Loss Function Estimation of Forecasting Models A Bayesian Perspective

Frank Schorfheide
University of Pennsylvania, Department of Economics
3718 Locust Walk, Philadelphia, PA 19104-6297
schorf@ssc.upenn.edu

Key Words: Bayesian, Forecasting, Loss Functions.

1 Introduction

Suppose a simple AR(1) model $y_t = \mu + \rho y_{t-1} + \epsilon_t$ is used to forecast GDP per capita growth h quarters ahead. This multi-step forecast is evaluated under a quadratic prediction error loss function. Parameter estimates can be obtained by minimizing $\sum_{t=1}^{T}(y_t - \mu - \rho y_{t-1})^2$, which corresponds to maximum likelihood estimation conditional on the initial observation if the ϵ_t have a Gaussian distribution. Based on $\hat{\mu}$ and $\hat{\rho}$, multi-step ahead predictions can be generated iteratively according to $\hat{y}_{T+h|T} = \hat{\mu} + \hat{\rho}\hat{y}_{T+h-1|T}$. In a Bayesian framework, the exact predictor depends on the prior distribution and takes into account the posterior parameter uncertainty. Asymptotically, the Bayes predictor and the maximum likelihood plug-in predictor are equivalent.

Alternatively, an AR(1) model forecast can be obtained by simply minimizing the in-sample h step ahead prediction errors $\sum_{t=h-1}^{T}(y_t - \tilde{\mu} - \tilde{\rho}y_{t-h})^2$. This loss function estimation approach leads to the predictor $\hat{y}_{T+h|T} = \hat{\tilde{\mu}} + \hat{\tilde{\rho}}y_T$. In practice it has to be decided whether to use the likelihood function or the loss function to estimate the model parameters. If GDP growth were correctly represented by an AR(1) process, then statistical theory implies that the likelihood based procedure is preferable to the loss function estimation. If the AR(1) model, however, is misspecified then the ranking can change.

In the context of h-step ahead forecasting the loss function estimators are also called dynamic estimators. The properties of such estimators have been examined, for instance, by Clements and Hendry (1998), Findlay (1983), Findlay *et al.* (1998), Weiss and Andersen (1984), Weiss (1991). 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 expectation of y_{T+h} conditional on time T information. This paper examines the choice between likelihood and loss function based predictors from a Bayesian perspective.

I wish to thank Peter Phillips and Chris Sims for valuable comments, and suggestions. Financial Assistance from the German Academic Exchange Service through "Doktorandenstipendium aus Mitteln des zweiten Hochschulsonderprogramms" and the Alfred P. Sloan Foundation are gratefully acknowledged.

2 A Framework

At time T, an econometrician faces a decision problem under uncertainty. He observes a time series $Y_T = \{y_t\}_{t=1}^{T}$ where y_t is a $n \times 1$ vector. He has to make a point prediction $\hat{\varphi}$ for a random variable φ that takes values in $\mathbb{R}^m$. Here, φ is composed of future observations y_{t+h}, $h \leq s$. The predictor $\hat{\varphi}$ is a function of the data Y_T, and is evaluated a loss function $L(\varphi, \hat{\varphi})$. The prediction loss is observable at time $T+s$ for some $s > 0$ and the sample up to time T can be used to evaluate hypothetical prediction losses that would have occurred if the predictor had been used at early times.

A family of parametric probability distributions $\{P_{\theta,T+s} : \theta \in \Theta\}$ is considered to characterize the joint distribution of Y_{T+s}. For expository purposes, it is assumed that $\Theta \subseteq \mathbb{R}^k$ and that the P_θ distributions and the prior distribution of θ have densities $p(Y_{T+s}|\theta)$ and $p(\theta)$, respectively. The candidate probability model defined by likelihood function and prior is denoted by $\mathcal{M}$. The random variable φ has some distribution conditional on Y_T with density $p(\varphi|Y_T, \theta, \mathcal{M})$. The marginal posterior density of φ is

$$p(\varphi|Y_T, \mathcal{M}) = \int p(\varphi|Y_T, \theta, \mathcal{M})p(\theta|Y_T)d\theta \qquad (1)$$

where $p(\theta|Y_T)$ is the posterior of θ given the sample information Y_T.

Since it is assumed that model $\mathcal{M}$ is potentially misspecified, an alternative family of distributions $\{Q_{\psi,T+s} : \psi \in \Psi\}$ with prior distribution $p(\psi)$ is

considered. The $Q_{\psi,T+s}$ distributions and a prior for ψ define the reference model $\mathcal{M}_*$. This reference model might be of a very general form. The likelihood of the data under $\mathcal{M}_*$ is denoted by $p(Y_T|\psi,\mathcal{M}_*)$ and the prior probability of $\mathcal{M}_*$ is $\pi_{*,0}$. The mixture of $\mathcal{M}$ and $\mathcal{M}_*$ can be regarded as a model that the econometrician is willing to accept as "true" for practical purposes. We say, that model $\mathcal{M}$ is potentially misspecified *a priori* if the prior probability of the reference model $\pi_{*,0}$ is strictly greater than zero. A posteriori, model $\mathcal{M}$ is misspecified if the posterior probability of $\mathcal{M}_*$ approaches unity in large samples.

We will assume that the econometrician finds it too onerous to conduct a full posterior analysis with the reference model. If the econometrician would evaluate posterior predictions and model probabilities for the reference model then there would be no reason to contemplate loss function estimation of the candidate model. In practical applications, one often chooses simple candidate models, such as linear autoregressive specifications, because they can be easily analyzed and used to compute forecasts. Thus, only predictions derived from mixtures of the form

$$\int p(\varphi|Y_T,\theta,\mathcal{M})f(\theta)d\theta \tag{2}$$

are considered. The function $f(\theta)$ is a normalized weight function with $\int_\Theta f(\theta)d\theta = 1$. The corresponding predictor is obtained by

$$\hat{\varphi}(f(\theta)) = \mathrm{argmin}_{\varphi\in\mathbb{R}^m} \int L(\varphi,\hat{\varphi})p(\varphi|Y_T,\theta,\mathcal{M})f(\theta)d\theta \tag{3}$$

Two special cases are of particular interest: (i) The weight function $f(\theta) \in F$ is equal to the posterior density $p(\theta|Y_T)$. Therefore, the predictor minimizes the posterior expected loss provided the data stem from $\mathcal{M}$. (ii) The weight function concentrates its mass on a finite set of points $\theta_i' \in \Theta$, $i = 1,\ldots,k$. This will be denoted by $f(\theta) = \sum_{i=1}^k \lambda_i \delta_{\{\theta=\theta_i'\}}$ where $\lambda_i \geq 0$, $\sum \lambda_i = 1$, and δ has the properties $\int \delta_{\{\theta=\theta'\}}d\theta = 1$ and $\delta_{\{\theta=\theta'\}} = 0$ for $\theta \neq \theta'$.

The goal is to find a weight function $f_0(\theta)$, such that predictions, that are optimal if future observations conditional on Y_T were generated from the corresponding mixture of P_θ distributions, lead to small prediction losses if the data actually stem from the mixture of $\mathcal{M}$ and $\mathcal{M}_*$. We will consider two procedures to estimate appropriate weights $f(\theta)$ from the data. The first procedure simply uses the posterior weights $p(\theta|Y_T)$ solely based on $\mathcal{M}$, ignoring the possibility that the data could have been generated from $\mathcal{M}_*$. In the second procedure, a θ' is estimated for a degenerate weight function $\delta_{\{\theta=\theta'\}}$ by minimizing hypothetical prediction losses during the sample period. Both procedures are suboptimal relative to the full Bayesian procedure that explicitly takes the contribution of $\mathcal{M}_*$ to the overall posterior distribution of φ into account.

2.1 Suboptimal Prediction Procedures

Let $\pi_{*,T}$ be the posterior probability of the reference model $\mathcal{M}_*$ conditional on the observations Y_T. Under the loss function $L(\varphi,\hat{\varphi})$, the posterior prediction risk is defined as

$$\mathcal{R}(\hat{\varphi}|Y_T) = (1-\pi_{*,T})\int L(\varphi,\hat{\varphi})p(\varphi|Y_T,\mathcal{M})d\varphi + \pi_{*,T}\int L(\varphi,\hat{\varphi})p(\varphi|Y_T,\mathcal{M}_*)d\varphi \tag{4}$$

In the standard Bayesian analysis, the posterior distribution of φ under the reference model is evaluated or at least approximated and the optimal predictor is

$$\hat{\varphi}_{opt} = \mathrm{argmin}_{\hat{\varphi}\in\mathbb{R}^m}\, \mathcal{R}(\hat{\varphi}|Y_T) \tag{5}$$

In a related paper on the evaluation of dynamic stochastic equilibrium models (Schorfheide, 1999b), the weights $f(\theta)$ of the predictor defined in Equation (3) are obtained by minimization of $\mathcal{R}(\hat{\varphi}(f(\theta))|Y_T)$ over a suitable set of weight functions F. However, this is infeasible without evaluating $p(\varphi|Y_T,\mathcal{M}_*)$. In this paper we will consider predictors based on the observed frequencies of $L(\varphi,\hat{\varphi})$ up to time T and judge them according to their integrated risk.

The integrated risk $\mathcal{R}(\varphi)$ is obtained by averaging the posterior risk over all trajectories Y_T, that is,

$$\mathcal{R}(\hat{\varphi}) = (1-\pi_{*,T})\int \mathcal{R}(\hat{\varphi}|Y_T,\mathcal{M})p(Y_T|\mathcal{M})dY_T + \pi_{*,T}\int \mathcal{R}(\hat{\varphi}|Y_T,\mathcal{M}_*)p(Y_T|\mathcal{M}_*)dY_T \tag{6}$$

It is assumed that the integrated risk exists. This assumption restricts the shape of the loss function $L(\varphi,\hat{\varphi})$ and the prior distributions $p(\theta)$ and $p(\psi)$. A predictor $\hat{\varphi}_1$ is preferable to $\hat{\varphi}_2$ if $\mathcal{R}(\hat{\varphi}_1) < \mathcal{R}(\hat{\varphi}_2)$. Note that the full information Bayes predictor $\hat{\varphi}_{opt}$ minimizes $\mathcal{R}(\hat{\varphi}|Y_T)$ on almost all trajectories Y_T under the mixture of $\mathcal{M}$ and $\mathcal{M}_*$. Thus, $\hat{\varphi}_{opt}$ also minimizes the integrated risk.

Suppose it were known that the data stem from model $\mathcal{M}$. Conditional on this information it is optimal to use the Bayes predictor based solely on the P_θ distributions, which will be denoted by

$$\hat{\varphi}_b = \hat{\varphi}(p(\theta|Y_T)) \tag{7}$$

If the data stem from the reference model $\mathcal{M}_*$, then it is conceivable to use either the Bayes predictor $\hat{\varphi}_b$ or an alternative predictor, that we regard as loss function predictor $\hat{\varphi}_l$, obtained from a weight function $f(\theta) \neq p(\theta|Y_T)$. Conditional on the $Q_{\psi,T+s}$ distribution, it is possible to compare

$$\mathcal{R}(\hat{\varphi}|\psi,\mathcal{M}_*) = \int \mathcal{R}(\hat{\varphi}|Y_T,\psi,\mathcal{M}_*)p(Y_T|\psi,\mathcal{M}_*)dY_T \tag{8}$$

for the predictors $\hat{\varphi}_b$ and $\hat{\varphi}_l$. $\mathcal{R}(\hat{\varphi}|\psi,\mathcal{M}_*)$ is the frequentist prediction risk conditional on ψ and $\mathcal{M}_*$. Consider the lower bound of the integrated risk achieved by the following prediction procedure.

Procedure 1 *(Infeasible) If the data are generated by model $\mathcal{M}$, use the Bayes predictor $\hat{\varphi}_b$. If the data are generated by the reference model $\mathcal{M}_*$, in particular, from a distribution $Q_{\psi_0,T+s}$, then use $\hat{\varphi}_l$ if*

$$\mathcal{R}(\hat{\varphi}_l|\psi_0,\mathcal{M}_*) < \mathcal{R}(\hat{\varphi}_b|\psi_0,\mathcal{M}_*)$$

and $\hat{\varphi}_b$ otherwise.

Procedure 1 is infeasible because it is unknown to the forecaster whether the data stem from $\mathcal{M}$ or $\mathcal{M}_*$. Two simple and feasible selection strategies for $\hat{\varphi}_b$ and $\hat{\varphi}_l$ are to use either the Bayes predictor or the loss function predictor on each trajectory Y_T.

Procedure 2 *(Feasible) Always use the Bayes predictor $\hat{\varphi}_b$.*

Procedure 3 *(Feasible) Always use the loss function predictor $\hat{\varphi}_l$.*

The Bayes predictor, or the related maximum likelihood plug-in predictor $\hat{\varphi}(\delta_{\{\theta=\hat{\theta}_{mle}\}})$, is usually justified by low prior probability of distributions $Q_{\psi,T+s}$ that are very different from the $P_{\theta,T+s}$ distribution. Procedure 3 is motivated by a large sample minimax argument. If the data stem from a particular $P_{\theta_0,T+s}$ distribution then the loss function estimator will converge to θ_0 in a large sample so that the efficiency loss of using $\hat{\varphi}_l$ instead of $\hat{\varphi}_b$ given $\mathcal{M}$ is small. However, the potential loss of using $\hat{\varphi}_b$ under $\mathcal{M}_*$ is large if there are parameters ψ in the support of the prior $p(\psi)$ such that the conditional distribution of φ given Y_T under $Q_{\psi,T+s}$ is very different from the conditional distribution under $P_{\theta,T+s}$.

2.2 Loss Function Based Procedures

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. This section develops a notion of pseudo-true values that is closely linked to the prediction problem and discusses the relationship between pseudo-true values and loss function estimators. Under the reference model $\mathcal{M}_*$, we can define an optimal predictor $\hat{\varphi}_\psi$ conditional on the parameters ψ and the observations Y_T as

$$\hat{\varphi}_\psi = \text{argmin}_{\varphi\in I\!R^m}\ \mathcal{R}(\varphi|Y_T,\psi,\mathcal{M}_*) \tag{9}$$

The additional risk of using any other predictor $\hat{\varphi}$ is

$$\begin{aligned} & R(\hat{\varphi}|Y_T,\psi,\mathcal{M}_*) \\ &= \mathcal{R}(\hat{\varphi}|Y_T,\psi,\mathcal{M}_*) - \mathcal{R}(\hat{\varphi}_\psi|Y_T,\psi,\mathcal{M}_*) \geq 0 \end{aligned} \tag{10}$$

Suppose that $\hat{\varphi}_f = \hat{\varphi}(f(\theta))$ is derived from a mixture of $P_{\theta,T+s}$ distributions with weights $f(\theta)$ according to Equation (3). This mixture is denoted by $\mathcal{P}_f$. The idea of deriving predictions from mixtures is an important aspect of the extensive literature on Bayesian and non-Bayesian approaches of combining forecasts, for instance, Bates and Granger (1968), Min and Zellner (1993), and references cited therein. Rather than interpreting $R(\hat{\varphi}|Y_T,\psi,\mathcal{M}_*)$ as risk differential it can also be interpreted as discrepancy $\Delta(\mathcal{P}_f|Q_\psi)$ between the P_θ mixture and the Q_ψ distribution.

Similar to the well-known Kullback-Leibler distance, the discrepancy $\Delta(\mathcal{P}_f|Q_\psi) = R(\hat{\varphi}|Y_T,\psi,\mathcal{M}_*)$ is not a metric since it is neither symmetric because $\Delta(\mathcal{P}_f|Q_\psi) \neq \Delta(Q_\psi|\mathcal{P}_f)$, nor does it satisfy the triangle inequality. While the Kullback-Leibler distance has a very general information-theoretic interpretation that is valid without postulating a specific decision problem, $\Delta(\mathcal{P}_f|Q_\psi)$ provides the relevant measure of discrepancy in situations where the econometrician does face a specific decision problem such as forecasting. We can now easily define pseudo-true parameter values θ_0 and more generally, a pseudo-true $f_0(\theta)$ mixture.

Definition 1 *(i) The pseudo-true parameter θ_0 is a solution to the problem*

$$\min_{\theta'\in\Theta}\ \Delta(P_{\theta',T+s}|Q_{\psi,T+s})$$

where $\Delta(P_{\theta',T+s}|Q_{\psi,T+s}) = R(\hat{\varphi}(\delta_{\{\theta=\theta'\}})|\psi,\mathcal{M}_*)$. (ii) The weight function $f_0(\theta)$ of the pseudo-true mixture $\mathcal{P}_{f_0}$ solves the problem

$$\min_{f(\theta)\in F}\ \Delta(\mathcal{P}_{f(\theta),T+s}|Q_{\psi,T+s})$$

This notion of pseudo-true values and mixtures depends on the prediction problem, the loss function, and the sample size. The solution to the minimization problems stated in Definition 1 need not be unique. However, if $\Delta(\mathcal{P}_f|Q_\psi) = \Delta(\mathcal{P}_{f'}|Q_\psi)$ then the predictors associated with the two mixtures perform equally well conditional on Q_ψ.

Suppose the set of weight functions F is restricted to the denegenerate weights $\delta_{\{\theta=\theta'\}}$, $\theta' \in \Theta$. Moreover, $\varphi = y_{T+h}$. The loss function estimator is of the form

$$\hat{\theta}_T = \text{argmin}_{\theta' \in \Theta} \sum_{t=t_0}^{T-h} L(y_{t+h}, \hat{y}_{t+h}(\delta_{\{\theta=\theta'\}})) \quad (11)$$

If $Q_{\psi,T+s}$ and $L(\varphi, \hat{\varphi})$ satisfy some weak regularity conditions that guarantee the uniform convergence of the average hypothetical prediction loss before time T to the frequentist risk at time T, that is,

$$R(\hat{y}_{t+h}|\psi, \mathcal{M}_*) - \frac{1}{T}\sum_{t=t_0}^{T-h} L(y_{t+h}, \hat{y}_{t+h}) \longrightarrow 0 \quad (12)$$

almost surely under $Q_{\psi,T+s}$ then the loss function estimator $\hat{\theta}_T$ converges to the pseudo-true value. The estimator belongs to the extensively studied class of extremum estimators.

Non-degenerate weight functions $f(\theta)$ lead to a combination of forecasts from different P_θ models. The function $f(\theta) = p(\theta|Y_T, \mathcal{M})$ corresponds to the Bayesian approach of combining forecasts. Provided that the $Q_{\psi,T+s}$ distributions satisfy some regularity conditions, the posterior $p(\theta|Y_T, \mathcal{M})$ will converge to a point mass under $\mathcal{M}_*$ as the sample size T tends to infinity. Thus, asymptotically, all forecasts are derived from a single P_θ model, instead from the pseudo-true mixture $\mathcal{P}_{f_0}$. Most non-Bayesian approaches can be interpreted as attempts to estimate a weight function that remains non-degenerate as $T \to \infty$. Consider the following class of functions F

$$\left\{ f_\lambda(\theta) = \lambda\delta_{\{\theta=\theta_1\}} + (1-\lambda)\delta_{\{\theta=\theta_2\}} : \lambda \in [0,1] \right\}$$

The parameter space Θ consists of only two parameters θ_1 and θ_2. For h-step ahead forecasting under quadratic loss, the predictor derived from the λ mixture of P_{θ_1} and P_{θ_2} is of the form

$$\hat{y}_{t+h}(f_\lambda) = \lambda\hat{y}_{t+h}(\delta_{\{\theta=\theta_1\}}) + (1-\lambda)\hat{y}_{t+h}(\delta_{\{\theta=\theta_2\}})$$

The loss function estimator can be defined as $\hat{\lambda}_T = \text{argmin}_{\lambda\in[0,1]}\ \hat{y}_{t+h}(f_\lambda)$ which leads to the forecast combination weights proposed by Bates and Granger (1969), except that our mixture interpretation requires λ to lie between zero and one.

3 Example: Multi-Step Forecasting of a Linear Process

The framework can be applied to the forecasting problem discussed in the Introduction. The $P_{\theta,T+h}$ distributions are given by the density through a Gaussian AR(1) process

$$\begin{aligned} & p(Y_{T+h}|\rho, \sigma, y_0) \quad (13) \\ &= \prod_{t=1}^{T+h} |\sigma^2|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2}(y_t - \rho y_{t-1})^2 \right\} \end{aligned}$$

where $\theta = (\rho, \sigma)$. Moreover, there is some prior distribution $p(\rho, \sigma)$. Let $\hat{\rho}_b$ denote the maximum likelihood estimate and $\hat{\rho}_l$ the loss function estimate of ρ.

The econometrician believes that the model $\mathcal{M}$ is potentially misspecified, and places prior probability one on a linear process as a reference model. Let $\psi = (\alpha, \rho_0, \sigma_\epsilon, \sigma_u, \sigma_{\epsilon,u}, a_0, a_1, \ldots)$. The distribution of y_t conditional on ψ is given by the model

$$y_t = x_t + \alpha T^{-1/2} z_t \quad (14)$$

where $x_t = \sum_{j=0}^{\infty} \rho_0^j \epsilon_{t-j}$, $z_t = \sum_{j=0}^{\infty} a_j u_{t-j}$, $\sum_{j=0}^{\infty} j^2 a_j^2 < \infty$, and $|\rho_0| < 1$. The innovations are distributed according to $\epsilon_t \sim iid(0, \sigma_\epsilon^2)$ and $u_t \sim iid(0, \sigma_u^2)$ with contemporaneous covariance $\sigma_{\epsilon u}^2$. Conditional on a disturbance process z_t, the parameter α controls the severeness of the misspecification. It can be shown that the Bayes predictor is asymptotically of the form $\hat{\varphi}_b = \hat{\rho}_b^h + O_p(1)$. The asymptotic sampling variance of the Bayes estimator, denoted by v_b, is smaller than the sampling variance v_l of the loss function estimator under the $Q_{\psi,T+s}$ distributions. Define

$$R_\iota^* = \lim_{T\longrightarrow\infty} TR(\hat{\varphi}(\rho_{T,\iota})|\alpha = 1, \psi, \mathcal{M}_*) \quad \iota = b, l$$

By definition, the pseudo-true value $\rho_{T,l}$ minimizes $R(\hat{\varphi}(\rho)|\alpha = 1, \psi, \mathcal{M}_*)$ with respect to ρ for all T. Therefore, $R_l^* \leq R_b^*$. The asymptotic frequentist risk $R(\hat{\varphi}_\iota|\psi, \mathcal{M}_*)$ for $\iota = b, l$ is:

$$\bar{R}(\hat{\varphi}_\iota|\psi, \mathcal{M}_*) = \alpha^2 R_\iota^* + \gamma_{xx}(0) v_\iota^0 \quad (15)$$

Since $R_l^* \leq R_b^*$ and $v_l^* \geq v_b^*$ there exists a trade-off between the two predictors. Conditional on the disturbance process z_t, the Bayes predictor is preferable if the misspecification parameter α is small.

3.1 A Pseudo Bayes Prediction Rule

Consider the following approach to select one of the predictors $\hat{\phi}_b$ and $\hat{\phi}_l$. 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. We will argue that this approach is flawed if the data are, in fact, generated from a mixture of $\mathcal{M}$ and $\mathcal{M}_*$. Let $\tilde{\theta} = (\tilde{\rho}, \tilde{\sigma})$. For h-step ahead forecasting under quadratic loss the alternative model is of the following form

$$\begin{aligned} & p(y_t|Y_{t-1}, \tilde{\theta}, \mathcal{M}_l) \\ & \propto \quad |\tilde{\sigma}^2|^{-1/2} \exp\left\{-\frac{1}{2\tilde{\sigma}^2}(y_t - \tilde{\rho}^h y_{t-h})^2\right\} \end{aligned} \tag{16}$$

with some prior distribution $p(\tilde{\theta})$. It is easily verified that for model $\mathcal{M}_l$ the Bayes predictor under the loss function $L(\varphi, \hat{\varphi}) = (y_{T+h} - \hat{y}_{T+h})^2$ behaves asymptotically like $\hat{\varphi}_l$. The model $\mathcal{M}_l$ implies that conditional on any parameter $\tilde{\theta}$ only the autocovariances of order $k \cdot h$, $k = 0, 1, \ldots$ can be non zero. Thus, the behavior of the time series Y_T under this loss function model $\mathcal{M}_l$ is quite different from the behavior under the actual reference model $\mathcal{M}_*$. Define

$$s^2 = \sum_{t=1}^{T}(y_t - \hat{\rho}_{T,b} y_{t-1})^2, \quad \tilde{s}^2 = \sum_{t=1}^{T}(y_t - \hat{\rho}^h_{T,l} y_{t-h})^2$$

Suppose the econometrician selects between $\hat{\varphi}_b$ and $\hat{\varphi}_l$ to minimize the posterior expected prediction loss under a mixture of $\mathcal{M}$ and $\mathcal{M}_l$, where $\mathcal{M}_l$ has prior probability $\pi_{l,T}$. In this case, it is optimal to choose the loss function predictor $\hat{\varphi}_l$ if the posterior odds $\pi_{l,T}/(1 - \pi_{l,T})$ favor model $\mathcal{M}_l$. The posterior odds are obtained as the ratio of the marginal data densities under $\mathcal{M}$ and $\mathcal{M}_l$. Asymptotically, the posterior odds ratio can be represented as

$$\frac{2}{T}\ln\left(\pi_{l,T}/(1 - \pi_{l,T})\right) = \ln(s^2/\tilde{s}^2) + O_p(T^{-1}) \tag{17}$$

which suggests the following selection rule as an approximation.

Procedure 4 *(Feasible) Choose the loss function predictor $\hat{\varphi}_l$ if $s^2 > \tilde{s}^2$ and the Bayes predictor $\hat{\varphi}_b$ otherwise.*

Suppose the data were generated under $\mathcal{M}_*$ conditional on parameters ψ. Then

$$\begin{aligned} \tilde{s}^2 - s^2 \quad &= \quad \frac{\left(\frac{1}{T}\sum y_t y_{t-h}\right)^2}{\frac{1}{T}\sum y^2_{t-h}} - \frac{\left(\frac{1}{T}\sum y_t y_{t-1}\right)^2}{\frac{1}{T}\sum y^2_{t-1}} \\ &\longrightarrow \quad \frac{\gamma_{xx}(h)^2 - \gamma_{xx}(1)^2}{\gamma_{xx}(0)^2} \\ &= \quad \rho_0^{2h} - \rho_0^2 < 0 \end{aligned} \tag{18}$$

almost surely under Q_ψ for $|\rho_0| < 1$. Thus, if the sample size is large, Procedure 4 suggests to almost always select the Bayes predictor $\hat{\varphi}_b$ based on model $\mathcal{M}$, regardless of the degree of the misspecification α. It is important to notice that posterior odds can be misleading if the data stem from neither $\mathcal{M}$ nor $\mathcal{M}_l$. The odds implicitly measure the success of $\hat{\rho}_{T,b} y_T$ to predict one-step ahead and of $\hat{\rho}^h_{T,l} y_T$ to forecast h-steps ahead, rather than comparing the h-step ahead performance of $\hat{\varphi}_b$ and $\hat{\varphi}_l$. Since under the reference model the expected h-step ahead forecast error is greater than the one-step ahead forecast error, Procedure 4 is not helpful in determining whether to use $\hat{\varphi}_b$ or $\hat{\varphi}_l$.

3.2 A Prediction Rule Based on 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. In the context of this paper, it is interesting to analyze whether such a model checking procedure can be helpful to choose between the Bayes predictor $\hat{\varphi}_b$ and the loss function estimator $\hat{\varphi}_l$. Unlike in many inferential situations, the nature of the decision problem requires a prediction. It is not possible to simply reject model $\mathcal{M}$ and search for a better representation of the data.

It is important to distinguish clearly between Y_T as a random variable and the observed time series which we will denote by $Y_{T,d}$. 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 $Y_{T,d}$. If $Y_{T,d}$ falls 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 $g(Y_T) \geq 0$. Suppose the density is of $g(Y_T)$ is unimodal. $\mathcal{M}$ passes the model check if

$$\int \mathcal{I}\Big\{g(Y_T) \geq g(Y_{T,d})\Big\} p(Y_T|\mathcal{M}) dY_T > \alpha \tag{19}$$

where $\mathcal{I}\{y \geq z\}$ is the indicator function that is one if $y \geq z$ and zero otherwise. If such a rule is embedded in a prediction procedure, the overall risk properties depend on the checking function $g(Y_T)$, the tail probability threshold level α, and at last, on the alternative $\mathcal{M}_*$. It is important to keep in

mind that the rejection of a model $\mathcal{M}$ does not automatically imply that the loss function predictor $\hat{\varphi}_l$ is preferable to the Bayes predictor $\hat{\varphi}_b$.

Procedure 5 *Use the Bayes predictor $\hat{\varphi}_b$ if $\mathcal{M}$ passes the model check at level α. Otherwise use the loss function predictor $\hat{\varphi}_l$.*

In Schorfheide (1999a) we analyze the asymptotic risk properties of a model check that is based on the idea of a Hausman (1978) test. The divergence of $\hat{\rho}^h_{T,b}$ and $\hat{\rho}^h_{T,l}$ 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^0_l and v^0_b 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 has a lot of power. The calibration of the procedure through the rejection threshold α depends on the prior distribution for $\mathcal{M}$ and $\mathcal{M}_*$. Conditional on a prior distribution it is possible to minimize $\bar{R}(\hat{\varphi}_c)$ with respect to α. In practice, it is more common to specify a "plausible" rejection level α. Given α, it is possible to determine prior distributions for which the choice of α is indeed sensible.

4 Conclusion

We can now return to the problem posed in the Introduction. Should the likelihood or the loss function estimator be used to compute h-step ahead forecasts of output growth with an AR(1) model. The previous sections demonstrated how the answer to this question is related to the kind of model that one is willing to accept as "true" for practical purposes. We provided framework to analyze prediction procedures for potentially misspecified models from a Bayesian perspective. The econometrician has a strong preference for a model $\mathcal{M}$ but believes that with some positive probability, the data stem from a reference model $\mathcal{M}_*$. He seeks a prediction procedure based on $\mathcal{M}$ that keeps the integrated prediction risk small. It is assumed that the Bayesian analysis of the reference model is too onerous. The alternative to the Bayes predictor is a loss function predictor that is well behaved under $\mathcal{M}_*$. A pseudo posterior odds selection rule between the two predictors is generally not helpful in reducing the prediction risk in the presence of misspecification. Choosing the predictor based on the outcome of a Bayesian model check can reduce the integrated risk of the forecasting procedure. A detailed analysis can be found in Schorfheide (1999a). Whether or not there are model checks that dominate other sampling tests for a large class of priors, and the analysis of the small sample risk properties of such prediction procedures through Monte Carlo studies is left for future research.

References

Bates, J.M. and Clive W.J. Granger (1969): "The Combination of Forecasts". *Operations Research Quarterly*, **20**, 451-468.

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

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

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.

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.

Schorfheide, Frank (1999a): "Loss Function vs. Likelihood Estimation of Forecasting Models: A pre-test Procedure and a Bayesian Interpretation". *Institute for Economic Research*, Working Paper 99-006, University of Pennsylvania.

Schorfheide, Frank (1999b): "A Unified Econometric Framework for the Evaluation of DSGE Models". *Institute for Economic Research*, Working Paper 99-007, University of Pennsylvania.

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