)	1.0	-0.25			0.76			2.96	2.35	0.13
M8	(11)	0.7	-0.10			0.61			1.70	1.60	0.01
M9	(12)	0.8		0.5	0.5	0.84	0.90	0.78	4.94	4.77	0.17
M10	(13)	0.8		2.5	0.5	0.86	0.84	3.68	56.8	54.1	2.70
M11	(14)	0.8		5.0	0.5	0.86	0.83	7.26	218.8	208.6	10.2
M12	(15)	0.8		0.5	0.8	0.90	0.68	1.22	11.6	10.1	1.50

Table 1: Reference Model Parameters for Monte Carlo Experiment. Sample Size $T = 100$, rejection threshold for pre-test $c = 2$, weight matrix of loss function $W = 1$. Number of Monte Carlo replications is 50,000.

Model	Proc	Simulated Loss Forecast Horizon				Asymptotic Loss Forecast Horizon			
		2	4	6	8	2	4	6	8
M1	B	0.0248	0.0394	0.0379	0.0306	0.0256	0.0419	0.0387	0.0281
	L	0.0294	0.0655	0.0874	0.0990	0.0292	0.0680	0.0949	0.1105
	H	0.0262	0.0475	0.0521	0.0477	0.0265	0.0488	0.0534	0.0497
M2	B	0.0095	0.0029	0.0006	0.0001	0.0100	0.0025	0.0004	3.9E-5
	L	0.0169	0.0205	0.0208	0.0212	0.0175	0.0217	0.0222	0.0222
	H	0.0115	0.0074	0.0050	0.0047	0.0120	0.0075	0.0061	0.0058
M3	B	0.0017	4.5E-5	1.3E-6	4.7E-8	0.0016	1.0E-5	3.7E-8	1.E-10
	L	0.0110	0.0111	0.0110	0.0112	0.0112	0.0113	0.0113	0.0113
	H	0.0043	0.0028	0.0026	0.0026	0.0041	0.0030	0.0030	0.0030
M4	B	1.8914	4.0567	6.1981	8.6738	1.8993	4.0572	6.0807	8.2545
	L	1.9089	3.6236	4.1219	4.0277	1.8681	3.4204	3.8562	3.8723
	H	1.9093	3.6254	4.1287	4.0416	1.8681	3.4204	3.8562	3.8723
M5	B	0.3932	0.7132	1.0662	1.3150	0.4029	0.7462	1.0794	1.2612
	L	0.3884	0.4812	0.4829	0.4853	0.3770	0.4522	0.4673	0.5123
	H	0.3889	0.4862	0.5109	0.5683	0.3773	0.4544	0.4873	0.5971
M6	B	0.0630	0.1200	0.1534	0.1552	0.0667	0.1223	0.1423	0.1287
	L	0.0665	0.1139	0.1463	0.1641	0.0628	0.1078	0.1499	0.1814
	H	0.0663	0.1208	0.1577	0.1679	0.0659	0.1249	0.1611	0.1633
M7	B	0.1181	0.1816	0.1606	0.1033	0.1219	0.1785	0.1417	0.0788
	L	0.1047	0.0913	0.0874	0.0886	0.1022	0.0934	0.0979	0.1039
	H	0.1075	0.1190	0.1308	0.1082	0.1080	0.1307	0.1411	0.1046
M8	B	0.0267	0.0222	0.0092	0.0030	0.0271	0.0186	0.0060	0.0014
	L	0.0279	0.0326	0.0333	0.0341	0.0277	0.0344	0.0366	0.0371
	H	0.0287	0.0291	0.0164	0.0095	0.0289	0.0262	0.0154	0.0111

Table 2: Monte Carlo Results, Part (i). Simulated and Asymptotic Prediction Losses for Bayes Predictor (B), Loss Function Predictor (L), and Hausman-Test Procedure (H). Columns contain Monte Carlo averages of $(\mathbb{E}^*_T[Y_{T+h}] - \hat{\varphi}_\iota)^2$ and asymptotic risk $T \cdot \bar{R}(\hat{\varphi}_\iota|\psi, \mathcal{M}_*)$.

Model	Proc	Simulated Loss				Asymptotic Loss			
		Forecast Horizon				Forecast Horizon			
		2	4	6	8	2	4	6	8
M9	B	0.1568	0.1958	0.1914	0.1744	0.1554	0.2033	0.1946	0.1636
	L	0.1640	0.2171	0.2238	0.2274	0.1568	0.2100	0.2272	0.2430
	H	0.1604	0.2109	0.2144	0.2062	0.1567	0.2135	0.2206	0.2058
M10	B	2.5972	4.4110	6.2791	7.5328	2.4279	4.4889	5.9821	6.2902
	L	2.5912	3.2091	3.2080	3.2238	2.2809	2.8798	2.9592	3.0965
	H	2.5977	3.2641	3.4573	3.8488	2.3373	3.3600	4.2375	5.0460
M11	B	10.115	17.822	26.157	31.965	9.5088	18.172	24.897	26.707
	L	10.036	12.428	12.488	12.571	8.8726	11.186	11.503	12.075
	H	10.053	12.572	13.280	14.741	9.0739	12.847	16.266	19.922
M12	B	0.6801	1.2670	1.4892	1.5596	0.5991	1.1878	1.4829	1.5996
	L	0.6905	1.2741	1.3937	1.3021	0.5975	1.1250	1.2702	1.2307
	H	0.6871	1.2712	1.4078	1.3452	0.5997	1.1699	1.4090	1.4840

Table 3: Monte Carlo Results, Part (ii). Simulated and Asymptotic Prediction Losses for Bayes Predictor (B), Loss Function Predictor (L), and Hausman-Test Procedure (H). Columns contain Monte Carlo averages of $(I\!E^*_T[Y_{T+h}] - \hat{\varphi}_\iota)^2$ and asymptotic risk $T \cdot \bar{R}(\hat{\varphi}_\iota | \psi, \mathcal{M}_*)$.

4 A Bayesian Perspective

In Sections 2 and 3 we demonstrated that a Hausman type pre-test can be helpful to choose between the Bayes and the loss function predictor. The analysis also showed that there is still no clear ranking of the three forecast rules in terms of their frequentist prediction risk. We will now argue, that an *a priori* judgement about likely parameter values of the reference model can rationalize the choice of prediction procedure.

Consider a Bayesian forecaster who uses the candidate model to compute forecasts, although he believes that it is misspecified. More formally, he places non-zero prior probability on the reference model and its parameters, yet finds it too onerous to evaluate the posterior predictive distribution of the reference model. An example for such a reference model is a VARMA process. Posterior predictive distributions for VARMA models are much more difficult to compute than for vector autoregressions. This is reflected in the dominance of VARs in empirical applications. The non-zero prior probability of the reference model implies that the forecaster is aware of the possibility that the candidate model is inappropriate. It enables the forecaster to determine the potential behavior of statistical procedures applied to the candidate model, and how to translate frequencies of past forecasting losses into expected future forecasting losses.

Without a predictive distribution from the reference model, the forecaster is not able to calculate the predictor that minimizes the overall expected loss conditional on the observed data. However, the prior distribution enables averaging over the different parametrizations of the reference model. Thus, the forecaster is able to evaluate and rank prediction procedures *a priori* based on their integrated risk.

The pre-test can be interpreted as model checking. Box (1980) argued in favor of a sampling approach to criticize a statistical model in the light of the available data, say model $\mathcal{M}$ in the context of this paper. This model criticism then can induce model modifications. Although conceptually not undisputed, model checking and sensitivity analysis plays an important role in applied Bayesian statistics, see for instance the discussion in Gelman *et al.* (1995). However, unlike in many inferential situations, the nature of the forecasting problem requires a prediction at time T. It is not possible to simply reject model $\mathcal{M}$ and search for a better representation of the data. Yet it is possible to use the loss function predictor if the candidate model appears to be discredited.[6]

The general idea of model checking in a Bayesian framework is to evaluate the marginal density of the data under the entertained model $\mathcal{M}$, at the observed data. If the observations fall in a region of low density, then model $\mathcal{M}$ is discredited. In practice, this approach is often implemented through the evaluation of tail probabilities for a function of the data $g(\Upsilon_T)$. In our case the check function is given by Equation (25). If the distribution of $g(\Upsilon_T)$ converges in distribution to a χ^2 random variable, conditional on every parameter θ in the support of the prior distribution of the candidate model, then the marginal distribution of $g(\Upsilon_T)$ will also converge to a χ^2 distribution, and the pre-test in Procedure 1 can indeed be interpreted as Bayesian model check.

4.1 A Numerical Illustration

Suppose the candidate model is a simple autoregression of order one and the reference model is an ARMA(1,1) process of the form

$$(1 - FL)y_t = (1 + \alpha T^{-1/2} L)\epsilon_t \qquad (31)$$

where L denotes the lag operator and $\epsilon_t \sim iid\mathcal{N}(0,1)$. The reference model has prior probability one. The prior for F places uniform probability on the grid $0.05, 0.10, \ldots, 0.95$. The misspecification parameter α is equal to $\alpha_j = j/2$ with

[6] Consider the following approach to select one of the predictors $\hat{\varphi}_b$ and $\hat{\varphi}_l$. Define a third model $\mathcal{M}_l$ such that the likelihood function under $\mathcal{M}_l$ embeds the loss function $L(\varphi, \hat{\varphi})$ and the Bayes predictor under $\mathcal{M}_l$ leads to the loss function predictor $\hat{\varphi}_l$. The posterior odds ratio of $\mathcal{M}$ and $\mathcal{M}_l$ could then be used to determine which predictor to choose. Although plausible at first glance, it can be shown that this strategy will in large samples always lead to the selection of the Bayes predictor.

probability

$$\mathbb{P}\{\alpha = \alpha_j\} \propto \alpha_j^{\mu/2-1} e^{-\alpha_j/2} \tag{32}$$

for $j = 1, 2, \ldots, J_\alpha$ and $J_\alpha = 100$. Thus, the prior of α has the form of a discretized Gamma $\mathcal{G}(\mu/2, 2)$ distribution. The risk properties of the different predictors will be compared for various choices of μ.

Both R_ι and V_ι^0, which appear in the prediction risk formula of Proposition 2, are functions of the parameters of the reference model. The first step of our analysis is to average the terms R_ι and V_ι^0 with respect to the prior distribution of F.[7] The resulting marginal risk $\bar{R}(\hat{\varphi}_\iota|\alpha, \mathcal{M}_*)$ conditional on the misspecification parameter α is depicted in Figure 1(i), normalized by by $\bar{R}(\hat{\varphi}_b|\alpha, \mathcal{M}_*)$. Due to the normalization the risk of the Bayes predictor is equal to 100 for all α. If the misspecification parameter α is small, in particular $\alpha < 3$, then the Bayes predictor is preferable. For large misspecifications, the risk is minimized by the loss function predictor $\hat{\varphi}_l$. Figure 1(i) also shows the risk properties of the model check predictor for the rejection levels $c = 2$ and $c = 5$. The level $c = 2$ implies that in about five percent of the cases, the loss function estimator is chosen if the misspecification $\alpha = 0$. For $c = 5$, this frequency is almost zero. Since the model check procedure chooses the loss function predictor more often if c is small, $\bar{R}(\hat{\varphi}_{c=2}|\alpha, \mathcal{M}_*) < \bar{R}(\hat{\varphi}_{c=5}|\alpha, \mathcal{M}_*)$ for large values of α. The prior distribution of α and the absolute prediction loss $\bar{R}(\hat{\varphi}_b|\alpha, \mathcal{M}_*)$ are plotted in Figure 1(ii).

The risk $\bar{R}(\hat{\varphi}|\alpha, \mathcal{M}_*)$ can now be integrated with respect to the prior distribution of α to obtain $\bar{R}(\hat{\varphi}|\mathcal{M}_*) = \bar{R}(\hat{\varphi})$. Figure 2(i) shows the relative integrated risk $\bar{R}(\hat{\varphi}|\mathcal{M}_*)$ as a function of the parameter μ of the prior distribution of α. The relative risk is defined as

$$r(\hat{\varphi}, \mu) = 100 \frac{\bar{R}(\hat{\varphi}|\mathcal{M}_*)}{\min\{\bar{R}(\hat{\varphi}_l|\mathcal{M}_*), \bar{R}(\hat{\varphi}_b|\mathcal{M}_*)\}} \tag{33}$$

[7] If the prior distribution of the parameters is continuous, the exchange of integration and limit operation requires some uniformity in the convergence of the frequentist risk to its limit. One way to establish such uniformity is to limit the support of the prior distribution for Φ and F to a compact subset of the stationary region of the parameter space.

For the predictors $\hat{\varphi}_c$ the graph shows by how much the model checking procedure improves on the risk of the Bayes or loss function predictor. As the prior parameter μ increases, the probability of severe misspecification of model $\mathcal{M}$ rises and the loss function predictor $\hat{\varphi}_l$ has the best integrated risk properties. Correspondingly, if μ is small the Bayes predictor minimizes the integrated risk. In both cases, the model check becomes obsolete.

Figure 2(ii) depicts a contour plot of $r(\hat{\varphi}_c, \mu)$ as a function of the prior parameter μ and the rejection level c. The contour plot can be read in two ways. Suppose an econometrician is willing to specify a prior distribution for α by choosing a μ. Conditional on μ, the diagram enables us to determine the optimal rejection level. If $\mu = 1.3$ then $c_{opt} = \min_{c \in \boldsymbol{R}^+} \; r(\hat{\varphi}_c, \mu) \approx 2.8$. If μ is very small then c_{opt} is large (almost always use the Bayesian predictor) and as μ increases, c_{opt} decreases which implies to use the loss function predictor more often. However, many econometricians might be reluctant to choose a prior distribution through picking a value for μ and skeptical about the prior for ψ in the first place. Yet it is important to recognize that choosing a rejection level for the model check can be interpreted as an implicit prior distribution for the size of the misspecification α, and more generally for the parameters of the reference model under which the prediction procedure is sensible. A rejection level of, say, $c = 3$ is only reasonable if μ is neither very small, that is, the misspecification is negligible, or very large, that is, the misspecification is severe.

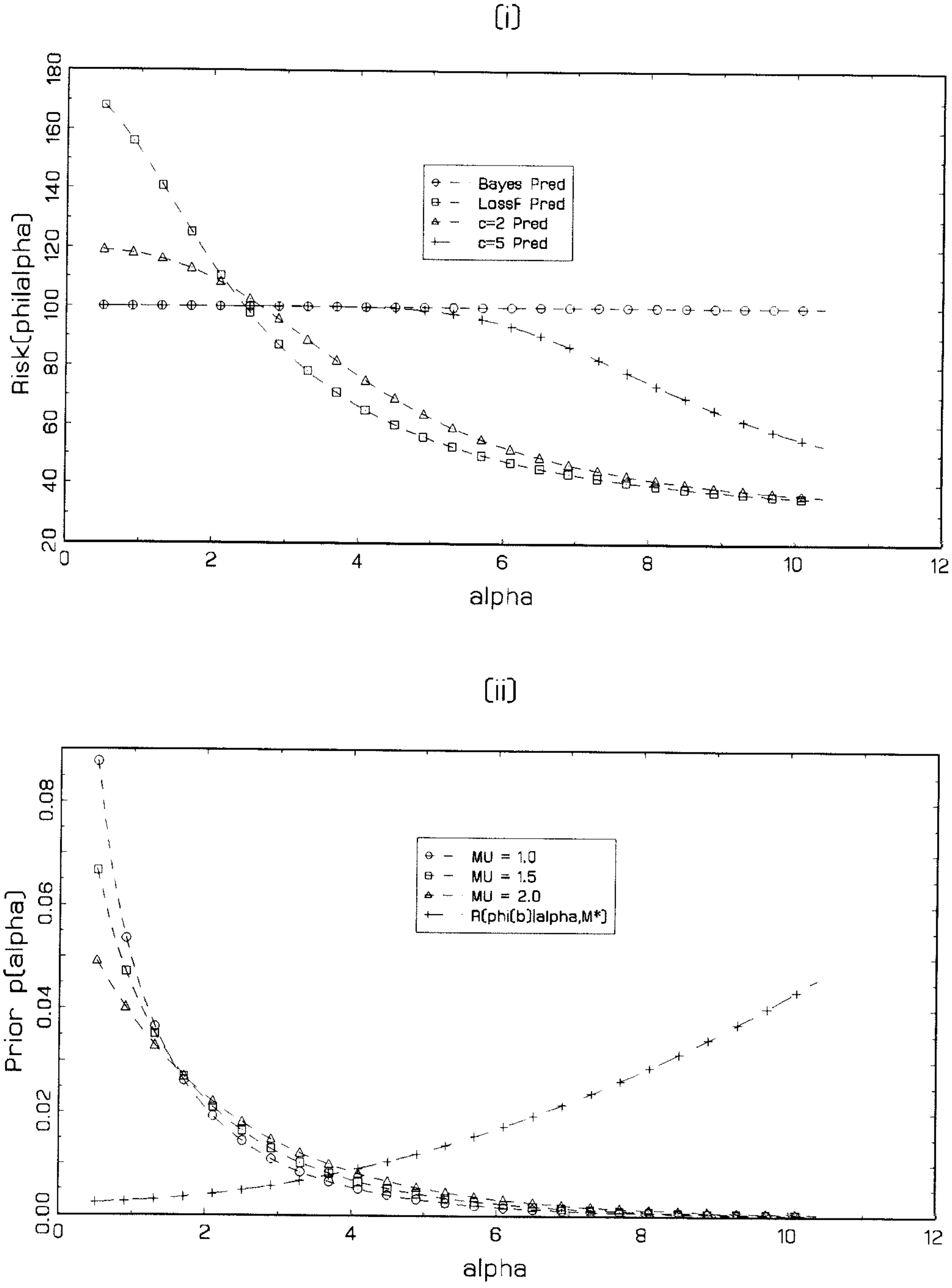

Figure 1: (i) Relative prediction risk $100 \cdot \bar{R}(\hat{\varphi}|\alpha, \mathcal{M}_*)/\bar{R}(\hat{\varphi}_b|\alpha, \mathcal{M}_*)$ for the predictors $\hat{\varphi}_l$, $\hat{\varphi}_b$, $\hat{\varphi}_{c=2}$, and $\hat{\varphi}_{c=5}$. (ii) Prior distribution $P\{\alpha = \alpha_j\}$ and absolute prediction risk $\bar{R}(\hat{\varphi}_b|\alpha, \mathcal{M}_*)/1000$.

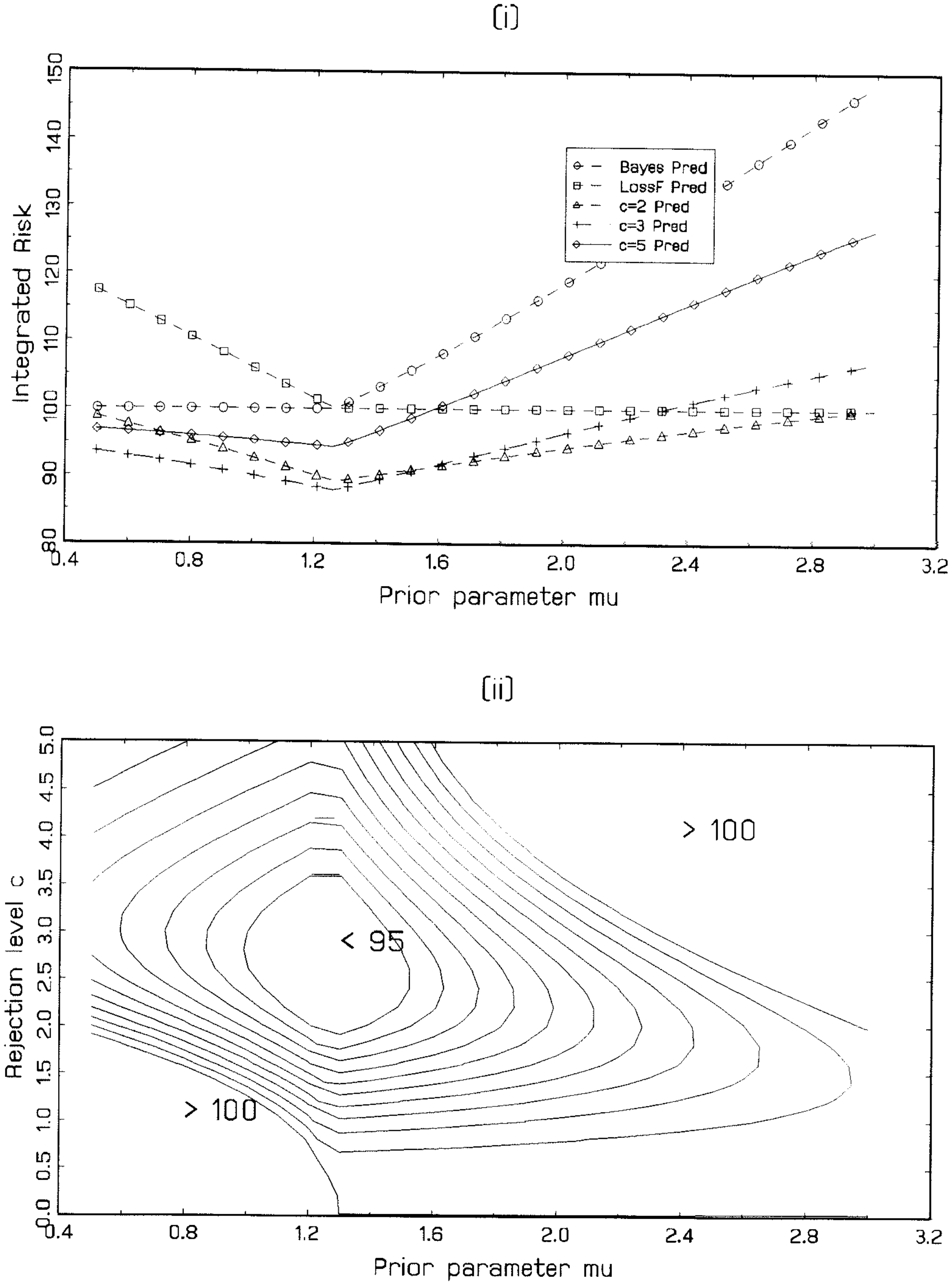

Figure 2: Normalized integrated risk $r(\hat{\varphi}, \mu)$ for the predictors $\hat{\varphi}_l$, $\hat{\varphi}_b$, $\hat{\varphi}_{c=2}$, $\hat{\varphi}_{c=3}$, and $\hat{\varphi}_{c=5}$. (ii) Contour plot of $r(\hat{\varphi}_c, \mu)$.

5 Conclusion

The paper considers the problem of multi-step ahead forecasting with a potentially misspecified VAR(p) model. It is well known that if the forecasting model is severely misspecified, parameters should be estimated by minimization of h-step in-sample forecasting errors. Under the assumption that a stationary time series is generated from a linear process that approaches the candidate model at rate $T^{-1/2}$ we calculate asymptotic forecast error losses for a loss function predictor and a Bayes, or maximum likelihood plug-in, predictor. The results illustrate that the ranking of the two predictors is not only determined by whether or not the candidate model is misspecified, but rather by the size of the misspecification relative to the available sample size. An easy to implement procedure to choose between the two predictors is to compare the discrepancy of the loss function based parameter estimate and the Bayes or maximum likelihood estimate. If the discrepancy is large, then the candidate model appears inadequate and the loss function predictor is chosen. The asymptotic prediction risk for this pre-test procedure is derived. Both the Monte Carlo results and the asymptotic calculations demonstrate that the pre-test can avoid choosing the inferior predictor among the Bayes and the loss function predictor. As an alternative to the usual mini-max argument that is used to justify loss function estimation, we offer a Bayesian interpretation of the loss function predictor and the pre-testing. If the misspecification of the candidate model is believed to be large, then the integrated prediction risk is small for a prediction rule based on a low threshold level c.

References

Box, George E.P. (1980): "Sampling and Bayes' Inference in Scientific Modelling and Robustness". *Journal of the Royal Statistical Society A*, **143**, 383-430.

Christoffersen, Peter F. and Francis X. Diebold (1996): "Further Results on Forecasting and Model Selection under Asymmetric Loss". *Journal of Applied Econometrics*, **11**, 561-71.

——— (1997): "Optimal Prediction under Asymmetric Loss". *Econometric Theory*, **13**, 808-817.

Clements, Michael P. and David F. Hendry (1998): *"Forecasting Economic Time Series"*. Cambridge University Press.

Crowder, Martin (1988): "Asymptotic Expansions of Posterior Expectations, Distributions, and Densities for Stochastic Processes". *Annals of the Institute for Statistical Mathematics*, **40**, 297-309.

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

Findley, David F., Benedict M. Pötscher, and Ching-Zong Wei (1998): "Convergence Results for the Modeling of Time Series Arrays by Multistep Prediction or Likelihood Methods". *Mimeographed*, University of Vienna.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (1995): *"Bayesian Data Analysis"*, Chapman & Hall.

Gourieroux, Christian, Alain Monfort, and E. Renault (1993): "Indirect Inference". *Journal of Applied Econometrics*, **8**, S85-118.

Granger, Clive W.J. (1993): "On the Limitations of Comparing Mean Squared Forecast Errors: Comment". *Journal of Forecasting*, **12**, 651-652.

Hausman, J.A. (1978): "Specification Tests in Econometrics". *Econometrica*, **46**, 1251-71.

Lin, Jin-Lung and Ruey S. Tsay (1996): "Co-Integration Constraint and Forecasting: An Empirical Examination". *Journal of Applied Econometrics*, **11**, 519-538.

Phillips, Peter C.B. and Victor Solo (1992): "Asymptotic Theory for Linear Processes". *Annals of Statistics*, **20**, 971-1001.

Tiao, George C. and Ruey S. Tsay (1994): "Some Advances in Non-linear and Adaptive Modelling in Time-series". *Journal of Forecasting*, **13**, 109-131.

Tsay, Ruey S. (1993): "Comment: Adaptive Forecasting". *Journal of Business and Economics Statistics*, **11**, 140-142.

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

——————— (1996): "Estimating Time Series Models Using the Relevant Cost Function". *Journal of Applied Econometrics*, **11**.

Weiss, Andrew (1984) and A.P. Andersen (1984): "Estimating Forecasting Models Using the Relevant Forecast Evaluation Criterion". *Journal of the Royal Statistical Society A*, **137**, 484-487.

A Derivation and Proofs

Throughout the Appendix, we will analyze the VAR(1) case. However, we will use the companion form notation. The derivations can be easily generalized to higher order vector autoregressions by the appropriate insertion of the selection matrix M_n.

A.1 Calculation of the Pseudo-true VAR parameters

The difference between the conditional expectation of Y_{T+h} and the prediction function ΦY_{T+h} is

$$\mathbb{E}^*_T[Y_{T+h}] - \Phi^h Y_T = (F^h - \Phi^h)\sum_{j=0}^{\infty} F^j E_{T-j} + \alpha T^{-1/2}\sum_{j=0}^{\infty}(A_{j+h} - F^h A_j)U_{T-j}$$

The expected quadratic deviation of the prediction function from the conditional expectation is

$$\begin{aligned}
&\mathbb{E}^*\Big[tr[W(\mathbb{E}^*_T[Y_{T+h}] - \Phi^h Y_T)(\mathbb{E}^*_T[Y_{T+h}] - \Phi^h Y_T)']\\
&= \mathbb{E}^*\Big[tr\Big[W\Big((F^h - \Phi^h)\Gamma_{XX}(0)(F^h - \Phi^h)' + \alpha^2 T^{-1}\Phi^h\Gamma_{ZZ}(0)\Phi^{h'} - 2\alpha^2 T^{-1}\Phi^h\sum_{j=0}^{\infty} A_j\Sigma_{UU}A'_{j+h}\\
&\quad +2\alpha T^{-1/2}\Phi^h(\sum_{j=0}^{\infty} A_j\Sigma_{UE}F^{j'})(\Phi^h - F^h)' - 2\alpha T^{-1/2}(\sum_{j=0}^{\infty} A_{j+h}\Sigma_{UE}F^{j'})(\Phi^h - F^h) + (\cdots)\Big)\Big]\Big]
\end{aligned}$$

The omitted terms do not depend on Φ^h and therefore do not affect the calculation of pseudo-true values. The pseudo-true $\Phi^h_{T,l}$ has to satisfy the first order condition

$$\begin{aligned}
&\sqrt{T}(\Phi^h_{T,l} - F^h)\Big[\Gamma_{XX}(0) + \alpha T^{-1/2}(\sum_{j=0}^{\infty} F^j\Sigma_{UE}A'_j) + \alpha T^{-1/2}(\sum_{j=0}^{\infty} A_j\Sigma_{UE}F^{j'})\Big]\\
&= \alpha\sum_{j=0}^{\infty} A_{j+h}\Sigma_{UE}F^{j'} - \alpha\sum_{j=0}^{\infty} F^h A_j\Sigma_{UE}F^{j'} + \alpha T^{-1/2}\sum_{j=0}^{\infty} A_{j+h}\Sigma_{UU}A'_j - \alpha^2 T^{-1/2}F\Gamma_{ZZ}(0)
\end{aligned}$$

which implies that

$$\sqrt{T}(\Phi^h_{T,l} - F^h) = \alpha[\Gamma_{ZX}(h) - F^h\Gamma_{ZX}(0)]\Gamma_{XX}(0)^{-1} + o(1) \quad \square$$

A.2 Proof of Proposition 1

We will begin with the asymptotic analysis of the sample autocovariance and then use the δ-method to derive the limit distribution for $\hat{\Phi}^h_{T,b}$, $\hat{\Phi}^h_{T,l}$, and $\hat{\Phi}^h_{T,b} - \hat{\Phi}^h_{T,l}$.

Denote the sample autocovariance matrices of Y_t by $\hat{\Gamma}_{YY}(T,h) = \frac{1}{T}\sum_{t=h}^{T} Y_t Y_{t-h}'$. Under the reference model, the sample autocovariances have the following asymptotic approximation:

$$\begin{aligned}
\hat{\Gamma}_{YY}(T,h) &= \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=0}^{\infty}(F^jE_{t-j}+\alpha T^{-1/2}A_jU_{t-j}\right)\left(\sum_{j=0}^{\infty}(F^jE_{t-h-j}+\alpha T^{-1/2}A_jU_{t-h-j}\right)' \\
&= \frac{1}{T}\sum_{t=1}^{T}\Big[(\sum F^jE_{t-j})(\sum F^jE_{t-h-j})' + \alpha T^{-1/2}(\sum F^jE_{t-j})(\sum A_jU_{t-h-j}) \\
&\quad +\alpha T^{-1/2}(\sum F^jE_{t-h-j})(\sum A_jU_{t-j})' + \alpha^2T^{-1}(\sum A_jU_{t-j})(\sum A_jU_{t-h-j})'\Big] \\
&\xrightarrow{p} \Gamma_{XX}(h)
\end{aligned}$$

since the last three terms converge to zero as sample size tends to inifinity. Now consider the asymptotic behavior of $\sqrt{T}$ standardized sample autocovariances

$$\begin{aligned}
&\sqrt{T}(\hat{\Gamma}_{YY}(T,h)-\Gamma_{XX}(h)) \\
&= \sqrt{T}(\hat{\Gamma}_{XX}(T,h)-\Gamma_{XX}(h)) + \frac{1}{T}\sum_{t=1}^{T}\Big[(\sum F^jE_{t-j})(\sum A_jU_{t-h-j})' \\
&\quad +(\sum F^jE_{t-h-j})(\sum A_jU_{t-j})'\Big] + o_p(1)
\end{aligned}$$

where

$$\frac{1}{T}\sum_{t=1}^{T}\Big[(\sum F^jE_{t-j})(\sum A_jU_{t-h-j})' + (\sum F^jE_{t-h-j})(\sum A_jU_{t-j})'\Big] \xrightarrow{p} \Gamma_{XZ}(h)+\Gamma_{ZX}(h)$$

Following the proof of Theorem 3.7 in Phillips and Solo (1992) it can be deduced that

$$\sqrt{T}\begin{bmatrix} vech[\hat{\Gamma}_{YY}(T,0)-\Gamma_{XX}(0)] \\ vech[\hat{\Gamma}_{YY}(T,1)-\Gamma_{XX}(1)] \\ vech[\hat{\Gamma}_{YY}(T,h)-\Gamma_{XX}(h)] \end{bmatrix} \Longrightarrow \mathcal{N}\left(\begin{bmatrix} G_0 \\ G_1 \\ G_h \end{bmatrix}, \begin{bmatrix} S_{00} & S_{01} & S_{0h} \\ S_{10} & S_{11} & S_{1h} \\ S_{h0} & S_{h1} & S_{hh} \end{bmatrix}\right)$$

where $G_i = \Gamma_{XZ}(i)+\Gamma_{ZX}(i)$ for $i = 0,1,h$. Define $F_k^*(1) = \sum_{j=0}^{\infty} F^j \otimes F^{j+k}$ and $F_{-k}^*(1) = \sum_{j=k}^{\infty} F^j \otimes F^{j-k}$, for $k \geq 0$. The limit covariance matrix consists of the sub-matrices S_{kl} ($k = 0,1,h$; $l = 0,1,h$)

$$\begin{aligned}
S_{k,l} &= F_k^*(1)\mathbb{E}^*[(EE'-\Sigma_{EE})(EE'-\Sigma_{EE})']F_l^{*'}(1) \\
&\quad +\sum_{r=1}^{\infty}\big(F_{k+r}^*(1)+F_{k-r}^*(1)\big)\mathbb{E}^*[(E_{t-r}E_t')(E_{t-r}E_t')']\big(F_{l+r}^*(1)+F_{l-r}^*(1)\big)'
\end{aligned}$$

To apply the δ-method we approximate the functions $\hat{\Phi}_{T,b}^h$ and $\hat{\Phi}_{T,l}^h$ by a first order Taylor expansion around $\Gamma_{XX}(0)$, $\Gamma_{XX}(1)$, $\Gamma_{XX}(h)$, respectively. Define $d\hat{\Gamma}_{YY}(T,s) = \hat{\Gamma}_{YY}(T,s)-$

$\Gamma_{XX}(s)$

$$\begin{aligned}
&vech(\hat{\Phi}^h_{T,b} - [\Gamma_{XX}(1)\Gamma_{XX}(0)^{-1}]^h) \\
&= -\left\{\sum_{k=0}^{h-1}[\Gamma_{XX}(1)\Gamma_{XX}(0)^{-1}]^{h-k} \otimes [\Gamma_{XX}(0)^{-1}\Gamma_{XX}(-1)]^k\Gamma_{XX}(0)^{-1}\right\} vec(d\hat{\Gamma}_{YY}(0)) \\
&\quad +\left\{\sum_{k=0}^{h-1}[\Gamma_{XX}(1)\Gamma_{XX}(0)^{-1}]^{h-k-1} \otimes [\Gamma_{XX}(0)^{-1}\Gamma_{XX}(1)]^{k+1}\Gamma_{XX}(-1)^{-1}\right\} vec(d\hat{\Gamma}_{YY}(-1)) \\
&\quad +R_b(d\hat{\Gamma}_{YY}(0), d\hat{\Gamma}_{YY}(h))
\end{aligned}$$

$$\begin{aligned}
&vech(\hat{\psi}_T(h) - \Gamma(h)\Gamma(0)^{-1}) \\
&= -\left\{\Gamma_{XX}(h)\Gamma_{XX}(0)^{-1} \otimes \Gamma_{XX}(0)^{-1}\right\} vec(d\hat{\Gamma}_{YY}(0)) \\
&\quad + \left\{I \otimes \Gamma_{XX}(0)^{-1}\Gamma_{XX}(-h)\Gamma(-h)^{-1}\right\} vec(d\hat{\Gamma}_{YY}(-h)) + R_l(d\hat{\Gamma}_{YY}(0), d\hat{\Gamma}_{YY}(h))
\end{aligned}$$

Tedious but straightforward algebraic manipulations reveal that $\hat{\Phi}^h_{T,b}$, and $\hat{\Phi}^h_{T,l} - \Phi^h_{T,b}$ are asymptotically independent with limit variances V_b^0, and $V_l^0 - V_b^0$, respectively. □

A.3 Proof of Proposition 2

Consider the three terms of the prediction risk decomposition in Equation (22). Note that

$$\begin{aligned}
I\!E^*_T\tilde{Y}_{T+h} - \Phi^h_{T,i}\tilde{Y}_T &= \alpha T^{-1/2}\sum_{j=0}^{\infty} A_{j+h}\tilde{U}_{T-j} - \alpha T^{-1/2}\sum_{j=0}^{\infty} F^h A_j\tilde{U}_{T-j} \\
&\quad -\alpha T^{-1/2}\mu_\iota\left(\sum_{j=0}^{\infty} F^j\tilde{E}_{T-j} + \alpha T^{-1/2}\sum_{j=0}^{\infty} A_j\tilde{U}_{T-j}\right)
\end{aligned}$$

Therefore, the first term converges to

$$\begin{aligned}
&T \cdot I\!E^*\left[tr[W(I\!E^*_T\tilde{Y}_{T+h} - \Phi^h_{T,i}\tilde{Y}_T)(I\!E^*_T\tilde{Y}_{T+h} - \Phi^h_{T,i}\tilde{Y}_T)']\right] \\
&\longrightarrow \alpha^2 tr\Bigg[W\Bigg(\mu_\iota\Gamma_{XX}(0)\mu_\iota' - 2\sum_{j=0}^{\infty}(A_{j+h} - F^hA_j)\Sigma_{UE}F^{j'}\mu_\iota' \\
&\qquad +\sum_{j=0}^{\infty}(A_{j+h} - F^hA_j)\Sigma_{UU}(A_{j+h} - F^hA_j)'\Bigg)\Bigg]
\end{aligned}$$

Since $\sum_{j=0}^{\infty}(A_{j+h} - F^hA_j)\Sigma_{UE}F^{j'} = \mu_l\Gamma_{XX}(0)$, see Equation (14), the first term converges to $\alpha^2 R_\iota$, defined in Equation (23).

The second term can be manipulated as follows

$$\begin{aligned}
&T \cdot I\!E^*_T\left[tr[W(\hat{\Phi}_{T,\iota} - \Phi_{T,\iota})\tilde{Y}_T\tilde{Y}_T'(\hat{\Phi}_{T,\iota} - \Phi_{T,\iota})']\right] \\
&= I\!E^*_T\left[tr(W \otimes \tilde{Y}_T\tilde{Y}_T')vech(\hat{\Phi}^h_{T,\iota} - \Phi^h_{T,\iota})vech'(\hat{\Phi}^h_{T,\iota} - \Phi^h_{T,\iota})\right]]
\end{aligned}$$

If the sequence $\left\{T \cdot vech(\hat{\Phi}^h_{T,\iota} - \Phi^h_{T,\iota})vech'(\hat{\Phi}^h_{T,\iota} - \Phi^h_{T,\iota})\right\}_{T=T_0}^{\infty}$ is uniformly integrable for some T_0, then it can be deduced that

$$T \cdot \mathbb{E}^*_T\left[tr[W(\hat{\Phi}_{T,\iota} - \Phi_{T,\iota})\tilde{Y}_T\tilde{Y}'_T(\hat{\Phi}_{T,\iota} - \Phi_{T,\iota})']\right] \longrightarrow tr[(W \otimes \Gamma_{XX}(0))V_\iota]$$

The third term converges to zero because $\hat{\Phi}_{T,\iota}$ was assumed to be independent of $\tilde{Y}_T$, $\mathbb{E}^*_T\tilde{Y}_T - \Phi^h_{T,\iota}\tilde{Y}_T$ is $O_p(T^{-1/2})$, and $\mathbb{E}_*[\sqrt{T}(\hat{\Phi}^h_{T,\iota} - \Phi_{T,\iota})] \longrightarrow 0$. □

A.4 Proof of Proposition 3

Define $D_T = \Phi^h_{T,l} - \Phi^h_{T,b} = \alpha T^{-1/2}(\mu_l - \mu_b)$. We proceed by analyzing the three terms of the prediction risk decomposition in Equation (22). However, we have to take into account that the choice of predictor depends on the pre-test. Let $\{g(\Upsilon_T) > c^2\}$ denote the indicator function that is equal to one if $g(\Upsilon_T) > c^2$. The first component of the frequentist risk is

$$\begin{aligned}
&T \cdot \mathbb{E}^*\Big[tr[(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,b}\tilde{Y}_T)(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,b}\tilde{Y}_T)']\{g(\Upsilon_T) \leq c^2\} \\
&+tr[(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,b}\tilde{Y}_T)(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,l}\tilde{Y}_T)']\{g(\Upsilon_T) \leq c^2\}\Big] \\
&\quad = \quad T \cdot \mathbb{E}^*\Big[tr[(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,b}\tilde{Y}_T)(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,b}\tilde{Y}_T)'] + tr[D_T\tilde{Y}_T\tilde{Y}'_TD'_T]\{g > c^2\} \\
&\qquad\qquad -2tr[(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,b}\tilde{Y}_T)\tilde{Y}'_TD'_T]\{g > c^2\}\Big] \\
&\quad \longrightarrow \quad \alpha^2R^*_b + \alpha^2tr[W(\mu_l - \mu_b)\Gamma_{XX}(0)(\mu_l - \mu_b)']I_{(0)}(c^2, \alpha V_d^{-1/2}(\mu_l - \mu_b)) \\
&\qquad\qquad -2\alpha^2tr[W(\mu_l - \mu_b)\Gamma_{XX}(0)(\mu_l - \mu_b)']I_{(0)}(c^2, \alpha V_d^{-1/2}(\mu_l - \mu_b)) \\
&\quad = \quad \alpha^2R^*_b + \alpha^2(R^*_l - R^*_b)I_{(0)}(c^2, \alpha V_d^{-1/2}(\mu_l - \mu_b))
\end{aligned}$$

because

$$\begin{aligned}
&T \cdot \mathbb{E}^*\left[tr[W(\mathbb{E}^*_T\tilde{Y}_{T+h} - \Phi^h_{T,b}\tilde{Y}_T)\tilde{Y}'_TD'_T]\right] \\
&\quad = \quad T \cdot \mathbb{E}^*\Bigg[tr\Bigg[W\left(\alpha T^{-1/2}\sum_{j=0}^{\infty}(A_{j+h} - F^hA_j)\tilde{U}_{T-j} - \alpha T^{-1/2}\mu_b(\sum_{j=0}^{\infty}F^j\tilde{E}_{T-j} + \alpha T^{-1/2}\sum_{j=0}^{\infty}A_j\tilde{U}_{T-j})\right) \\
&\qquad\qquad \times\left((\sum_{j=0}^{\infty}F^j\tilde{E}_{T-j} + \alpha T^{-1/2}\sum_{j=0}^{\infty}A_j\tilde{U}_{T-j})'(\mu_l - \mu_b)'\right)\Bigg]\Bigg] \\
&\quad = \quad \alpha^2tr\left[W\left(\sum_{j=0}^{\infty}(A_{j+h} - F^hA_j)\Sigma_{UE}F^{j'}\right)(\mu_l - \mu_b)'\right] - \alpha^2tr\left[W\mu_b\Gamma_{XX}(0)(\mu_l - \mu_b)'\right] \\
&\quad = \quad \alpha^2tr[W(\mu_l - \mu_b)\Gamma_{XX}(0)(\mu_l - \mu_b)']
\end{aligned}$$

The second term is of the form

$$\begin{aligned}
& T \cdot \mathbb{E}^* \Big[tr[W(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b})\tilde{Y}_T \tilde{Y}'_T(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b}\tilde{Y}_T)']\{g(\Upsilon_T) \leq c^2\} \\
& + tr[W(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b} + (\hat{D}_T - D_T))\tilde{Y}_T \tilde{Y}'_T(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b}\tilde{Y}_T)' + (\hat{D}_T - D_T)]\{g(\Upsilon_T) > c^2\} \Big] \\
= \quad & T \cdot \mathbb{E}^* \Big[tr[W(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b})\tilde{Y}_T \tilde{Y}'_T(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b}\tilde{Y}_T)'] \\
& + tr[W(\hat{D}_T - D_T)\tilde{Y}_T \tilde{Y}'_T(\hat{D}_T - D_T)']\{g(\Upsilon_T) > c^2\} \\
& + 2tr[W(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b})\tilde{Y}_T \tilde{Y}'_T(\hat{D}_T - D_T)']\{g(\Upsilon_T) > c^2\} \Big] \\
\longrightarrow \quad & tr[(W \otimes \Gamma_{XX}(0))V^0_b] + tr[(W \otimes \Gamma_{XX}(0))(V_d^{1/2} I_{(2)}(c^2, \alpha V_d^{-1/2}(\mu_l - \mu_b))V_d^{1/2'}]
\end{aligned}$$

The cross product term drops out because $(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b})$ is asymptotically independent of $\tilde{Y}_T$ and $\hat{D}_T$ and has mean zero.

The third term component is of the form

$$\begin{aligned}
& T \cdot \mathbb{E}^* \Big[tr[W(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b})\tilde{Y}_T(\mathbb{E}^*_T \tilde{Y}_{T+h} - \Phi^h_T \tilde{Y}_T)'] \\
- \quad & tr[W(\hat{\Phi}^h_{T,b} - \Phi^h_{T,b})\tilde{Y}_T \tilde{Y}'_T D_T]\{g(\Upsilon_T > c^2)\} \\
+ \quad & tr[W(\hat{D}_T - D_T)\tilde{Y}_T(\mathbb{E}^*_T \tilde{Y}_{T+h} - \Phi^h_{T,l}\tilde{Y}_T)']\{g(\Upsilon_T > c^2)\} \Big]
\end{aligned}$$

The first two term converges to zero because $\hat{\Phi}^h_{T,b} - \Phi^h_{T,b}$ has asymptotically mean zero and is independent of the other expressions. The last term converges to zero because

$$\begin{aligned}
& \mathbb{E}^* \Big[(\mathbb{E}^*_T \tilde{Y}_{T+h} - \Phi^h_{T,l}\tilde{Y}_T)\tilde{Y}'_T \Big] \\
= \quad & \alpha T^{-1/2} \mathbb{E}^* \Bigg[\left(\sum_{j=0}^{\infty} (A_{j+h} - F^h A_j)\tilde{U}_{T-j} - \mu_l (\sum_{j=0}^{\infty} F^j \tilde{E}_{T-j} + \alpha T^{-1/2} \sum_{j=0}^{\infty} A_j \tilde{U}_{T-j}) \right) \\
& \times \left(\sum_{j=0}^{\infty} F^j \tilde{E}_{T-j} + \alpha T^{-1/2} \sum_{j=0}^{\infty} A_j \tilde{U}_{T-j} \right) \Bigg]' \\
\longrightarrow \quad & \alpha T^{-1/2} [\sum_{j=0}^{\infty} (A_{j+h} - F^h A_j)\Sigma_{UE} F^{j'} - \mu_l \Gamma_{XX}(0)] = 0
\end{aligned}$$

This completes the prediction risk calculations for the pre-test predictor. □