Available online at www.sciencedirect.com
ScienceDirect

Journal of Econometrics 140 (2007) 884–918

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Properties of optimal forecasts under asymmetric loss and nonlinearity

Andrew J. Patton[a,*], Allan Timmermann[b]

[a]*London School of Economics, HoughtonStreet, London WC2A 2AE, UK*
[b]*University of California, San Diego, 9500 Gilman Drive, La Jolla CA 92093-0508, USA*

Available online 6 September 2006

Abstract

Evaluation of forecast optimality in economics and finance has almost exclusively been conducted under the assumption of mean squared error loss. Under this loss function optimal forecasts should be unbiased and forecast errors serially uncorrelated at the single period horizon with increasing variance as the forecast horizon grows. Using analytical results we show that standard properties of optimal forecasts can be invalid under asymmetric loss and nonlinear data generating processes and thus may be very misleading as a benchmark for an optimal forecast. We establish instead that a suitable transformation of the forecast error—known as the generalized forecast error—possesses an equivalent set of properties. The paper also provides empirical examples to illustrate the significance in practice of asymmetric loss and nonlinearities and discusses the effect of parameter estimation error on optimal forecasts.

JEL classification: C53; C22; C52

Keywords: Loss function; Nonlinear data generating process; Rationality; Market efficiency; Prediction

1. Introduction

Properties of optimal forecasts under mean squared error (MSE) loss include unbiasedness of the forecast, lack of serial correlation in one-step-ahead forecast errors, serial correlation of (at most) order $h-1$ in h-step-ahead forecast errors, and

[*] Corresponding author.
E-mail addresses: a.patton@lse.ac.uk (A.J. Patton), atimmerm@ucsd.edu (A. Timmermann).

0304-4076/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2006.07.018

non-decreasing forecast error variance as the forecast horizon grows. Although such properties seem sensible, they are in fact established under a set of very restrictive assumptions on the forecaster's loss function. While bias in the optimal forecast under asymmetric loss has been established by Granger (1969, 1999) and characterized analytically for certain classes of loss functions and forecast error distributions by Christoffersen and Diebold (1997), failure of the remaining optimality properties has not previously been demonstrated.[1]

In this paper we first show that, in general, asymmetric loss and nonlinear dynamics in the data generating process (DGP) leads to violations of standard properties that an optimal forecast must possess. We next provide new analytical results under linear-exponential (linex) loss and a Markov switching DGP. Linex loss is frequently used in the literature on asymmetric loss, c.f. Batchelor and Peel (1998), Varian (1974), Zellner (1986), Sancetta and Satchell (2004) and Elliott and Timmermann (2004). Similarly, following Hamilton (1989), Markov switching processes have been widely used as a way to identify nonlinear dynamics in macroeconomic and financial time series. We show that not only can the optimal forecast be biased, but forecast errors can be serially correlated of arbitrarily high order and both the unconditional and conditional forecast error variance may be *decreasing* functions of the forecast horizon. We thus demonstrate that none of the properties traditionally associated with tests of optimal forecasts carry over to a more general setting with asymmetric loss and possible nonlinear dynamics in the DGP.[2]

Our results are of particular interest due to the revived interest in forecasting and econometric modeling under asymmetric loss, c.f. Christoffersen and Diebold (1996, 1997), Diebold (2004), Granger and Newbold (1986), Granger and Pesaran (2000), Nobay and Peel (2003) and Skouras (2006). When coupled with common findings of nonlinearities and dynamics in higher order moments, e.g. in the form of ARCH effects, in most macroeconomic and financial time series this means that the typical situation facing an economic forecaster may be well represented by our stylized example.

The insights provided in the paper also have important empirical implications. For example, it is common to test for rationality by testing for serial correlation in forecast errors.[3] Our paper shows that this implication of rationality fails to hold under even minor deviations from symmetric loss when combined with nonlinear dynamics in the DGP, and that serial correlation can be expected in the forecast errors from optimal forecasts under very plausible assumptions. This suggests that many of the conclusions in the empirical literature concerning suboptimality of forecasts may be premature.

The contributions of our paper are as follows. First, we establish conditions under which the standard properties that an optimal forecast possesses under MSE loss fail to hold.

[1] Under asymmetric loss functions such as lin–lin and linex and assuming a conditionally Gaussian process, Christoffersen and Diebold (1997) characterize the optimal bias analytically. Their study does not, however, consider the other properties of optimal forecast errors such as limited serial correlation and non-decreasing variance.

[2] Hoque et al. (1988), and Magnus and Pesaran (1989, 1991) discuss violations of the standard properties of optimal forecasts caused by estimation error, rather than by a choice of a loss function different from MSE. In this paper we focus on the case of zero estimation error to rule this out as a cause of apparent violations. However, we also discuss the effect of estimation error on forecasts in Section 6.

[3] The literature on forecast rationality testing is extensive. For a list of references to papers that use the Survey of Professional Forecasters data set maintained by the Federal Reserve Bank of Philadelphia, see ⟨http://www.phil.frb.org/econ/spf/spfbib.html⟩.

Second, we provide tractable, analytical results for the multi-step forecast errors under a nonlinear DGP in common use and a popular way to represent asymmetric loss. Due to difficulties with time-aggregation of many nonlinear processes, such as GARCH processes, this has not previously been done. Third, we demonstrate, again through analytical results, that the "generalized forecast error" (which is a simple transformation of the forecast error) inherits some of the standard properties such as lack of bias and serial correlation. Finally, we provide empirical examples that demonstrate the importance of these effects using data on stock returns and output growth and we show how to account for the effect of parameter estimation errors on the optimal forecast.

The outline of the paper is as follows. Section 2 reviews the properties of optimal forecasts under MSE loss. Section 3 provides some general results on forecast optimality under asymmetric loss. Given the difficulty in obtaining results in very general cases, Section 4 provides detailed analytical results for a particular combination of asymmetric loss and DGP, establishing the violation of the standard properties on (absence of) bias and serial correlation in forecast errors. Section 5 shows which properties actually do hold for a transformation of the forecast error known as the generalized forecast error. Section 6 discusses the impact of estimation error on our results, and Section 7 provides empirical illustrations using data on stock returns and GDP growth. Section 8 concludes. An Appendix contains proofs.

2. Optimality properties under MSE loss

We are interested in studying optimal h-period ahead forecasts of some univariate time-series process, $\{Y_\tau\}$, computed conditional on some information set, $\mathscr{F}_t$. The information set is assumed to include all lagged values of Y_t and possibly lagged values of other variables. The conditional distribution function of Y_{t+h} given $\mathscr{F}_t$ is $F_{t+h,t}$ while the associated density is $f_{t+h,t}$. Generic forecasts are denoted by $\hat{Y}_{t+h,t}$ and forecast errors are $e_{t+h,t} = Y_{t+h} - \hat{Y}_{t+h,t}$, giving rise to a loss, $L(Y_{t+h}, \hat{Y}_{t+h,t})$. Under MSE loss, $L(Y, \hat{Y}) = L(e) = e^2$. The optimal h-period forecast computed at time t is denoted by $\hat{Y}^*_{t+h,t}$ and its forecast error is $e^*_{t+h,t}$.

An optimal forecast (in a population sense) is defined as the value of $\hat{Y}_{t+h,t}$ that, conditional on the information set at time t and taking the parameter values and the DGP to be known, minimizes the expected loss:

$$\hat{Y}^*_{t+h,t} \equiv \underset{\hat{Y}_{t+h,t}}{\arg\min}\, \mathrm{E}[L(Y_{t+h}, \hat{Y}_{t+h,t})|\mathscr{F}_t] \tag{1}$$

$$= \underset{\hat{Y}_{t+h,t}}{\arg\min} \int L(y, \hat{Y}_{t+h,t})\, \mathrm{d}F_{t+h,t}(y). \tag{2}$$

Provided that the forecast errors are covariance stationary,[4] under MSE loss optimal forecasts satisfy the following four properties (c.f. Diebold and Lopez, 1996):

1. Forecasts are unbiased (so the forecast error, $e^*_{t+h,t}$, has zero mean).
2. The variance of the forecast error ($e^*_{t+h,t}$) is a non-decreasing function of the forecast horizon, h.

[4] This is implied by, but does not itself require, covariance stationarity of $[Y_{t+h}, \hat{Y}_{t+h,t}]'$. For example, both Y_{t+h} and $\hat{Y}_{t+h,t}$ could contain unit roots, but the linear combination $Y_{t+h} - \hat{Y}_{t+h,t}$ still be covariance stationary.

3. h-period forecast errors ($e^*_{t+h,t}$) are not correlated with information dated by h periods or more ($\mathscr{F}_t$). Hence the h-period forecast error is at most serially correlated of order $h-1$.
4. Single-period forecast errors ($e^*_{t+1,t}$) are serially uncorrelated.

These properties are easily derived under MSE loss. Suppose that Y_t is covariance stationary, and without loss of generality assume a zero unconditional mean. Wold's representation theorem then establishes that it can be represented as a linear combination of serially uncorrelated white noise terms:

$$Y_t = \sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i}, \tag{3}$$

where $\varepsilon_t = Y_t - \mathscr{P}(Y_t | y_{t-1}, y_{t-2}, \ldots)$ is the projection error which is serially uncorrelated white noise, $WN(0, \sigma^2)$. The h-period forecast thus becomes

$$\hat{Y}^*_{t+h,t} = \sum_{i=0}^{\infty} \theta_{h+i} \varepsilon_{t-i}. \tag{4}$$

This is an optimal forecast provided that ε_t is Gaussian or can alternatively be viewed as an optimal linear forecast given $\{y_{t-1}, y_{t-2}, \ldots\}$. The associated forecast errors are

$$e^*_{t+h,t} = \sum_{i=0}^{h-1} \theta_i \varepsilon_{t+h-i}. \tag{5}$$

It follows directly that

$$\mathrm{E}[e^*_{t+h,t}] = 0,$$

$$Var(e^*_{t+h,t}) = \sigma^2 \sum_{i=0}^{h-1} \theta_i^2,$$

$$Cov(e^*_{t+h,t}, e^*_{t+h-j,t-j}) = \begin{cases} \sigma^2 \sum_{i=j}^{h-1} \theta_i \theta_{i-j} & \text{for } j < h, \\ 0 & \text{for } j \geqslant h. \end{cases} \tag{6}$$

This accounts for properties 1–4, respectively.

3. Some general results

It is difficult to establish general results for the properties of optimal forecasts under nonlinear DGP and asymmetric loss. The reason is that the form of the DGP interacts with the shape of the loss function in determining the optimal forecast. Assuming that $L(.)$ only depends on the forecast error, $e = Y - \hat{Y}$, is analytic and hence can be represented by a power series (c.f. Rudin, 1964, p. 158), this interaction is perhaps best seen from a simple Taylor series expansion of the loss function $L(e)$ about the point μ:

$$\mathrm{E}[L(e_{t+h})|\mathscr{F}_t] = L(\mu) + L'(\mu)(\mathrm{E}[e_{t+h}|\mathscr{F}_t] - \mu) + \sum_{m=2}^{\infty} L^{(m)}(\mu) \sum_{i=0}^{m} \frac{(-1)^i}{i!(m-i)!} \mu^i \mathrm{E}[e_{t+h}^{m-i}|\mathscr{F}_t],$$

where $L^{(m)}(\mu)$ is the mth derivative of L evaluated at $e=\mu$. Clearly all higher order moments of the predictive density as well as higher order derivatives of the loss function will matter in determining the optimal forecast. Conditions on both the shape of the loss function and the density function are therefore required to obtain analytical results.

The first optimality property (unbiasedness) is a direct consequence of the first-order condition characterizing an optimal forecast under MSE loss, i.e. $\mathrm{E}_t[\partial(Y_{t+h}-\hat{Y}^*_{t+h,t})^2/\partial\hat{y}]=-2\mathrm{E}_t[L(Y_{t+h}-\hat{Y}^*_{t+h,t})]=0$, so that $\hat{Y}^*_{t+h,t}=\mathrm{E}_t[Y_{t+h}]$, where $\mathrm{E}_t[\cdot]$ is short for $\mathrm{E}[\cdot|\mathscr{F}_t]$. It need not hold for other symmetric families of loss functions. For example under absolute error loss (so $L(e)\propto|e|$), the optimal forecast is the conditional median of Y_{t+h}. This will differ from the conditional mean unless f is symmetric. Hence double symmetry, i.e. symmetric loss *and* a symmetric conditional density, $f_{t+h,t}$, are required to obtain unbiasedness of the optimal forecast.[5]

To better understand the properties that an optimal forecast must have under more general conditions than the standard ones, assume that the loss function, $L(.)$, can be represented as the sum of a symmetric component, $\tilde{L}(.)$, and a component, $\zeta(.)$, that captures the asymmetry:

$$L(e)=\tilde{L}(e)+\zeta(e).$$

Without loss of generality, assume that $\zeta(.)$ only applies to positive forecast errors:

Assumption L1. $\zeta=0$ for $e\leqslant 0$, and $\zeta\geqslant 0$ for $e>0$ with ζ convex and $\zeta'>0$ on a set with probability greater than zero.

Assumption L2. $\tilde{L}(e)$ is convex.

The results can easily be extended to modify these assumptions. All standard asymmetric loss functions in common use, including lin–lin and linex, can be written in this form.

Furthermore, let the DGP for Y_{t+h} take the form

$$Y_{t+h}=\mu_{t+h,t}+\sigma_{t+h,t}\varepsilon_{t+h}, \tag{7}$$

where $\varepsilon_{t+h}|\mathscr{F}_t{\sim}F_{t+h,t}$

$$\mathrm{E}[\varepsilon_{t+h}|\mathscr{F}_t]=0,\quad \mathrm{E}[\varepsilon^2_{t+h}|\mathscr{F}_t]=1,$$

where $\mu_{t+h,t}=\mathrm{E}_t[Y_{t+h}]$ and $\sigma^2_{t+h,t}=\mathrm{E}_t[(Y_{t+h}-\mu_{t+h,t})^2]$ are the conditional mean and variance given information at time $t,\mathscr{F}_t$. We shall also be making use of the following assumptions on the DGP:

Assumption D1. ε_{t+h} has a symmetric and unimodal conditional density, $f_{t+h,t}$.

Assumption D2. The conditional second moment of Y_{t+h} is time-varying, i.e.,

$$\mathrm{E}[(Y_{t+h}-\mathrm{E}[Y_{t+h}|\mathscr{F}_t])^2|\mathscr{F}_t]\neq\mathrm{E}[(Y_{t+h}-\mathrm{E}[Y_{t+h}|\mathscr{F}_t])^2]\text{ for some }t,\ \text{ and all }h<\infty$$

Assumption D3. Y_{t+h} exhibits declining "volatility of volatility". That is, $V[\sigma^2_{t+h,t}]$ is a decreasing function of h.

[5] More specifically, Granger (1969) has shown that the conditional expectation of Y_{t+h} is an optimal forecast provided that $L(e)$ is symmetric about $e=0$ and $f_{t+h,t}$ is symmetric about the conditional mean and either of two conditions hold: (1) $L'(e)$ exists almost everywhere and is strictly increasing; or (2) $f_{t+h,t}$ is continuous and unimodal.

D1 is the traditional density symmetry assumption. We use "unimodal" in the sense that $f_{t+h,t}$ has a unique local maximum. Time-varying conditional variance is widely observed in both financial and macroeconomic data, and so D2 is not a particularly strong assumption. Assumption D3 is satisfied by many common volatility models such as the GARCH(1, 1) process and the Markov switching volatility model, as we show in the next section.

Using these assumptions we have the following result:

Proposition 1. (1) *Let assumptions* D1, L1 *and* L2 *hold. Then the optimal forecast*, $\hat{Y}^*_{t+h,t}$, *is biased.*

(2) *Let assumptions* D1, D2, L1 *and* L2 *hold. Then, generically* $\mathrm{E}[e^*_{t+h,t}|Z_t] \neq \mathrm{E}[e^*_{t+h,t}]$ *for some* $Z_t \in \mathscr{F}_t$, *and so the optimal forecast error is predictable using* $\mathscr{F}_t$.

(3) *Let assumptions* D2 *and* D3 *hold. Then the optimal forecast error variance need not be a weakly increasing function of the forecast horizon*, *h*.

This proposition, along with other results, is proved in the Appendix. The assumed DGP is nonlinear in the sense that there is dynamics in the second moment of the process, as indicated by the model in Eq. (7) if $\sigma_{t+h,t}$ varies with t, but the conditional mean is left unrestricted. Predictability of the forecast error, $e^*_{t+h,t}$, may or may not take the form of serial correlation. We emphasize that the conditions used to establish the results are sufficient but not necessary—for example it is possible to construct examples where the optimal forecast is biased when both the density and the loss function are asymmetric, so D1 fails to hold. Nevertheless, Proposition 1 establishes basic conditions under which violations of the standard properties of an optimal forecast can be shown.

While the first part of Proposition 1 follows readily, establishing the second and third points requires more structure on the problem due to the interaction between higher order moments and higher order derivatives of the loss function. Even simple types of nonlinearities and asymmetric loss tend to lead to complicated expressions for expected loss that are difficult to evaluate. For example, our assumptions on the signs of the derivatives, $\partial\hat{Y}^*/\partial\sigma > 0$ and $\partial^2\hat{Y}^*/(\partial\sigma\partial\kappa) \leqslant 0$, are not primitive and cannot be derived without substantially more structure on the DGP and loss function.

The absence of a set of analytical results demonstrating the second and third points in Proposition 1 poses a severe limitation to our understanding of the factors contributing to violations of standard optimality properties of an optimal forecast, let alone their empirical relevance. For this reason we next provide analytical results with reasonable assumptions about the forecaster's loss function and a nonlinear DGP which show that each of the optimality properties (1)–(4) cease to be valid when the assumption of MSE loss is relaxed.

4. Asymmetric loss and a nonlinear process

We establish our results in the context of the linex loss function that allows for asymmetries

$$L(Y_{t+h} - \hat{Y}_{t+h,t}; a) = \exp\{a(Y_{t+h} - \hat{Y}_{t+h,t})\} - a(Y_{t+h} - \hat{Y}_{t+h,t}) - 1, \quad a \neq 0. \tag{8}$$

This loss function has been used extensively to demonstrate the effect of asymmetric loss, see e.g. Varian (1974), Zellner (1986) and Christoffersen and Diebold (1997). It is included in the family of loss functions in Proposition 1. To see this, note that (8) can be written as

$$L(e) = \tilde{L}(e) + \zeta(e),$$

where

$$\tilde{L}(e) = \exp\{-a|e|\} + a|e| - 1$$

$$\zeta(e) = (\exp\{ae\} - \exp\{-ae\} - 2ae)\mathbf{1}(e>0) \tag{9}$$

where $\mathbf{1}(e>0)$ is an indicator function equaling one if $e>0$, zero otherwise.

Assuming that we may interchange the expectation and differentiation operators, the first-order condition for the optimal forecast, $\hat{Y}^*_{t+h,t}$, under linex loss takes the form

$$\mathrm{E}_t\left[\frac{\partial L(Y_{t+h} - \hat{Y}^*_{t+h,t}; a)}{\partial \hat{Y}_{t+h,t}}\right] = a - a\mathrm{E}_t[\exp\{a(Y_{t+h} - \hat{Y}^*_{t+h,t})\}] = 0. \tag{10}$$

We derive analytical expressions for the optimal forecast and the expected loss using a popular nonlinear DGP, namely a regime switching model of the type proposed by Hamilton (1989).[6] Thus, suppose that $\{Y_t\}$ is generated by a simple Gaussian mixture model, with constant mean and volatility driven by some underlying state process, S_t, that takes a finite number (k) of values:

$$Y_{t+1} = \mu + \sigma_{s_{t+1}} v_{t+1},$$

$$v_{t+1} \sim \text{i.i.d. } \mathrm{N}(0,1),$$

$$S_{t+1} = 1, \ldots, k. \tag{11}$$

We assume that the state indicator function, S_{t+1}, is independently distributed of all past, current and future values of v_{t+1}. To establish analytical results we assume that the parameters of this DGP are known but we allow S_t to be unobservable, i.e. S_t is *not* adapted to $\mathcal{F}_t$. The state-specific means and variances can be collected in $k \times 1$ vectors, $\boldsymbol{\mu} = \mu\boldsymbol{\iota}$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)'$, where $\boldsymbol{\iota}$ is a $k \times 1$ vector of ones. Conditional on a given realization of the future state variable, $S_{t+1} = s_{t+1}$, Y_{t+1} is Gaussian with mean μ and variance $\sigma^2_{s_{t+1}}$, but future states are unknown at time t so Y_{t+1} can be strongly non-Gaussian given current information, $\mathcal{F}_t$.

At each point in time the state variable, S_{t+1}, takes an integer value between 1 and k. Following Hamilton (1989), we assume that the states are generated by a first-order

[6] Christoffersen and Diebold (1997) characterize analytically the optimal bias under linex loss and a conditionally Gaussian process with ARCH disturbances. They derive analytically the optimal time-varying bias as a function of the conditional variance. For our purposes, however, this process is less well-suited to show violation of all four properties forecast errors have in the standard setting since this requires characterizing the forecast error distribution at many different horizons, h. The problem is that while the one-step-ahead forecast error distribution is Gaussian for a GARCH(1,1) process, this typically does not hold at longer horizons, c.f. Drost and Nijman (1993).

stationary and ergodic Markov chain with transition probability matrix

$$\mathbf{P}(s_{t+1}|s_t) = \mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & p_{k-1k} \\ p_{k1} & \cdots & p_{kk-1} & p_{kk} \end{bmatrix}, \tag{12}$$

where each row of **P** sums to one. To allow for unobservable states, define a $k \times 1$ random vector, ξ_t, of zeros except for the jth element which is equal to unity if $s_t = j$:

$$\xi_t = \begin{cases} (1\ 0\ 0 \cdots 0)' & \text{for } s_t = 1, \\ (0\ 1\ 0 \cdots 0)' & \text{for } s_t = 2, \\ \vdots & \vdots \\ (0\ 0 \cdots 0\ 1)' & \text{for } s_t = k. \end{cases} \tag{13}$$

It follows from this definition and (12) that $\mathrm{E}[\xi_{t+1}|\xi_t] = \mathbf{P}\xi_t$. However, when states are unobserved this is not the relevant conditioning information set which is instead often given by past realizations, i.e. $\mathcal{F}_t = \{y_t, y_{t-1}, \ldots, y_1\}$. In cases where the parameters, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P})$, are known, the unknown state probabilities can be derived as follows. Define the vectors of conditional state probabilities as $\hat{\xi}_{t|t} = (Pr(s_t = 1|\mathcal{F}_t; \boldsymbol{\theta})\ Pr(s_t = 2|\mathcal{F}_t; \boldsymbol{\theta})\ Pr(s_t = k|\mathcal{F}_t; \boldsymbol{\theta}))'$, $\hat{\xi}_{t+1|t} = (Pr(s_{t+1} = 1|\mathcal{F}_t; \boldsymbol{\theta})\ \ Pr(s_{t+1} = 2|\mathcal{F}_t; \boldsymbol{\theta})\ Pr(s_{t+1} = k|\mathcal{F}_t; \boldsymbol{\theta}))'$ and let $\boldsymbol{\eta}_t = (f(y_t|s_t = 1, \mathcal{F}_{t-1}, \boldsymbol{\theta}),\ f(y_t|s_t = 2, \mathcal{F}_{t-1}, \boldsymbol{\theta}), \ldots, f(y_t|s_t = k, \mathcal{F}_{t-1}, \boldsymbol{\theta}))'$ be a vector of probability densities conditional on the underlying states and past information, $\mathcal{F}_{t-1}$. For a given starting value, $\hat{\xi}_{1|0}$, an optimal estimate of the unknown state probabilities can then be derived by iterating on the equations (c.f. Hamilton, 1994)

$$\hat{\xi}_{t|t} = \frac{(\hat{\xi}_{t|t-1} \odot \boldsymbol{\eta}_t)}{\boldsymbol{\iota}'(\hat{\xi}_{t|t-1} \odot \boldsymbol{\eta}_t)},$$

$$\hat{\xi}_{t+1|t} = \mathbf{P}\hat{\xi}_{t|t}, \tag{14}$$

where $\odot$ represents the Hadamard (element-by-element) product and $\boldsymbol{\iota}$ is a $k \times 1$ vector of ones. We refer to $\bar{\xi}$ as the vector of unconditional or ergodic state probabilities that solves the equation $\bar{\xi}'\mathbf{P} = \bar{\xi}'$. Uncertainty about the parameters can be handled in a Bayesian context as we show in Section 6.

4.1. Unbiasedness

We will consider an idealized h-step-ahead forecasting problem where, in addition to knowing that the form of the DGP is the Markov switching process in (11) and (12), the forecaster is assumed to also know the parameters of this DGP, removing estimation error from the problem. Forecasts in this example are thus perfectly optimal and easy to construct: neither estimation error nor irrationality is driving our results.

Using the conditional normality of v_{t+h} and letting $\hat{\xi}_{(s_{t+h}|t)}$ be the s_{t+h}th element of the vector $\hat{\boldsymbol{\xi}}_{t+1|t}$, i.e. the probability that $S_{t+h} = s_{t+h}$ given $\mathscr{F}_t$, the expected loss is[7]

$$\begin{aligned}
\mathrm{E}_t[L(e_{t+h,t}; a)] &= \mathrm{E}_t[\exp\{a(Y_{t+h} - \hat{Y}_{t+h,t})\}] - a\mathrm{E}_t[Y_{t+h}] + a\hat{Y}_{t+h,t} - 1 \\
&= \sum_{s_{t+h}=1}^{k} \hat{\xi}_{(s_{t+h}|t)} \mathrm{E}_t[\exp\{a(Y_{t+h} - \hat{Y}_{t+h,t})\} | S_{t+h} = s_{t+h}] \\
&\quad - a \sum_{s_{t+h}=1}^{k} \hat{\xi}_{(s_{t+h}|t)} \mathrm{E}_t[Y_{t+h} | S_{t+h} = s_{t+h}] + a\hat{Y}_{t+h,t} - 1 \\
&= \hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h \exp\left\{a\boldsymbol{\mu} - a\iota\hat{Y}_{t+h,t} + \frac{a^2}{2}\boldsymbol{\sigma}^2\right\} - a\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h\boldsymbol{\mu} + a\hat{Y}_{t+h,t} - 1. \qquad (15)
\end{aligned}$$

Differentiating with respect to $\hat{Y}_{t+h,t}$ and setting the resulting expression equal to zero gives the first-order condition

$$1 = \hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h \exp\left\{a\boldsymbol{\mu} - a\iota\hat{Y}^*_{t+h,t} + \frac{a^2}{2}\boldsymbol{\sigma}^2\right\}.$$

Solving for $\hat{Y}^*_{t+h,t}$ we get an expression that is easier to interpret:

$$\hat{Y}^*_{t+h,t} = \mu + \frac{1}{a}\log(\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h\boldsymbol{\varphi}), \qquad (16)$$

where $\boldsymbol{\varphi} \equiv \exp\{(a^2/2)\boldsymbol{\sigma}^2\}$. The h-step forecast error associated with the optimal forecast, denoted $e^*_{t+h,t}$, is

$$e^*_{t+h,t} = \sigma_{s_{t+h}} v_{t+h} - \frac{1}{a}\log(\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h\boldsymbol{\varphi}).$$

This expression makes it easy for us to establish the violation of property 1 (unbiasedness) in our setup:

Proposition 2. *The unconditional and conditional bias in the optimal forecast error arising under linex loss* (8) *for the Markov switching process* (11)–(12) *is given by*:

$$\mathrm{E}_t[e^*_{t+h,t}] = -\frac{1}{a}\log(\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h\boldsymbol{\varphi}),$$

$$\mathrm{E}\left[e^*_{t+h,t}\right] = -\frac{1}{a}\bar{\boldsymbol{\xi}}'\boldsymbol{\lambda}_h \to -\frac{1}{a}\log(\bar{\boldsymbol{\xi}}'\boldsymbol{\varphi}) \quad \text{as } h \to \infty, \qquad (17)$$

where $\boldsymbol{\lambda}_h \equiv \log(\mathbf{P}^h\boldsymbol{\varphi})$. *Thus, generically, the optimal forecast is conditionally and unconditionally biased at all forecast horizons, h, and the bias persists even as h goes to infinity.*

The proof of the proposition is given in the Appendix. For purposes of exposition, we present some results for a specific form of the loss function ($a = 1$) and regime switching process:

$$\boldsymbol{\mu} = [0, 0]',$$

$$\boldsymbol{\sigma} = [0.5, 2]',$$

[7] All $\exp\{\cdot\}$ and $\log(\cdot)$ operators are applied element-by-element to vector and matrix arguments.

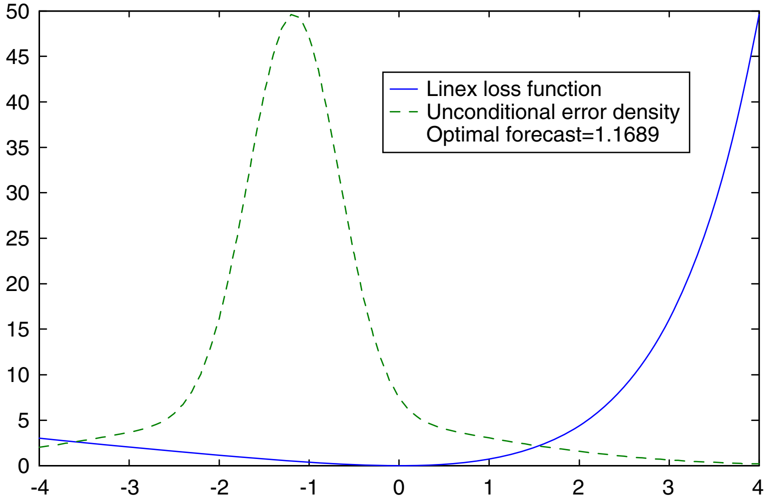

Fig. 1. Linear–exponential loss function and unconditional optimal forecast error density, two-state regime switching example.

$$\mathbf{P} = \begin{bmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{bmatrix},$$

$$\text{so} \quad \bar{\xi} = \left[\frac{2}{3}, \frac{1}{3}\right]'$$

The unconditional mean of Y_t is zero, and the unconditional variance is $\bar{\xi}'\boldsymbol{\sigma}^2 = 1.5$. This parameterization is not dissimilar to empirical results frequently obtained when this model is estimated on financial data, see Section 7 for example. For this particular parameterization the optimal unconditional bias in $e^*_{t+1,t}$ is -1.17, indicating that it is optimal to over-predict. Fig. 1 shows the unconditional density of $e_{t+h,t}$ and also plots the linex loss function. The density function has been re-scaled so as to match the range of the loss function. This figure makes it clear why the optimal bias is negative: the linex loss function with $a = 1$ penalizes positive errors (under-predictions) more heavily than negative errors (over-predictions). The optimal forecast is in the tail of the unconditional distribution of Y_t: the probability mass to the right of the optimal forecast is only 10%. Under symmetric loss the optimal forecast is the mean, and so for symmetric distributions the amount of probability mass either side of the forecast would be 50%. In Fig. 2 we plot the optimal unconditional bias as a function of the forecast horizon (using the steady-state weights as initial probabilities), as well as the optimal bias when the state is known. The unconditional bias for this case is an increasing (in absolute value) function of h and asymptotes to -1.17.

4.2. Non-decreasing variance

We next demonstrate the violation of property 2 (non-decreasing variance). Let $\odot$ be the Hadamard (element-by-element) product. The result is as follows:

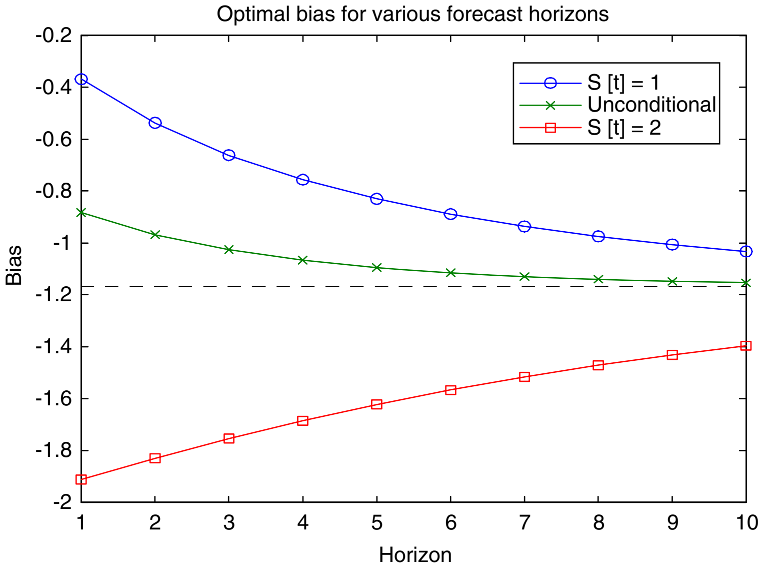

Fig. 2. Bias in the optimal forecast for various forecast horizons, two-state regime switching example.

Proposition 3. *The variance of the forecast error arising under linex loss* (8) *for the Markov switching process* (11)–(12) *associated with the optimum forecast is given by*

$$Var(e^*_{t+h,t}) = \bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2 + \frac{1}{a^2}\boldsymbol{\lambda}'_h((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}') \odot \mathbf{I} - \bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}')\boldsymbol{\lambda}_h. \tag{18}$$

This variance need not be a decreasing function of the forecast horizon, h. In the limit as h goes to infinity, the forecast error variance converges to the steady-state variance, $\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2$.

Conversely, under MSE loss, the unconditional forecast error variance is constant across forecast horizons, h:

$$\begin{aligned} Var(e^*_{t+h,t}) &= \mathrm{E}[\sigma^2_{s_{t+h}} v^2_{t+h}] \\ &= \sum_{s_{t+h}=1}^{k} \bar{\xi}_{(s_{t+h})}\sigma^2_{s_{t+h}}\mathrm{E}[v^2_{t+h}|S_{t+h} = s_{t+h}] \\ &= \bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2, \end{aligned}$$

where $\bar{\xi}_{(s_{t+h})}$ is the s_{t+h}th element of $\bar{\boldsymbol{\xi}}$.

It is also common to consider the MSE of the forecast. This is closely related to the forecast error variance but differs by the squared bias. The corresponding result for the mean squared forecast error (MSFE) is as follows:

Corollary 1. *The MSFE arising under linex loss* (8) *for the Markov switching process* (11)–(12) *associated with the optimum forecast is given by*

$$MSFE(e^*_{t+h,t}) = \bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2 + \frac{1}{a^2}\boldsymbol{\lambda}'_h((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}') \odot \mathbf{I})\boldsymbol{\lambda}_h. \tag{19}$$

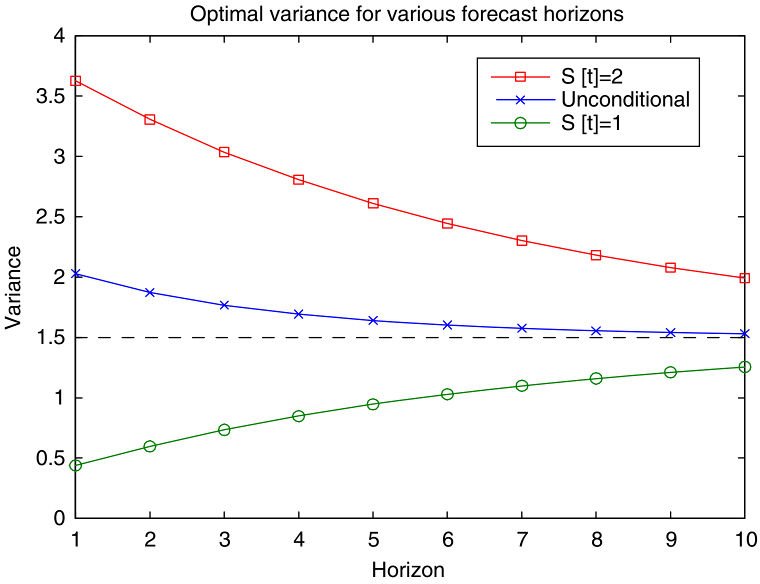

Fig. 3. Variance of the optimal h-step forecast error for various forecast horizons, two-state regime switching example.

The MSFE need not be a decreasing function of the forecast horizon, h. *In the limit as* h *goes to infinity*, *the MSFE converges to* $\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2 + ((1/a)\log(\bar{\boldsymbol{\xi}}'\boldsymbol{\varphi}))^2$.

A surprising implication of Proposition 3 is that it is not always true that $Var(e^*_{t+h,t})$ will converge to $\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2$ from below, that is, $Var(e^*_{t+h,t})$ need not be increasing in h. Depending on the form of $\mathbf{P}$ and $\boldsymbol{\sigma}^2$, it is possible that $Var(e^*_{t+h,t})$ actually *decreases* towards the unconditional variance of Y_t. Corollary 1 shows that a similar result is true for the MSFE.

Using the numerical example described above, the unconditional variance of the optimal forecast error as a function of the forecast horizon is shown in Fig. 3. For comparison the figure also shows the optimal forecast error variance when the state is observable. It is clearly possible that the forecast error at the distant future has a lower variance than at the near future.[8] The reason for this surprising result lies in the mismatch of the forecast objective function, L, and the variance of the forecast error, $Var(e_{t+h,t})$, which does not occur when using quadratic loss (see next section). A similar mismatch of the objective function and the performance metric has been discussed by Christoffersen and Jacobs (2004), Corradi and Swanson (2002) and Sentana (2005).

Using the expression for $Var(e^*_{t+h,t})$ in Proposition 3, we can consider two interesting special cases. First, suppose that $\sigma_1 = \sigma_2 = \sigma$ so the variable of interest is i.i.d. normally distributed with constant mean and variance. In this case we have

$$Var(e^*_{t+h,t}) = \bar{\boldsymbol{\xi}}'\boldsymbol{\iota}\sigma^2 + \frac{1}{a^2}\log(\bar{\boldsymbol{\xi}}'\boldsymbol{\varphi})\boldsymbol{\iota}'((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}')\odot\mathbf{I} - \bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}')\boldsymbol{\iota}\log(\bar{\boldsymbol{\xi}}'\boldsymbol{\varphi}) = \sigma^2.$$

[8] Using the same numerical example it can be shown that the MSFE decreases when moving from $h = 1$ to 2, but increases with h for $h \geqslant 2$. We do not report this figure in the interests of parsimony.

And so the optimal forecast error variance is constant for all forecast horizons as we would expect.

The second special case arises when the transition matrix takes the form $\mathbf{P} = \boldsymbol{\iota}\bar{\xi}'$. That is, the probability of being in a particular state is independent of past information, so the density of the variable of interest is a constant mixture of two normal densities and thus is i.i.d but may exhibit arbitrarily high kurtosis. In this case we have $\boldsymbol{\lambda}_h = \boldsymbol{\iota}\log(\bar{\xi}'\boldsymbol{\varphi})$ for all h, so

$$\begin{aligned} Var(e^*_{t+h,t}) &= \bar{\xi}'\boldsymbol{\sigma}^2 + \frac{1}{a^2}\log(\bar{\xi}'\boldsymbol{\varphi})\boldsymbol{\iota}'((\bar{\xi}\boldsymbol{\iota}') \odot \mathbf{I} - \bar{\xi}\bar{\xi}')\boldsymbol{\iota}\log(\bar{\xi}'\boldsymbol{\varphi}) \\ &= \bar{\xi}'\boldsymbol{\sigma}^2. \end{aligned}$$

Thus the optimal forecast error variance is constant for all forecast horizons. This special case shows that it is not the fat tails of the mixture density that drives the curious result regarding decreasing forecast error variance in our example. Rather, it is the combination of asymmetric loss and persistence in the conditional variance. This is consistent with Proposition 1 from the previous section.

In Proposition 1 we obtained a more general non-decreasing variance result by making the high-level assumption that the volatility of volatility for the variable of interest is declining in h, i.e. that $V[\sigma^2_{t+h,t}]$ is decreasing in h. Below we derive the volatility of volatility for the Markov switching process in Eqs. (11)–(12).

Proposition 4. *The volatility of volatility for the Markov switching process* (11)–(12) *is given by*

$$V[V[Y_{t+h}|\mathscr{F}_t]] = \bar{\xi}'(P^h\boldsymbol{\sigma}^2\boldsymbol{\sigma}^{2\prime}P^{h\prime} \odot I)\boldsymbol{\iota} - \bar{\xi}'P^h\boldsymbol{\sigma}^2\boldsymbol{\sigma}^{2\prime}P^{h\prime}\bar{\xi}$$

and asymptotes to zero as $h \to \infty$.

The volatility of volatility under the numerical example described above is presented in Fig. 4, and clearly shows that it is decreasing with the forecast horizon.

4.3. Serial correlation in optimal forecast errors

In the standard linear, quadratic loss framework an optimal h-step forecast is an *MA* process of order no greater than $(h-1)$. This implies that all autocovariances beyond the $(h-1)$th lag are zero. This particular property of an optimal forecast is usually tested by regressing the observed forecast error on its own lagged values of order h and higher:

$$e_{t+h,t} = \alpha_0 + \sum_{i=0}^{n}\beta_i e_{t-i,t-i-h} + u_{t+h,t}.$$

Here it is assumed that $u_{t+h,t}$ has mean zero and is serially uncorrelated. Under MSE loss and forecast rationality, $\alpha_0 = \beta_0 = \cdots = \beta_n = 0$.

Outside the standard setting with MSE loss this property need no longer hold. In particular, we have:

Proposition 5. *The h-step-ahead forecast error arising under linex loss* (8) *for the Markov switching process* (11)–(12) *is serially correlated with autocovariance*

$$Cov[e^*_{t+h,t}, e^*_{t+h-j,t-j}] = \bar{\xi}'\boldsymbol{\sigma}^2\mathbf{1}_{\{j=0\}} + \frac{1}{a^2}\boldsymbol{\lambda}_h'((\bar{\xi}\boldsymbol{\iota}') \odot \mathbf{P}^j - \bar{\xi}\bar{\xi}')\boldsymbol{\lambda}_h. \tag{20}$$

Although this converges to zero as j goes to infinity, it can be non-zero at lags larger than h.

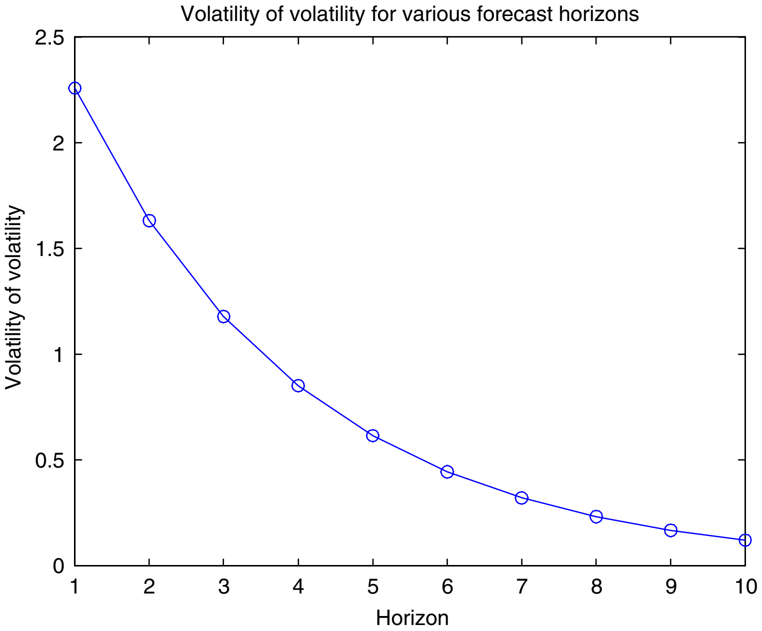

Fig. 4. Unconditional variance of the h-step conditional variance under the Markov switching DGP, for various forecast horizons.

Using the same parameterization as in the earlier example, the autocorrelation function for various forecast horizons is presented in Fig. 5. Note the strong autocorrelation even at lags much longer than the forecast horizon, h. Thus the optimal forecast error in our set-up need not follow an $MA(h-1)$ process and the one-step-ahead forecast error need not be serially uncorrelated (property 3).[9]

It is easily verified that, under MSE loss, the autocovariance of the optimal forecast errors under the regime switching process is zero:

$$\begin{aligned} Cov(e^*_{t+h,t}, e^*_{t+h-j,t-j}) &= \mathrm{E}[\sigma_{s_{t+h-j}}\sigma_{s_{t+h}}v_{t+h-j}v_{t+h}] \\ &= \sum_{s_{t+h-j}=1}^{k}\sum_{s_{t+h}=1}^{k} \bar{\xi}_{(s_{t+h-j})}\hat{\xi}_{(s_{t+h}|s_{t+h-j})}\sigma_{s_{t+h-j}}\sigma_{s_{t+h}} \\ &\quad \times \mathrm{E}[v_{t+h-j}v_{t+h}|S_{t+h-j}=s_{t+h-j}, S_{t+h}=s_{t+h}] \\ &= 0 \quad \text{for } j\neq 0. \end{aligned}$$

Thus optimal forecast errors under MSE loss are conditionally and unconditionally unbiased, have constant unconditional variance as a function of the forecast horizon, and

[9] We can again consider the two special cases: iid Normal ($\sigma_1 = \sigma_2 = \sigma$), and iid mixture of normals ($\mathbf{P} = \boldsymbol{\iota}\bar{\xi}'$). Following the same logic as for the analysis of forecast error variance, it can be shown that in both of these cases the autocorrelation function equals zero for all lags greater than zero.

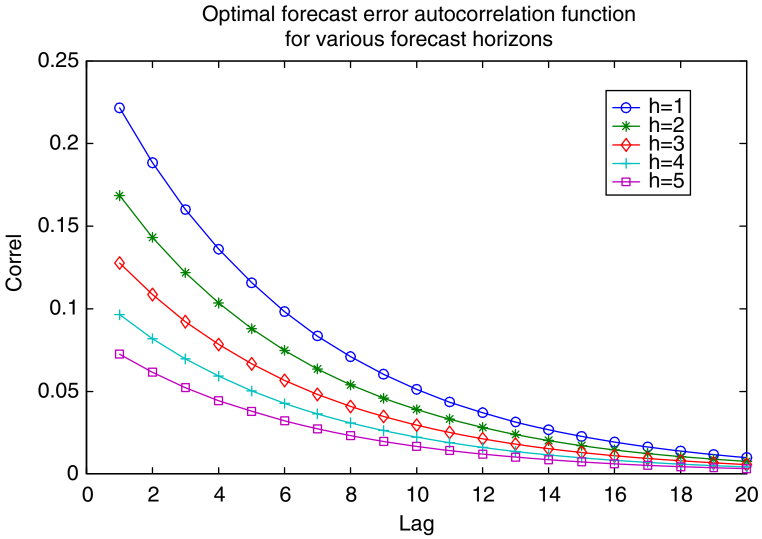

Fig. 5. Autocorrelation in the optimal *h*-step forecast error for various forecast horizons, two-state regime switching example.

are serially uncorrelated at all lags. This is because under MSE loss the optimal forecast error in this example is simply the term $\sigma_{s_{t+h}} v_{t+h}$, which is (heteroskedastic) white noise.

4.4. An alternative DGP

Readers might wonder whether the results established thus far are peculiar to the regime switching process in Eqs. (11)–(12). This is not the case as we show here for the most widely-used model for conditional variance, namely the GARCH(1,1) process, c.f. Bollerslev (1986):

$$Y_{t+1} = \mu + \varepsilon_{t+1},$$

$$\varepsilon_{t+1} = \sigma_{t+1} v_{t+1}, \quad v_{t+1} \sim \text{i.i.d. N}(0,1),$$

$$\sigma^2_{t+1} = \omega + \alpha\varepsilon^2_t + \beta\sigma^2_t, \tag{21}$$

where $\omega > 0, \alpha, \beta \geqslant 0$. We assume that $(\alpha+\beta)^2 + 2\alpha^2 < 1$, which is sufficient for both the unconditional second and fourth moments to exist:

$$\mathrm{E}[\sigma^2_{t+1}] = \frac{\omega}{1-\alpha-\beta},$$

$$\mathrm{E}[\sigma^4_{t+1}] = \frac{\omega^2(1+\alpha+\beta)}{(1-3\alpha^2-\beta^2-2\alpha\beta)(1-\alpha-\beta)}.$$

Under linex loss the one-step-ahead optimal forecast has a bias of $-(a/2)\sigma^2_{t+1}$, and so the optimal forecast error is $e^*_{t+1,t} = \sigma_{t+1}v_{t+1} - (a/2)\sigma^2_{t+1}$ and

$$\mathrm{E}[e^*_{t+1,t}] = \frac{-a\omega}{2(1-\alpha-\beta)}, \tag{22}$$

which establishes the violation of property 1 (unbiasedness). Using this equation the first order auto-covariance of the forecast error is

$$\begin{aligned} Cov(e^*_{t+1,t}, e^*_{t,t-1}) &= \mathrm{E}\left[\left(\sigma_{t+1}v_{t+1} - \frac{a}{2}\sigma^2_{t+1} + \frac{a}{2}\sigma^2_{\varepsilon}\right)\left(\sigma_t v_t - \frac{a}{2}\sigma^2_t + \frac{a}{2}\sigma^2_{\varepsilon}\right)\right] \\ &= \frac{a^2}{4}(\alpha+\beta)\left(\mathrm{E}[\sigma^4_t] - \frac{\omega^2}{(1-\alpha-\beta)^2}\right) \\ &= \frac{a^2\omega^2\alpha^2(\alpha+\beta)}{2(1-(\alpha+\beta)^2-2\alpha^2)(1-\alpha-\beta)^2} \\ &\geqslant 0. \end{aligned} \qquad (23)$$

Serial correlation is absent when $\alpha = 0$ so the optimal bias does not depend on past (squared) innovations.

For multi-step-ahead predictions we can no longer obtain closed-form expressions since the innovations are not conditionally Gaussian. However, we are able to show that the volatility of volatility for this process is decreasing for reasonable parameter values, suggesting that the optimal forecast error variance can be decreasing in h for this process, using part (3) of Proposition 1. First, note that we can write the multi-step ahead conditional variance as

$$\begin{aligned} \sigma^2_{t+h,t} &= \omega\left(\frac{1-(\alpha+\beta)^h}{1-\alpha-\beta}\right) + \beta(\alpha+\beta)^{h-1}\sigma^2_{t,t-1} + \alpha(\alpha+\beta)^{h-1}y^2_t \\ &\equiv \omega_h + \beta_h\sigma^2_{t,t-1} + \alpha_h y^2_t. \end{aligned}$$

We then obtain

$$\mathrm{E}[\sigma^2_{t+h,t}] = \mathrm{E}[\omega_h + \beta_h\sigma^2_{t,t-1} + \alpha_h y^2_t] = \omega_h + \frac{\omega(\alpha_h+\beta_h)}{1-\alpha-\beta} = \frac{\omega}{1-\alpha-\beta},$$

$$\begin{aligned} \mathrm{E}[\sigma^4_{t+h,t}] &= \omega^2_h + \frac{2\omega\omega_h(\alpha_h+\beta_h)}{1-\alpha-\beta} + \frac{\omega^2(1+\alpha+\beta)(3\alpha^2_h+\beta^2_h+2\alpha_h\beta_h)}{(1-3\alpha^2-\beta^2-2\alpha\beta)(1-\alpha-\beta)} \\ &\to \frac{\omega^2}{(1-\alpha-\beta)^2} \quad \text{as } h \to \infty. \end{aligned}$$

In Fig. 6 we plot the volatility of volatility for the above process using $(\omega, \alpha, \beta) = (1, 0.05, 0.9)$, which are common parameter values for data such as daily stock returns.

Given the absence of a closed-form expression for the multi-step ahead conditional density in the GARCH case, we do not examine the autocorrelations for multi-step ahead forecasts or the forecast error variance as a function of forecast horizon for the GARCH example. Again this serves to show how difficult it is, in general, to characterize the moments of multi-step forecast errors under asymmetric loss and nonlinear DGPs.

5. Properties of a "generalized forecast error"

We demonstrated in the previous sections that all the properties of optimal forecasts established under MSE loss can be violated under asymmetric loss and a nonlinear DGP.

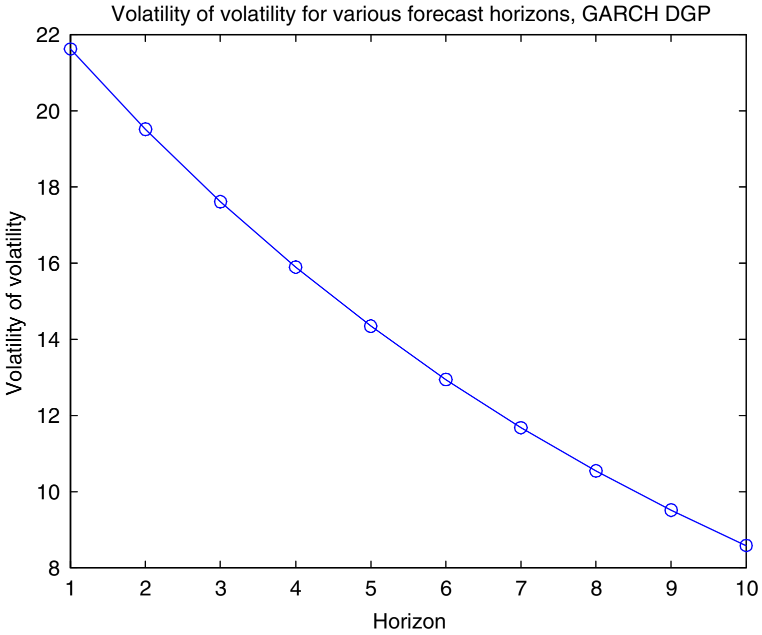

Fig. 6. The unconditional variance of h-step conditional variance for a GARCH(1,1) process with $(\omega, \alpha, \beta) = (1, 0.05, 0.9)$.

In this section we establish the properties that a simple transformation of the optimal forecast error—known as the generalized forecast error—must have.

Following Granger (1999) and Patton and Timmermann (2005), the generalized forecast error, $\psi^*_{t+h,t}$, is defined from the first-order condition of the loss function

$$\psi^*_{t+h,t} \equiv \frac{\partial L(Y_{t+h} - \hat{Y}^*_{t+h,t}; a)}{\partial \hat{Y}_{t+h,t}}, \tag{24}$$

which in the linex case simplifies to

$$\psi^*_{t+h,t} = a - a \exp\{a\sigma_{s_{t+h}} v_{t+h} - \log(\hat{\xi}'_{t|t}\mathbf{P}^h \boldsymbol{\varphi})\}. \tag{25}$$

The generalized forecast error is thus related to the "generalized residual" of Gourieroux et al. (1987) and Chesher and Irish (1987), and can be alternatively interpreted as the "score" at time $t+h$, evaluated using the forecaster's loss as opposed to the likelihood function. Clearly a correctly specified density model is a prerequisite for computing good (point) forecasts. The generalized forecast error exploits the property of an efficient forecast that, at the optimum, there can be no gain from changing the forecast by an amount that is a function of variables in the forecaster's current information set. For example, if bias is traded off against variance in the forecaster's loss function, at the optimum a reduction in one component due to a change in $\hat{Y}_{t+h,t}$ must exactly be offset by an increase in the other component. It is easy to establish that, although the forecast error, $e^*_{t+h,t}$, need not be mean zero, the generalized forecast

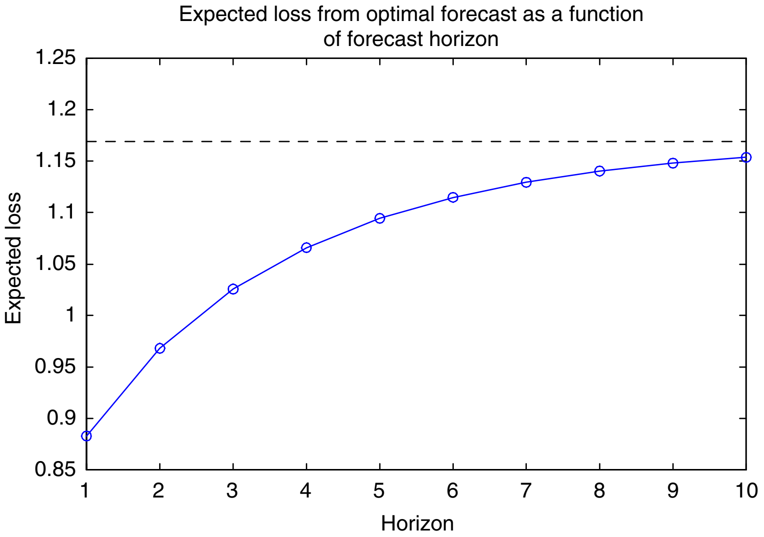

Fig. 7. Expected loss from the optimal forecast for various forecast horizons, two-state regime switching example.

error, $\psi^*_{t+h,t}$, has mean zero:

$$\begin{aligned}\mathrm{E}_t[\psi^*_{t+h,t}] &= a - a(\hat{\xi}'_{t|t}\mathbf{P}^h\boldsymbol{\varphi})^{-1}\mathrm{E}_t[\exp\{a\sigma_{s_{t+h}}v_{t+h}\}]\\ &= a - a(\hat{\xi}'_{t|t}\mathbf{P}^h\boldsymbol{\varphi})^{-1}\hat{\xi}'_{t|t}\mathbf{P}^h\exp\left\{\frac{a^2}{2}\boldsymbol{\sigma}^2\right\}\\ &= 0,\end{aligned}$$

and $\mathrm{E}[\psi^*_{t+h,t}] = 0$ by the law of iterated expectations. Thus the generalized forecast error has conditional and unconditional mean zero for all forecast horizons.

Turning to the second optimality property, while we saw that the variance of the optimal forecast error need not be non-decreasing with the forecast horizon, thus violating the second standard optimality property, the unconditional *expected loss* will always be non-decreasing in the forecast horizon. The expected loss corresponds exactly to the forecast error variance under MSE loss, but not generally.

Proposition 6. *Under linex loss* (8) *and the regime switching process* (11)–(12) *the unconditional expected loss is*

$$\mathrm{E}[L(Y_{t+h}, \hat{Y}^*_{t+h,t}; a)] = \bar{\xi}'\boldsymbol{\lambda}_h \to \log(\bar{\xi}'\boldsymbol{\varphi}) \quad \textit{as } h \to \infty.$$

For the numerical example used above, Fig. 7 shows the expected loss as a function of the forecast horizon. As expected it is a non-decreasing function of h.

For the third property, we saw that in a nonlinear framework, optimal h-step forecast errors need not follow an $MA(h-1)$ process. However, the generalized forecast error will display serial correlation of, at most, $(h-1)$th order:

Proposition 7. *The generalized forecast error from an optimal h-step forecast made at time t under the regime switching process* (11)–(12) *and assuming linex loss* (8) *has the following*

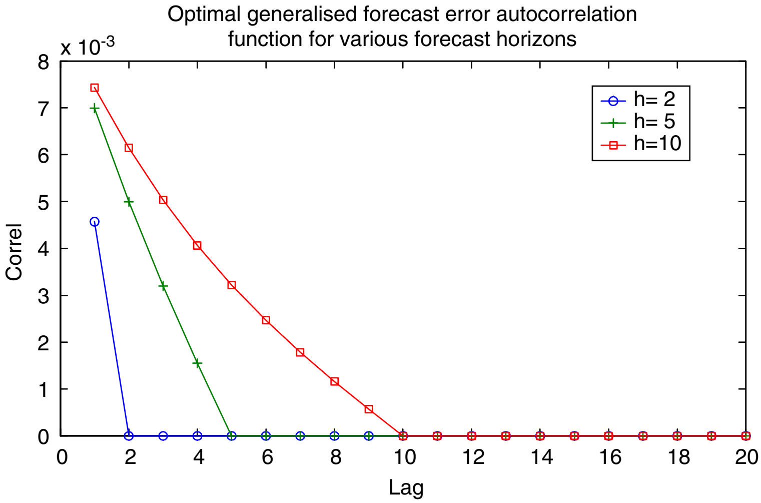

Fig. 8. Autocorrelation in the optimal generalized forecast error for various forecast horizons, two-state regime switching example.

autocovariance function:

$$Cov[\psi^*_{t+h,t}, \psi^*_{t+h-j,t-j}] = \begin{cases} -a^2 + a^2\sum_{s_t=1}^{k}\bar{\xi}_{(s_t)}(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi})^{-2}(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi}^4) & j=0, \\ -a^2 + a^2\sum_{s_{t-j}=1}^{k}\bar{\xi}_{(s_{t-j})}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi})^{-1} & \\ \quad \times\sum_{s_t=1}^{k}\xi_{(s_t|s_{t-j})}(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi})^{-1}\cdot(\boldsymbol{\varphi}'\odot(\boldsymbol{\iota}'_{s_t}\mathbf{P}^{h-j}))\mathbf{P}^j\boldsymbol{\varphi} & 0<j<h, \\ 0 & j\geqslant h, \end{cases}$$

where $\boldsymbol{\varphi}^4 \equiv \exp\{2a^2\boldsymbol{\sigma}^2\}$.

Using the numerical example above, Fig. 8 presents the autocorrelation function for the optimal generalized forecast error. As implied by Proposition 6, the autocorrelations are non-zero for $j<h$ and equal zero when $j\geqslant h$.

An alternative approach to evaluate the optimality of a forecast that is often used is based on the so-called probability integral transform proposed, in this context, by Rosenblatt (1952). This is simply the one-step forecast *cdf* associated with a given model evaluated at the observed data point. If the model is correctly specified, these probability integral transforms should constitute a sequence of i.i.d. uniformly distributed random variables. This approach requires, however, that the density model used by the forecaster is known in order to compute the probability integrals. It is not applicable in the common situation where the forecast evaluator does not know the forecaster's model and only a sequence of point forecasts generated under some loss function is observed.

6. Effect of parameter estimation error

So far we have established our theoretical results ignoring parameter estimation error. However, in practice this source of error can significantly influence optimality properties. Hoque et al. (1988), for example, show that for small sample sizes the MSFE is a decreasing function of h when the DGP is a mean zero AR(1) process with a small autoregressive coefficient (less than 0.2 in absolute value in their example), and the unknown coefficient is estimated using least squares. Further, although least squares methods provide the best linear unbiased estimator, for forecasting purposes least squares estimators are in fact inadmissible and dominated by shrinkage estimators and predictors that are biased though consistent. Requiring each of the elements of $\hat{\boldsymbol{\beta}}$ to be unbiased may be undesirable if the objective is to minimize the MSE of the estimator of $\boldsymbol{\beta}$. By dispensing with this requirement, shrinkage estimators can reduce the MSE.

More specifically, consider a linear forecasting model based on the n-vector of predictor variables, $\mathbf{x}_t = (x_{1t}, \ldots, x_{nt})$,

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \tag{26}$$

where $\mathbf{Y} = (Y_1 \cdots Y_T)'$, $\mathbf{X} = (\mathbf{x}'_0 \cdots \mathbf{x}'_{T-1})'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1 \cdots \varepsilon_T)'$. In the simple case where $\mathbf{X}'\mathbf{X} = \mathbf{I}_n$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and letting $\hat{\boldsymbol{\beta}}_{\mathrm{ls}}$ be the least squares estimator, the James and Stein (1961) estimator, $\hat{\boldsymbol{\beta}}_{\mathrm{JS}}$, that shrinks the least squares estimator towards a vector of zeros is

$$\hat{\boldsymbol{\beta}}_{\mathrm{JS}} = \left(1 - \frac{a}{\hat{\boldsymbol{\beta}}'_{\mathrm{ls}}\hat{\boldsymbol{\beta}}_{\mathrm{ls}}}\right)\hat{\boldsymbol{\beta}}_{\mathrm{ls}},$$

for some scalar, a. Under squared error loss the risk $R(\hat{\boldsymbol{\beta}}_{\mathrm{JS}}, \boldsymbol{\beta}) = \mathrm{E}[(\hat{\boldsymbol{\beta}}_{\mathrm{JS}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{\mathrm{JS}} - \boldsymbol{\beta})]$ associated with this estimator is (c.f. Judge et al., 1985)

$$R(\hat{\boldsymbol{\beta}}_{\mathrm{JS}}, \boldsymbol{\beta}) = n - a[2(n-2) - a]\mathrm{E}[1/\chi^2_{n,\lambda}],$$

where $\chi^2_{n,\lambda}$ is a non-central chi-square variable with degree of freedom parameter, n, and non-centrality parameter $\lambda = \boldsymbol{\beta}'\boldsymbol{\beta}/2$. The bias of this estimator, $\boldsymbol{\beta} - \mathrm{E}[\hat{\boldsymbol{\beta}}_{\mathrm{JS}}]$, equals $a\mathrm{E}[1/\chi^2_{n,\lambda}]\boldsymbol{\beta}$. Provided that we choose $0 \leqslant a \leqslant 2(n-2)$, it follows that $R(\hat{\boldsymbol{\beta}}_{\mathrm{JS}}, \boldsymbol{\beta}) < R(\hat{\boldsymbol{\beta}}_{\mathrm{ls}}, \boldsymbol{\beta})$ for all $\boldsymbol{\beta}$ so the least-squares estimator is inadmissible. Furthermore, the James–Stein estimator with the smallest risk sets $a = n-2$, i.e.[10]

$$\hat{\boldsymbol{\beta}}_{\mathrm{JS}} = \left(1 - \frac{n-2}{\hat{\boldsymbol{\beta}}'_{\mathrm{ls}}\hat{\boldsymbol{\beta}}_{\mathrm{ls}}}\right)\hat{\boldsymbol{\beta}}_{\mathrm{ls}}.$$

It follows from these results that biased forecasts can be optimal even under MSE loss, once parameter estimation error is accounted for. Empirical evidence seems to confirm this: Zellner and Hong (1989) find that shrinkage helps in forecasting growth rates of output in a large international data set, while Zellner and Chen (2001) find that a prior that shrinks disaggregate (sector) parameters in the direction of a common mean improves forecasts of US GDP growth.

[10] In the more general case where $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}$ and σ^2 is unknown, the optimal (feasible) James–Stein estimator takes the form

$$\hat{\boldsymbol{\beta}}^*_{\mathrm{JS}} = (1 - \frac{(T-n)(n-2)\hat{\sigma}^2}{(T-n+2)(\hat{\boldsymbol{\beta}}'_{\mathrm{ls}}\hat{\boldsymbol{\beta}}_{\mathrm{ls}})})\hat{\boldsymbol{\beta}}_{\mathrm{ls}},$$

where $\hat{\sigma}^2$ is the usual variance estimator. This estimator has risk:

$$R(\hat{\boldsymbol{\beta}}_{\mathrm{JS}}, \boldsymbol{\beta}) = n - \frac{(n-2)^2(T-n)}{T-n+2}\mathrm{E}[1/\chi^2_{n,\lambda}].$$

Outside the framework that assumes a linear forecasting model and a quadratic loss function, parameter estimation uncertainty is best dealt with in a Bayesian context. Under the Bayesian approach the evaluation of a set of predictions is no different from evaluation of a set of unknown parameters. If the data is taken as given (non-random), expected loss can be used to evaluate predictions, while the risk function and average or Bayes risk can be used to this end when the data is viewed as random. Moreover, Bayesian predictions that minimize average or Bayes risk have the desirable property that they are admissible so that no other predictor has lower average or Bayes risk.

More specifically, let $\mathbf{X}^t = (\mathbf{x}_0' \cdots \mathbf{x}_t')'$ be the data up to time t and let $\mathcal{Y}$ be the set of possible outcomes of Y, while $\boldsymbol{\theta}$ is again the parameters of the underlying model assumed to lie in some space Θ. The Bayesian approach chooses a forecast, $\tilde{Y}_{t+h,t}$, that minimizes expected loss given the available data, $\mathbf{X}^t$:

$$\tilde{Y}^*_{t+h,t} \equiv \underset{\tilde{Y}_{t+h,t}}{\arg\min} \int_{\mathcal{Y}} L(y_{t+h}, \tilde{Y}_{t+h,t}) p(y_{t+h}|\mathbf{X}^t)\, \mathrm{d}y_{t+h}, \tag{27}$$

where the predictive density is given by

$$p(y_{t+h}|\mathbf{X}_t) = \int_{\Theta} p(\boldsymbol{\theta}|\mathbf{X}^t) p(y_{t+h}|\mathbf{X}^t, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}, \tag{28}$$

and the posterior parameter distribution is proportional to the prior of the parameters and the conditional density of the data: $p(\boldsymbol{\theta}|\mathbf{X}^t) \propto p(\boldsymbol{\theta})p(\mathbf{X}^t|\boldsymbol{\theta})$. Under quadratic loss the optimal forecast is simply the conditional mean, $\tilde{Y}^*_{t+h,t} = \int_{\mathcal{Y}} yp(y|\mathbf{X}^t)\, \mathrm{d}y$. However, under asymmetric loss the forecast will generally not be easy to characterize in closed form. One exception pointed out by Zellner (1986) is for linex loss and a Gaussian density, $p(y_{t+h}|\mathbf{X}^t)$, in which case

$$\tilde{Y}^*_{t+h,t} = \mathrm{E}\left[Y_{t+h}|\mathbf{X}^t\right] - \frac{a}{2}\, Var(Y_{t+h}|\mathbf{X}^t). \tag{29}$$

Again it is clear that, in the presence of parameter estimation error, the optimal forecast is biased under asymmetric loss. Serial correlation of the optimal forecast error will also continue to hold even under MSE loss, c.f. Magnus and Pesaran (1989, 1991). We are not aware of any demonstrations of how it affects the variance of multi-period forecast errors, however, where the bias depends in a complicated manner on the persistence of the true DGP.

7. Empirical illustrations

In this section we present two illustrations of the above results for financial and economic data. The purpose of these empirical illustrations is to show that the simple model employed in the paper is a reasonable case to consider. Further, the results emphasize that using the standard properties of optimal forecast errors (unbiasedness, increasing variance, restricted serial correlation) in the presence of asymmetric loss and time-varying volatility could lead to the rejection of a perfectly optimal forecast.

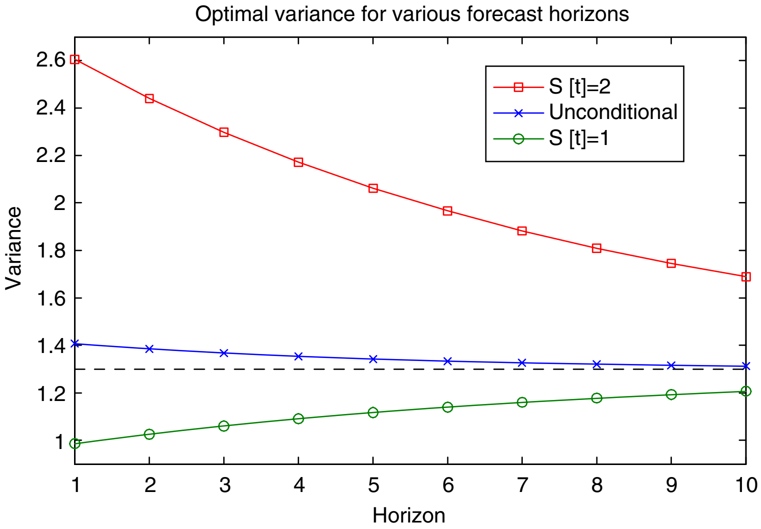

Fig. 9. Variance of the optimal h-step forecast error for various forecast horizons, Exxon stock returns, January 1970–December 2003.

7.1. Forecasts of stock returns

Using maximum likelihood methods we estimated a two-state regime switching model, with unobservable state variable S_t:

$$Y_t = \mu + \sigma_{s_t}\varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0,1), \tag{30}$$

on weekly returns for Exxon over the period January 1970 to December 2003, and obtained the following parameter estimates:

$$\hat{\mu} = 0.1323$$

$$\hat{\boldsymbol{\sigma}} = [0.9698, 1.6711]',$$

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.9756 & 0.0244 \\ 0.1014 & 0.8986 \end{bmatrix}.$$

Thus the annualized return on Exxon over this period was 6.9%. The steady state probabilities implied by the estimated transition matrix $\widehat{\bar{\xi}} = [0.8061, 0.1939]'$ indicate that returns were drawn from the low volatility regime four times as often as from the high volatility regime.

If we assume a linex loss function for the forecast user with $a = 1$, then Proposition 2 shows that the unconditional mean of the optimal Exxon return forecast errors is -0.7292. Given the asymmetry of the loss function it is not surprising that it is optimal to over-predict, leading to forecast errors that are negative on average.

Using the expression given in Proposition 3 we obtain the unconditional variance of the optimal forecast errors for different horizons, and plot these in Fig. 9 which is the empirical counterpart to Fig. 3. We again see, under linex loss and a regime switching model for Exxon stock returns, that the unconditional variance is a decreasing

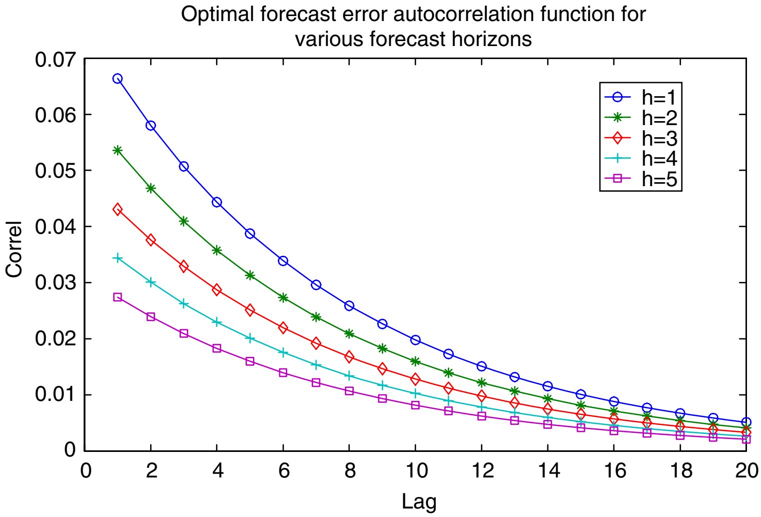

Fig. 10. Autocorrelation in the optimal h-step forecast error for various forecast horizons, Exxon stock returns, January 1970 to December 2003.

function of the forecast horizon. The MSE of the optimal forecast errors (not shown) is similarly monotonically downward sloping. Taking the standard properties of optimal forecasts as a guide, this figure would suggest that it is about 7% *easier* (variance of 1.31 versus 1.41) to predict the one-week return on Exxon in the period between nine and ten weeks from today, than the return on Exxon over the coming week. As we discussed above, this interpretation is misleading when the forecast user's loss function is not the MSE loss function. If we instead use expected loss to measure forecast difficulty, we find that it is about 8% *harder* (expected loss of 0.72 versus 0.67) to forecast the 10-step ahead one-week return than it is to forecast the return over the coming week.

Using Proposition 5 we obtain the autocorrelation function of the optimal forecast errors in this illustration, and plot this in Fig. 10 for $h = 1, \ldots, 5$. The one-step ahead weekly return forecast, for example, has first-order serial correlation of almost 0.07, and even at the 20th lag the serial correlation is still greater than zero. This implies that the forecast error is predictable using lagged forecast errors, albeit only weakly in this case. It does *not* imply that the forecast is suboptimal, however: the generalized forecast error exhibits the serial correlation properties presented in Proposition 7.

7.2. Forecasts of output growth

We next show the effect of asymmetric loss in a two-state example that accounts both for parameter estimation errors and an unobservable state, S_t. Given the strong evidence of regime-dependence in the mean of output growth (c.f. Hamilton, 1989), we extend (30) as follows:

$$Y_t = \mu_{s_t} + \sigma_{s_t}\varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0,1). \tag{31}$$

When both the underlying state and the parameters $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_{11}, p_{22})$ are unknown, a Gibbs sampling approach can be used to compute the predictive density, c.f. Albert and Chib (1993) and Kim and Nelson (1999, Chapter 9). The first step generates a vector of states, $\tilde{\mathbf{S}}_T = (S_1, \ldots, S_T)'$, from the conditional density $f(\tilde{\mathbf{S}}_T|\boldsymbol{\theta}, \tilde{\mathbf{y}}_T)$, where $\tilde{\mathbf{y}}_T = (y_1, y_2, \ldots, y_T)'$. The second step generates $(\mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$ from $f(\mu_0, \mu_1, \sigma_0^2, \sigma_1^2|\tilde{\mathbf{y}}_T, \tilde{\mathbf{S}}_T)$. Finally, the third step generates the state transition parameters, p_{11}, p_{22} from $f(p_{11}, p_{22}|\tilde{\mathbf{S}}_T)$ using that, conditional on $\tilde{\mathbf{S}}_T$, these state transitions are independent of $\tilde{\mathbf{y}}_T$.[11]

Following Chib (1996) we assume that the priors of the state transition probabilities follow Dirichlet distributions. Multiplying by the likelihood function for $\mathbf{P}$ conditional on a simulated vector of states this means that the posterior is also Dirichlet. We set the prior mean of each element equal to 0.5 and the standard deviation equal to 0.25. Assuming a Gaussian prior for μ_i conditional on σ_i^2, i.e., $\mu_i|\sigma_i^2 \sim \mathrm{N}(b_0, B_0)$, the posteriors for the mean parameters are (for $i = 1, 2$)

$$\begin{aligned}
&\mu_i|\sigma_i^2, \tilde{\mathbf{S}}_T, \tilde{\mathbf{y}}_T \sim \mathrm{N}(b_{1i}, B_{1i}),\\
&\qquad b_{1i} = (B_0^{-1} + \sigma_i^{-2}\boldsymbol{\iota}_T'\boldsymbol{\iota}_T)^{-1}(B_0^{-1}b_0 + \sigma_i^{-2}\boldsymbol{\iota}_T'\mathbf{Y}_T),\\
&\qquad B_{1i} = (B_0^{-1} + \sigma_i^{-2}\boldsymbol{\iota}_T'\boldsymbol{\iota}_T)^{-1}.
\end{aligned}$$

In both cases we use basically uninformative priors, i.e. $b_0 = 0, B_0 = 1000$. With μ_i in place we assume that the priors for the variance terms are inverse Gaussian, $\sigma_i^2|\mu_i \sim IG(v_0/2, \delta_0/2)$, where $v_0 = 2$ and $\delta_0 = 0.001$ so we again have basically uninformative priors. Conditional on $\tilde{\mathbf{S}}_T, \tilde{\mathbf{y}}_T$ this yields an inverse Gaussian posterior, $\sigma_i^2|\mu_i, \tilde{\mathbf{S}}_T, \tilde{\mathbf{y}}_T \sim IG(v_1/2, \delta_{1i}/2)$, where $v_1 = v_0 + T$ and $\delta_1 = \delta_0 + \sum_{t=1}^{T}(y_t - \mu_i)^2\mathbf{1}(s_t = i)$.

In our empirical illustration we use quarterly data on US GDP growth from 1952Q1 to 1994Q3, and save the last 20 observations of this data set to compare against our predictions. This data was also used by Kim and Nelson (1999). We compute the one through 20-step ahead forecasts starting from the last date in our estimation sample, 1989Q3. Using 7500 draws from the Gibbs sampler, the posterior means of the growth rate parameters in the two states were 1.09 and 0.05, respectively, with variance parameters equal to 0.49 and 1.00, and transition probability parameters centered at 0.74 and 0.62 (each associated with a standard deviation of 0.12).

In Fig. 11 we plot the optimal forecasts under MSE and linex loss (with $a = 1$) against the realizations of GDP growth. These forecasts are computed using Eq. (27). Note that actual GDP growth is greater than the linex forecast of GDP growth for only three out of 20 cases; this reflects the greater penalty this loss function applies to under-predictions relative to over-predictions. In contrast, the MSE forecasts are exceeded by the actual GDP growth figures in nine out of 20 cases. Hence the optimal forecast under linex loss is clearly above that under MSE loss even after accounting for parameter estimation error.[12] Fig. 12 shows how the one-step predictive density from model (28) differ from the normal distribution. Note also the pronounced negative skew in the regime switching density forecast.

[11] We are grateful to Davide Pettenuzzo for providing us with his Gibbs sampling code for this application.

[12] In related work based on the lin–lin loss function, Whiteman (1996) finds that asymmetric loss can significantly change the mean of the optimal forecast. His application is to budget surpluses for Iowa.

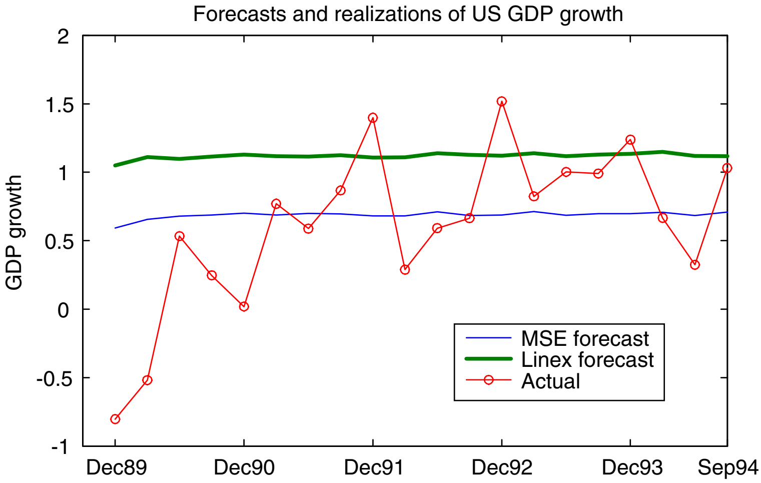

Fig. 11. Optimal one- to 20-step ahead forecasts of US GDP growth, as at 1989Q3, under MSE and Linex loss.

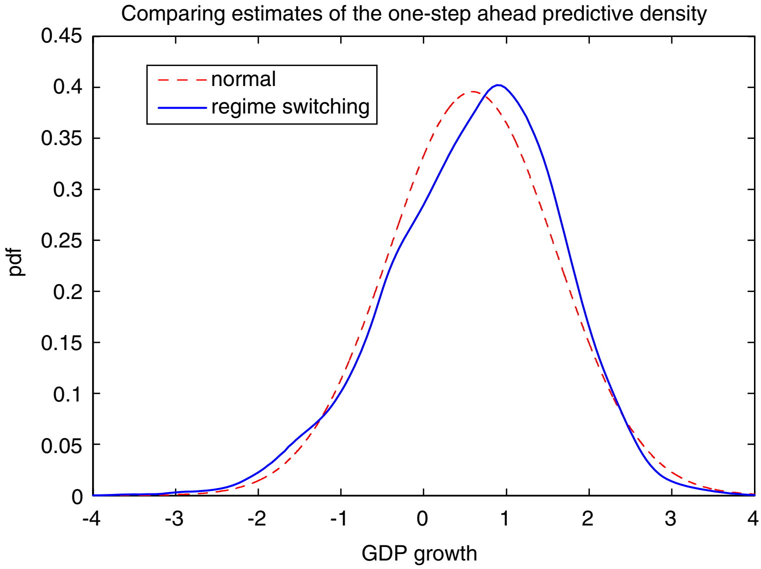

Fig. 12. Comparing estimates of the one-step ahead predictive density of GDP growth.

8. Conclusion

This paper demonstrated that the properties of optimal forecasts that are almost always tested in the empirical literature hold only under very restrictive assumptions, and are not generally robust to even slight departures from these assumptions. We showed analytically how they are violated under more general assumptions about the loss function, extending the seminal work of Granger (1969) and Christoffersen and Diebold (1997). The properties that optimal forecasts must possess were generalized to consider situations where the loss function may be asymmetric and the DGP may be nonlinear but strictly stationary. Illustrations using data on weekly stock returns and quarterly output growth confirmed that our findings have much empirical relevance.

A number of implications follow from this paper. Most importantly, our results suggest the need to develop new and more general methods for forecast evaluation that are robust to deviations from MSE loss. Most economic time series have some dynamics in the conditional variance. This means that even small deviations from MSE loss will overturn the standard properties that an optimal forecast must have. One approach—proposed by Elliott et al. (2005)—is to assume that the loss function belongs to a family of parsimoniously parameterized functions that nests MSE loss as a special case. This approach can be used even with the small sample sizes typically obtainable with forecasting data. Furthermore, the parameters of the loss function can be estimated by GMM and forecast optimality can be tested by means of a standard *J*-test when the model is overidentified. Another approach, suggested by Patton and Timmermann (2005), is to adopt robust test procedures that are valid for broad classes of loss functions and DGPs.

Acknowledgments

The authors would like to thank the editor, Arnold Zellner, an associate editor, two anonymous referees, Jeremy Berkowitz, Sean Campbell, Carlos Capistran, Peter Christoffersen, Frank Diebold, Graham Elliott, Raffaella Giacomini, Clive Granger, Peter Hansen, Oliver Linton, Mark Machina, Francisco Penaranda, Davide Pettenuzzo, Kevin Sheppard, Hal White, Stanley Zin and seminar participants at Brown, CORE, LSE, McGill, UC Riverside, Stanford, Warwick Business School and the 2002 EC^2 Conference for useful comments. All remaining deficiencies are the responsibility of the authors.

Appendix A

Proof of Proposition 1. Part 1: We know from Christoffersen and Diebold (1997) that $\hat{Y}^*_{t+h,t} = \mu_{t+h,t} + \alpha_{t+h,t}$, where $\alpha_{t+h,t}$ depends on the loss, $L(.)$, $F_{t+h,t}$ and $\sigma_{t+h,t}$ but not on $\mu_{t+h,t}$. So without loss of generality, consider a process with conditional mean zero:

$$Y_{t+h} = \sigma_{t+h,t}\varepsilon_{t+h}.$$

Proving that $\hat{Y}^*_{t+h,t}$ is biased then amounts to showing that $\hat{Y}^*_{t+h,t} \neq 0$. Under assumptions L1 and L2 we have $\tilde{L}(e) = \tilde{L}(-e)$, and $\zeta(e) \geqslant \zeta(-e)$ for all $e > 0$ and so the optimal forecast satisfies $\hat{Y}^*_{t+h,t} > 0$. To see this, define

$$\begin{aligned}\Gamma(\hat{Y}_{t+h,t}, \sigma_{t+h,t}) &\equiv \frac{\partial}{\partial \hat{Y}} \mathrm{E}_t[L(\sigma_{t+h,t}\varepsilon_{t+h} - \hat{Y}_{t+h,t})] \\ &= -\int \tilde{L}'(\sigma_{t+h,t}\varepsilon_{t+h} - \hat{Y}_{t+h,t}) f(\varepsilon_{t+h})\,\mathrm{d}\varepsilon_{t+h} \\ &\quad - \int_{\hat{y}_{t+h,t}/\sigma_{t+h,t}}^{\infty} \zeta'(\sigma_{t+h,t}\varepsilon_{t+h} - \hat{Y}_{t+h,t}) f(\varepsilon_{t+h})\,\mathrm{d}\varepsilon_{t+h}, \end{aligned} \tag{A.1}$$

assuming that we may interchange expectation and differentiation operators. $\Gamma(\hat{Y}, \sigma)$, evaluated at $\hat{Y} = 0$, is negative since the first term in (A.1) equals zero (if $\tilde{L}$ were the loss function then zero would be the optimal forecast under the given assumptions) while the second term is negative (the integral term is positive since $\zeta' \geqslant 0$ and $\zeta' > 0$ on a set with measure greater than zero). In the following we drop time-subscripts where these are not needed. Using the convexity of $\tilde{L}$ and ζ we can show that $\partial \Gamma / \partial \hat{Y} > 0$:

$$\begin{aligned}\frac{\partial \Gamma(\hat{Y}, \sigma)}{\partial \hat{Y}} &= \frac{\partial}{\partial \hat{Y}} \left(-\int \tilde{L}'(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon - \int_{\hat{y}/\sigma}^{\infty} \zeta'(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon \right) \\ &= \int \tilde{L}''(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon + \int_{\hat{y}/\sigma}^{\infty} \zeta''(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon + \zeta'(0)\frac{1}{\sigma} \\ &> 0. \end{aligned} \tag{A.2}$$

The existence of an interior optimum (by the convexity of $L(.)$, c.f. Lehmann and Casella, 1998) then implies that $\hat{Y}^*_{t+h,t} > 0$, proving that the optimal forecast is biased.

Part 2: To prove that the optimal forecast error is not in general a martingale difference sequence with respect to $\mathscr{F}_t$ note that $e^*_{t+h,t} = \sigma_{t+h,t}\varepsilon_{t+h} - \hat{Y}^*_{t+h,t}$, where $\hat{Y}^*_{t+h,t}$ is that value of $\hat{Y}_{t+h,t}$ for which (A.1) equals zero. Provided that $\partial \hat{Y}^*_{t+h,t} / \partial \sigma_{t+h,t} \neq 0$, time-variations in $\sigma_{t+h,t}$ will translate into predictable time-variations in the forecast error by means of the same variables in $\mathscr{F}_t$ that forecast $\sigma_{t+h,t}$. To show that $\partial \hat{Y}^*_{t+h,t} / \partial \sigma_{t+h,t} \neq 0$ we use the implicit function theorem: $\partial \hat{Y}^*(\sigma)/\partial \sigma = -(\partial \Gamma / \partial \sigma)/(\partial \Gamma / \partial \hat{Y})$. The denominator equals:

$$\begin{aligned}\frac{\partial \Gamma(\hat{Y}, \sigma)}{\partial \hat{Y}} &= \frac{\partial}{\partial \hat{Y}} \left(-\int \tilde{L}'(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon - \int_{\hat{y}/\sigma}^{\infty} \zeta'(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon \right) \\ &= \int \tilde{L}''(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon + \int_{\hat{y}/\sigma}^{\infty} \zeta''(\sigma\varepsilon - \hat{Y}) f(\varepsilon)\, \mathrm{d}\varepsilon + \zeta'(0)\frac{1}{\sigma}. \end{aligned} \tag{A.2}$$

Since both $\tilde{L}$ and ζ are convex, $\partial \Gamma / \partial \hat{Y}$ is guaranteed to be positive. (This further implies that $\hat{Y}^* > 0$ as $\Gamma(\hat{Y}_{t+h,t} = 0, \sigma_{t+h,t}) < 0$.) To derive $\partial \hat{Y}^*(\sigma)/\partial \sigma$ note that

$$\frac{\partial \Gamma(\hat{Y}^*(\sigma), \sigma)}{\partial \sigma} = -\int \tilde{L}''(\sigma\varepsilon - \hat{Y}^*)\varepsilon f(\varepsilon)\, \mathrm{d}\varepsilon - \int_{\hat{y}/\sigma}^{\infty} \zeta''(\sigma\varepsilon - \hat{Y}^*)\varepsilon f(\varepsilon)\, \mathrm{d}\varepsilon - \zeta'(0)\frac{1}{\sigma^2}\hat{Y}^*.$$

The last term in this equation is negative under assumption L1. The terms in the above expression will only cancel out in very special cases and so, generically, $\partial \Gamma(\hat{Y}^*(\sigma), \sigma)/\partial \sigma \neq 0$. For example, if $\tilde{L}(e) = ae^2$ the first term equals zero and the second term is negative, so $\partial \Gamma / \partial \sigma < 0$ which implies $\partial \hat{Y}^* / \partial \sigma > 0$. If $\partial \Gamma / \partial \sigma \neq 0$ then serial dependencies in $\sigma_{t+h,t}$ will generally translate into serial dependencies in the optimal bias, $\hat{Y}^*_{t+h,t}$, and hence in the forecast error. To see this, let σ_1 and σ_2 be two distinct possible values for $\sigma_{t+h,t}$, which is time-varying by assumption D2. If we know $\partial \hat{Y}^* / \partial \sigma > 0$, then $\hat{Y}^*(\sigma_1) \neq \hat{Y}^*(\sigma_2)$, and so we have $\mathrm{E}[e^*_{t+h,t} | \sigma_{t+h,t} = \sigma_1] = \mathrm{E}[\sigma_{t+h,t}\varepsilon_{t+h} - \hat{Y}^*_{t+h,t} | \sigma_{t+h,t} = \sigma_1] = \mathrm{E}[\sigma_{t+h,t}\mathrm{E}_t[\varepsilon_{t+h}] - \hat{Y}^*(\sigma_{t+h,t}) | \sigma_{t+h,t} = \sigma_1] = -\hat{Y}^*(\sigma_1) \neq -\hat{Y}^*(\sigma_2) = \mathrm{E}[e^*_{t+h,t} | \sigma_{t+h,t} = \sigma_2]$. $\mathrm{E}[e^*_{t+h,t}]$ can equal at most one of $\mathrm{E}[e^*_{t+h,t} | \sigma_{t+h,t} = \sigma_1]$ or $\mathrm{E}[e^*_{t+h,t} | \sigma_{t+h,t} = \sigma_2]$, which proves the claim.

Part 3: To understand the factors bringing about the possibility of a declining variance of the forecast error in h, consider first the case of MSE loss. In that case, $\hat{Y}^*_{t+h,t} = \mu_{t+h,t}$, and so $\mathrm{E}_t[e^*_{t+h,t}] = 0$. Using the law of iterated expectations and assuming the process is covariance stationary, the optimality of $\hat{Y}^*_{t+h,t}$ implies:

$$\mathrm{E}_t[e^{*2}_{t+h,t}] \leqslant \mathrm{E}_t[e^{*2}_{t+h,t-j}] \quad \forall j \geqslant 0,$$

so

$$\mathrm{E}[e^{*2}_{t+h,t}] \leqslant \mathrm{E}[e^{*2}_{t+h,t-j}],$$

and

$$V[e^{*}_{t+h,t}] \leqslant V[e^{*}_{t+h,t-j}] = V[e^{*}_{t+h+j,t}].$$

Thus the unconditional variance of the optimum forecast error is weakly increasing in the horizon. Now consider the general case of asymmetric loss: note that $e^*_{t+h,t} = \sigma_{t+h,t}\varepsilon_{t+h} - \hat{Y}^*_{t+h,t}$, so $V(e^*_{t+h,t}) = \mathrm{E}[V_t(e^*_{t+h,t})] + V(\mathrm{E}_t[e^*_{t+h,t}]) = \mathrm{E}[\sigma^2_{t+h,t}] + V(\hat{Y}^*_{t+h,t})$. Comparing this across different values of h, under covariance stationarity the first term is constant, so we focus on the second term, $V(\hat{Y}^*_{t+h,t})$. We establish the result through a Taylor series expansion under the assumption that Y_{t+h} is covariance stationary and $\varepsilon_{t+h}|\mathscr{F}_t \sim F(0,1,\kappa_h)$, where κ_h is independent of $\mathscr{F}_t$ and non-decreasing in h. We demonstrate that under sign restrictions on the derivatives, $\partial\hat{Y}^*/\partial\sigma > 0$ and $\partial^2\hat{Y}^*/(\partial\sigma\partial\kappa) \leqslant 0$, $V(e^*_{t+2,t}) < V(e^*_{t+1,t})$. Note that

$$\begin{aligned}\hat{Y}^*_{t+h,t} &= \hat{Y}^*(\sigma^2_{t+h,t}, \kappa_h)\\ &\approx \hat{Y}^*(\breve{\sigma}^2_h, \kappa_h) + \hat{Y}'^*_{\sigma}(\breve{\sigma}^2_h, \kappa_h)(\sigma^2_{t+h,t} - \breve{\sigma}^2_h)\\ &= \bar{Y}^*_h + \hat{Y}'^*_{\sigma}(\breve{\sigma}^2_h, \kappa_h)(\sigma^2_{t+h,t} - \breve{\sigma}^2_h),\end{aligned}$$

where $\breve{\sigma}^2_h$ is selected to satisfy $\hat{Y}^*(\breve{\sigma}^2_h, \kappa_h) = \bar{Y}^*_h \equiv \mathrm{E}[\hat{Y}^*_{t+h,t}]$. Then $\mathrm{E}[\hat{Y}^*_{t+h,t}] \approx \mathrm{E}[\bar{Y}^*_h + \hat{Y}^{*\prime}_{\sigma}(\breve{\sigma}^2_h, \kappa_h)(\sigma^2_{t+h,t} - \breve{\sigma}^2_h)] = \bar{Y}^*_h + \hat{Y}^{*\prime}_{\sigma}(\breve{\sigma}^2_h, \kappa_h)(\mathrm{E}[\sigma^2_{t+h,t}] - \breve{\sigma}^2_h) \Rightarrow \breve{\sigma}^2_h \approx \mathrm{E}[\sigma^2_{t+h,t}] \equiv \sigma^2_y\ \forall h$. So $\hat{Y}^*_{t+h,t} \approx \bar{Y}^*_h + \hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_h)(\sigma^2_{t+h,t} - \sigma^2_y)$ for all h. Then

$$V[\hat{Y}^*_{t+1,t}] \approx V[\bar{Y}^*_1 + \hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_1)(\sigma^2_{t+1,t} - \sigma^2_y)] = \hat{Y}_{\sigma}{}^{*\prime}(\sigma^2_y, \kappa_1)^2 V[\sigma^2_{t+1,t}]$$

Similarly,

$$V[\hat{Y}^*_{t+2,t}] \approx \hat{Y}_{\sigma}{}^{*\prime}(\sigma^2_y, \kappa_2)^2 V[\sigma^2_{t+2,t}].$$

So

$$\begin{aligned}V[\hat{Y}^*_{t+1,t}] - V[\hat{Y}^*_{t+2,t}] \approx\ & \hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_1)^2(V[\sigma^2_{t+1,t}] - V[\sigma^2_{t+2,t}])\\ &+ V[\sigma^2_{t+2,t}](\hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_1) + \hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_2))(\hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_1) - \hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_2))\\ &> 0.\end{aligned}$$

Here we used that $(V[\sigma^2_{t+1,t}] - V[\sigma^2_{t+2,t}]) > 0$ by $D3$, while $V[\sigma^2_{t+2,t}] \geqslant 0$ by assumption $D2$, $\hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_1)^2 > 0$ and $(\hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_1) - \hat{Y}^{*\prime}_{\sigma}(\sigma^2_y, \kappa_2)) \geqslant 0$ by the assumption that $\partial\hat{Y}^*/\partial\sigma > 0$ and $\partial^2\hat{Y}^*/\partial\sigma\partial\kappa < 0$ and the fact that $\kappa_1 \leqslant \kappa_2$. Thus the variance of $e^*_{t+1,t}$ is greater than that of $e^*_{t+2,t}$, completing the proof. □

Proof of Proposition 2. The h-step-ahead forecast error has a conditional expectation of

$$\mathrm{E}_t[e^*_{t+h,t}] = -\frac{1}{a}\log(\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h\boldsymbol{\varphi}),$$

which, since $\mathbf{P}$ is a probability matrix with one eigenvalue of unity, is different from zero even when $h \to \infty$. The unconditional expectation of the forecast error is

$$\begin{aligned}
\mathrm{E}[e^*_{t+h,t}] &= \mathrm{E}[\mathrm{E}_t[e^*_{t+h,t}]] \\
&= \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \mathrm{E}\left[-\frac{1}{a} \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) | S_t = s_t \right] \\
&= -\frac{1}{a} \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \log(\boldsymbol{\iota}'_{s_t,t} \mathbf{P}^h \boldsymbol{\varphi}) \\
&= -\frac{1}{a} \bar{\boldsymbol{\xi}}' \boldsymbol{\lambda}_h,
\end{aligned}$$

where $\boldsymbol{\lambda}_h = \log(\mathbf{P}^h \boldsymbol{\varphi})$ and $\boldsymbol{\iota}_{s_t}$ is a $k \times 1$ zero-one selection vector that is unity in the s_tth element and is zero otherwise.

The unconditional bias remains, in general, non-zero for all h. In the limit as $h \to \infty$,

$$\mathrm{E}[e^*_{t+h,t}] \to -\frac{1}{a} \bar{\boldsymbol{\xi}}' \log(\boldsymbol{\iota} \bar{\boldsymbol{\xi}}' \boldsymbol{\varphi}) = -\frac{1}{a} \bar{\boldsymbol{\xi}}' \boldsymbol{\iota} \log(\bar{\boldsymbol{\xi}}' \boldsymbol{\varphi}) = -\frac{1}{a} \log(\bar{\boldsymbol{\xi}}' \boldsymbol{\varphi}),$$

which is also, in general, non-zero. □

Proof of Proposition 3. From Proposition 2 we have

$$\begin{aligned}
Var(e^*_{t+h,t}) &= \mathrm{E}[e^{*2}_{t+h,t}] - \frac{1}{a^2} \boldsymbol{\lambda}'_h \bar{\boldsymbol{\xi}} \bar{\boldsymbol{\xi}}' \boldsymbol{\lambda}_h \\
&= \mathrm{E}\left[\left(\sigma_{s_{t+h}} v_{t+h} - \frac{1}{a} \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) \right)^2 \right] - \frac{1}{a^2} \boldsymbol{\lambda}'_h \bar{\boldsymbol{\xi}} \bar{\boldsymbol{\xi}}' \boldsymbol{\lambda}_h \\
&= \mathrm{E}[\sigma^2_{s_{t+h}} v^2_{t+h}] - \frac{2}{a} \mathrm{E}[\sigma_{s_{t+h}} v_{t+h} \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi})] + \frac{1}{a^2} \mathrm{E}[\log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi})^2] - \frac{1}{a^2} \boldsymbol{\lambda}'_h \bar{\boldsymbol{\xi}} \bar{\boldsymbol{\xi}}' \boldsymbol{\lambda}_h \\
&= \sum_{s_{t+h}=1}^{k} \bar{\xi}_{(s_{t+h})} \mathrm{E}[\sigma^2_{s_{t+h}} v^2_{t+h} | S_{t+h} = s_{t+h}] - \frac{1}{a^2} \boldsymbol{\lambda}'_h \bar{\boldsymbol{\xi}} \bar{\boldsymbol{\xi}}' \boldsymbol{\lambda}_h \\
&\quad + \frac{1}{a^2} \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \mathrm{E}[\log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) \cdot \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) | S_t = s_t] \\
&= \sum_{s_{t+h}=1}^{k} \bar{\xi}_{(s_{t+h})} \sigma^2_{s_{t+h}} - \frac{1}{a^2} \boldsymbol{\lambda}'_h \bar{\boldsymbol{\xi}} \bar{\boldsymbol{\xi}}' \boldsymbol{\lambda}_h + \frac{1}{a^2} \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \log(\boldsymbol{\varphi}' \mathbf{P}^{h\prime}) \boldsymbol{\iota}_{s_t} \boldsymbol{\iota}'_{s_t} \log(\mathbf{P}^h \boldsymbol{\varphi}) \\
&= \bar{\boldsymbol{\xi}}' \boldsymbol{\sigma}^2 + \frac{1}{a^2} \boldsymbol{\lambda}'_h \left(\sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \boldsymbol{\iota}_{s_t} \boldsymbol{\iota}'_{s_t} \right) \boldsymbol{\lambda}_h - \frac{1}{a^2} \boldsymbol{\lambda}'_h \bar{\boldsymbol{\xi}} \bar{\boldsymbol{\xi}}' \boldsymbol{\lambda}_h \\
&= \bar{\boldsymbol{\xi}}' \boldsymbol{\sigma}^2 + \frac{1}{a^2} \boldsymbol{\lambda}'_h ((\bar{\boldsymbol{\xi}} \boldsymbol{\iota}') \odot \mathbf{I} - \bar{\boldsymbol{\xi}} \bar{\boldsymbol{\xi}}') \boldsymbol{\lambda}_h.
\end{aligned}$$

Here $\bar{\xi}_{(i)}$ is the ith element of the vector $\bar{\boldsymbol{\xi}}$, the outer product $\boldsymbol{\iota}_{s_t} \boldsymbol{\iota}'_{s_t}$ is a $k \times k$ matrix of all zeros, except for the (s_t, s_t)th element, which equals one. To examine the variance of the optimal h-step ahead forecast as $h \to \infty$, note that

$$\lim_{h \to \infty} \boldsymbol{\lambda}_h = \boldsymbol{\iota} \log(\bar{\boldsymbol{\xi}}' \boldsymbol{\varphi}).$$

Furthermore, for any vector $\bar{\boldsymbol{\xi}}$ such that $\bar{\boldsymbol{\xi}}'\boldsymbol{\iota} = 1$,

$$\boldsymbol{\iota}'((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}') \odot \mathbf{I} - \bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}')\boldsymbol{\iota} = \boldsymbol{\iota}'((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}') \odot \mathbf{I})\boldsymbol{\iota} - \boldsymbol{\iota}'\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}'\boldsymbol{\iota} = \bar{\boldsymbol{\xi}}'\boldsymbol{\iota} - (\bar{\boldsymbol{\xi}}'\boldsymbol{\iota})'(\bar{\boldsymbol{\xi}}'\boldsymbol{\iota}) = 0.$$

So, as $h \to \infty$, the variance of the optimal h-step ahead forecast therefore converges to

$$\begin{aligned} Var[e^*_{t+h,t}] &\to \bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2 + \frac{1}{a^2}\log(\bar{\boldsymbol{\xi}}'\boldsymbol{\varphi})\boldsymbol{\iota}'((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}') \odot \mathbf{I} - \bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}')\boldsymbol{\iota}\log(\bar{\boldsymbol{\xi}}'\boldsymbol{\varphi}) \\ &= \bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2. \qquad \square \end{aligned}$$

Proof of Corollary 1. Follows directly from the proof of Proposition 3. □

Proof of Proposition 4. The conditional variance of the variable of interest is

$$\begin{aligned} V_t[Y_{t+h}] &= \sum_{s_{t+h}=1}^{k} \hat{\xi}_{(s_{t+h}|t)} V_t[\sigma_{s_{t+h}} v_{t+h} | S_{t+h} = s_{t+h}] \\ &= \sum_{s_{t+h}=1}^{k} \hat{\xi}_{(s_{t+h}|t)} \sigma^2_{s_{t+h}} = \hat{\boldsymbol{\xi}}'_{t|t} \boldsymbol{P}^h \boldsymbol{\sigma}^2, \end{aligned}$$

and so

$$\begin{aligned} \mathrm{E}[V_t[Y_{t+h}]] &= \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \mathrm{E}[\hat{\boldsymbol{\xi}}'_{t|t} \boldsymbol{P}^h \boldsymbol{\sigma}^2 | S_t = s_t] \\ &= \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \boldsymbol{\iota}'_{s_t} \boldsymbol{P}^h \boldsymbol{\sigma}^2 = \bar{\boldsymbol{\xi}}' \boldsymbol{P}^h \boldsymbol{\sigma}^2, \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}[V_t[Y_{t+h}]^2] &= \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \mathrm{E}[\hat{\boldsymbol{\xi}}'_{t|t} \boldsymbol{P}^h \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \boldsymbol{P}^{h\prime} \hat{\boldsymbol{\xi}}_{t|t} | S_t = s_t] \\ &= \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \boldsymbol{\iota}'_{s_t} \boldsymbol{P}^h \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \boldsymbol{P}^{h\prime} \boldsymbol{\iota}_{s_t} = \bar{\boldsymbol{\xi}}' (\boldsymbol{P}^h \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \boldsymbol{P}^{h\prime} \odot \boldsymbol{I}) \boldsymbol{\iota}, \end{aligned}$$

which yields the expression given in the proposition. As $h \to \infty$,

$$\begin{aligned} \bar{\boldsymbol{\xi}}' \boldsymbol{P}^h \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \boldsymbol{P}^{h\prime} \bar{\boldsymbol{\xi}} &\to \bar{\boldsymbol{\xi}}' \boldsymbol{\iota} \bar{\boldsymbol{\xi}}' \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \bar{\boldsymbol{\xi}} \boldsymbol{\iota}' \bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}' \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \bar{\boldsymbol{\xi}}, \\ \bar{\boldsymbol{\xi}}' (\boldsymbol{P}^h \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \boldsymbol{P}^{h\prime} \odot \boldsymbol{I}) \boldsymbol{\iota} &\to \bar{\boldsymbol{\xi}}' (\boldsymbol{\iota} \bar{\boldsymbol{\xi}}' \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \bar{\boldsymbol{\xi}} \boldsymbol{\iota}' \odot \mathbf{I}) \boldsymbol{\iota} = \bar{\boldsymbol{\xi}}' (\boldsymbol{\iota} \bar{\boldsymbol{\xi}}' \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \bar{\boldsymbol{\xi}}) = \bar{\boldsymbol{\xi}}' \boldsymbol{\sigma}^2 \boldsymbol{\sigma}^{2\prime} \bar{\boldsymbol{\xi}}, \end{aligned}$$

so $V[V_t[Y_{t+h}]] \to 0$. □

Proof of Proposition 5. The autocovariance function for an h-step forecast is

$$\begin{aligned} Cov[e^*_{t+h,t}, e^*_{t+h-j,t-j}] = \mathrm{E}[(\sigma_{s_{t+h-j}} \sigma_{s_{t+h}} v_{t+h-j} v_{t+h})] &- \frac{1}{a} \mathrm{E}[\sigma_{s_{t+h-j}} v_{t+h-j} \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi})] \\ &- \frac{1}{a} \mathrm{E}[\sigma_{s_{t+h}} v_{t+h} \log(\hat{\boldsymbol{\xi}}'_{t-j|t-j} \mathbf{P}^h \boldsymbol{\varphi})] \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{a^2}\,\mathrm{E}[\log(\hat{\boldsymbol{\xi}}_{t|t}'\mathbf{P}^h\boldsymbol{\varphi})\log(\hat{\boldsymbol{\xi}}_{t-j|t-j}'\mathbf{P}^h\boldsymbol{\varphi})]-\frac{1}{a^2}\,\boldsymbol{\lambda}_h'\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}'\boldsymbol{\lambda}_h\\
&=\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2\mathbf{1}_{\{j=0\}}-\frac{1}{a^2}\,\boldsymbol{\lambda}_h'\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}'\boldsymbol{\lambda}_h+\frac{1}{a^2}\sum_{s_{t-j}=1}^{k}\sum_{s_t=1}^{k}\bar{\xi}_{(s_{t-j})}\hat{\xi}_{(s_t|s_{t-j})}\\
&\quad\times\mathrm{E}[\log(\hat{\boldsymbol{\xi}}_{t|t}'\mathbf{P}^h\boldsymbol{\varphi})\log(\hat{\boldsymbol{\xi}}_{t-j|t-j}'\mathbf{P}^h\boldsymbol{\varphi})|S_{t-j}=s_{t-j},S_t=s_t]\\
&=\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2\mathbf{1}_{\{j=0\}}-\frac{1}{a^2}\,\boldsymbol{\lambda}_h'\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}'\boldsymbol{\lambda}_h\\
&\quad+\frac{1}{a^2}\sum_{s_{t-j}=1}^{k}\sum_{s_t=1}^{k}\bar{\xi}_{(s_{t-j})}\hat{\xi}_{(s_t|s_{t-j})}\log(\boldsymbol{\iota}_{s_t}'\mathbf{P}^h\boldsymbol{\varphi})\log(\boldsymbol{\iota}_{s_{t-j}}'\mathbf{P}^h\boldsymbol{\varphi})\\
&=\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2\mathbf{1}_{\{j=0\}}-\frac{1}{a^2}\,\boldsymbol{\lambda}_h'\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}'\boldsymbol{\lambda}_h\\
&\quad+\frac{1}{a^2}\boldsymbol{\lambda}_h'\left(\sum_{s_{t-j}=1}^{k}\bar{\xi}_{(s_{t-j})}\left(\sum_{s_t=1}^{k}\hat{\xi}_{(s_t|s_{t-j})}\boldsymbol{\iota}_{s_t}\right)\boldsymbol{\iota}_{s_{t-j}}'\right)\boldsymbol{\lambda}_h\\
&=\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2\mathbf{1}_{\{j=0\}}-\frac{1}{a^2}\,\boldsymbol{\lambda}_h'\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}'\boldsymbol{\lambda}_h\\
&\quad+\frac{1}{a^2}\,\boldsymbol{\lambda}_h'\left(\sum_{s_{t-j}=1}^{k}\bar{\xi}_{(s_{t-j})}\mathbf{P}^{j'}\boldsymbol{\iota}_{s_{t-j}}\boldsymbol{\iota}_{s_{t-j}}'\right)\boldsymbol{\lambda}_h\\
&=\bar{\boldsymbol{\xi}}'\boldsymbol{\sigma}^2\mathbf{1}_{\{j=0\}}+\frac{1}{a^2}\boldsymbol{\lambda}_h'((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}')\odot\mathbf{P}^j-\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}')\boldsymbol{\lambda}_h.
\end{aligned}
$$

For fixed h, as $j\to\infty$, $Cov[e_{t+h,t}^*,e_{t+h-j,t-j}^*]\to(1/a^2)\boldsymbol{\lambda}_h'((\bar{\boldsymbol{\xi}}\boldsymbol{\iota}')\odot(\boldsymbol{\iota}\bar{\xi}')-\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}')\boldsymbol{\lambda}_h=0$. □

Proof of Proposition 6.

$$
\begin{aligned}
&\mathrm{E}[L(Y_{t+h},\hat{Y}_{t+h,t}^*)]\\
&\quad=\mathrm{E}[\exp\{a\sigma_{s_{t+h}}v_{t+h}-\log(\hat{\boldsymbol{\xi}}_{t|t}'\mathbf{P}^h\boldsymbol{\varphi})\}]+\mathrm{E}[\log(\hat{\boldsymbol{\xi}}_{t|t}'\mathbf{P}^h\boldsymbol{\varphi})]-1\\
&\quad=\sum_{s_t=1}^{k}\sum_{s_{t+h}=1}^{k}\bar{\xi}_{(s_t)}\hat{\xi}_{(s_{t+h}|s_t)}\mathrm{E}[\exp\{a\sigma_{s_{t+h}}v_{t+h}-\log(\hat{\boldsymbol{\xi}}_{t|t}'\mathbf{P}^h\boldsymbol{\varphi})\}|S_{t+h}=s_{t+h},S_t=s_t]\\
&\qquad+\sum_{s_t=1}^{k}\bar{\xi}_{(s_t)}\mathrm{E}[\log(\hat{\boldsymbol{\xi}}_{t|t}'\mathbf{P}^h\boldsymbol{\varphi})|S_t=s_t]-1\\
&\quad=\sum_{s_t=1}^{k}\sum_{s_{t+h}=1}^{k}\bar{\xi}_{(s_t)}\hat{\xi}_{(s_{t+h}|s_t)}\exp\left\{-\log(\boldsymbol{\iota}_{s_t}'\mathbf{P}^h\boldsymbol{\varphi})+\frac{a^2}{2}\,\sigma_{s_{t+h}}^2\right\}+\sum_{s_t=1}^{k}\bar{\xi}_{(s_t)}\boldsymbol{\iota}_{s_t}'\log(\mathbf{P}^h\boldsymbol{\varphi})-1\\
&\quad=\sum_{s_t=1}^{k}\bar{\xi}_{(s_t)}\boldsymbol{\iota}_{s_t}'(\mathbf{P}^h\boldsymbol{\varphi})^{-1}\sum_{s_{t+h}=1}^{k}\hat{\xi}_{(s_{t+h}|s_t)}\exp\left\{\frac{a^2}{2}\sigma_{s_{t+h}}^2\right\}+\sum_{s_t=1}^{k}\bar{\xi}_{(s_t)}\boldsymbol{\iota}_{s_t}'\log(\mathbf{P}^h\boldsymbol{\varphi})-1\\
&\quad=\sum_{s_t=1}^{k}\bar{\xi}_{(s_t)}\boldsymbol{\iota}_{s_t}'(\mathbf{P}^h\boldsymbol{\varphi})^{-1}(\mathbf{P}^h\boldsymbol{\varphi})+\sum_{s_t=1}^{k}\bar{\xi}_{(s_t)}\boldsymbol{\iota}_{s_t}'\log(\mathbf{P}^h\boldsymbol{\varphi})-1
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} \boldsymbol{\iota}'_{s_t} \log(\mathbf{P}^h \boldsymbol{\varphi}) \\
&= \bar{\boldsymbol{\xi}}' \log(\mathbf{P}^h \boldsymbol{\varphi}).
\end{aligned}
$$

As $h \to \infty$ the expected loss becomes: $\mathrm{E}[L(Y_{t+h}, \hat{Y}^*_{t+h,t}; a)] = \bar{\boldsymbol{\xi}}' \log(\mathbf{P}^h \boldsymbol{\varphi}) \to \bar{\boldsymbol{\xi}}' \log(\boldsymbol{\iota}\bar{\boldsymbol{\xi}}' \boldsymbol{\varphi}) = \bar{\boldsymbol{\xi}}' \boldsymbol{\iota} \log(\bar{\boldsymbol{\xi}}' \boldsymbol{\varphi}) = \log(\bar{\boldsymbol{\xi}}' \boldsymbol{\varphi})$. □

Proof of Proposition 7.

$$
\begin{aligned}
&Cov[\psi^*_{t+h,t}, \psi^*_{t+h-j,t-j}] \\
&\quad = a^2 - a^2 \mathrm{E}[\exp\{a\sigma_{s_{t+h}} v_{t+h} - \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi})\}] \\
&\qquad - a^2 \mathrm{E}[\exp\{a\sigma_{s_{t+h-j}} v_{t+h-j} - \log(\hat{\boldsymbol{\xi}}'_{t-j|t-j} \mathbf{P}^h \boldsymbol{\varphi})\}] \\
&\qquad + a^2 \mathrm{E}[\exp\{a\sigma_{s_{t+h}} v_{t+h} + a\sigma_{s_{t+h-j}} v_{t+h-j} - \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) - \log(\hat{\boldsymbol{\xi}}'_{t-j|t-j} \mathbf{P}^h \boldsymbol{\varphi})\}] \\
&\quad = a\mathrm{E}[\psi^*_{t+h,t}] + a\mathrm{E}[\psi^*_{t+h-j,t-j}] - a^2 \\
&\qquad + a^2 \mathrm{E}[\exp\{a\sigma_{s_{t+h}} v_{t+h} + a\sigma_{s_{t+h-j}} v_{t+h-j} - \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) - \log(\hat{\boldsymbol{\xi}}'_{t-j|t-j} \mathbf{P}^h \boldsymbol{\varphi})\}] \\
&\quad = -a^2 + a^2 \mathrm{E}[\exp\{a\sigma_{s_{t+h}} v_{t+h} + a\sigma_{s_{t+h-j}} v_{t+h-j} - \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) - \log(\hat{\boldsymbol{\xi}}'_{t-j|t-j} \mathbf{P}^h \boldsymbol{\varphi})\}].
\end{aligned}
$$

If $j = 0$ we get the variance of the generalized forecast error:

$$
\begin{aligned}
&V[\psi^*_{t+h,t}] \\
&\quad = -a^2 + a^2 \sum_{s_t=1}^{k} \sum_{s_{t+h}=1}^{k} \bar{\xi}_{(s_t)} \hat{\xi}_{(s_{t+h}|s_t)} \exp\{-2\log(\boldsymbol{\iota}'_{s_t} \mathbf{P}^h \boldsymbol{\varphi}) + 2a^2 \sigma^2_{s_{t+h}}\} \\
&\quad = -a^2 + a^2 \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} (\boldsymbol{\iota}'_{s_t} \mathbf{P}^h \boldsymbol{\varphi})^{-2} \sum_{s_{t+h}=1}^{k} \hat{\xi}_{(s_{t+h}|s_t)} \exp\{2a^2 \sigma^2_{s_{t+h}}\} \\
&\quad = -a^2 + a^2 \sum_{s_t=1}^{k} \bar{\xi}_{(s_t)} (\boldsymbol{\iota}'_{s_t} \mathbf{P}^h \boldsymbol{\varphi})^{-2} (\boldsymbol{\iota}'_{s_t} \mathbf{P}^h \boldsymbol{\varphi}^4),
\end{aligned}
$$

where $\boldsymbol{\varphi}^4 \equiv \exp\{2a^2 \boldsymbol{\sigma}^2\} = \boldsymbol{\varphi} \odot \boldsymbol{\varphi} \odot \boldsymbol{\varphi} \odot \boldsymbol{\varphi}$.

For $0 < j < h$ we get:

$$
\begin{aligned}
&Cov[\psi^*_{t+h,t}, \psi^*_{t+h-j,t-j}] \\
&\quad = -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \sum_{s_t=1}^{k} \sum_{s_{t+h-j}=1}^{k} \sum_{s_{t+h}=1}^{k} \bar{\xi}_{(s_{t-j})} \hat{\xi}_{(s_t|s_{t-j})} \hat{\xi}_{(s_{t+h-j}|s_t)} \hat{\xi}_{(s_{t+h}|s_{t+h-j})} \cdot \\
&\qquad \cdot \mathrm{E}[\exp\{a\sigma_{s_{t+h}} v_{t+h} + a\sigma_{s_{t+h-j}} v_{t+h-j} - \log(\hat{\boldsymbol{\xi}}'_{t|t} \mathbf{P}^h \boldsymbol{\varphi}) \\
&\qquad - \log(\hat{\boldsymbol{\xi}}'_{t-j|t-j} \mathbf{P}^h \boldsymbol{\varphi})\} | S_{t-j} = s_{t-j}, S_{t+h-j} = s_{t+h-j}, S_t = s_t, S_{t+h} = s_{t+h}] \\
&\quad = -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \sum_{s_t=1}^{k} \sum_{s_{t+h-j}=1}^{k} \sum_{s_{t+h}=1}^{k} \bar{\xi}_{(s_{t-j})} \hat{\xi}_{(s_t|s_{t-j})} \hat{\xi}_{(s_{t+h-j}|s_t)} \hat{\xi}_{(s_{t+h}|s_{t+h-j})}
\end{aligned}
$$

$$\cdot \exp\left\{ -\log(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi}) - \log(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi}) + \frac{a^2}{2}\sigma^2_{s_{t+h}} + \frac{a^2}{2}\sigma^2_{s_{t+h-j}} \right\}$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi})^{-1} \sum_{s_t=1}^{k} \hat{\xi}_{(s_t|s_{t-j})}(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi})^{-1}$$

$$\cdot \sum_{s_{t+h-j}=1}^{k} \hat{\xi}_{(s_{t+h-j}|s_t)} \exp\left\{\frac{a^2}{2}\sigma^2_{s_{t+h-j}}\right\} \sum_{s_{t+h}=1}^{k} \hat{\xi}_{(s_{t+h}|s_{t+h-j})} \exp\left\{\frac{a^2}{2}\sigma^2_{s_{t+h}}\right\}$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi})^{-1} \sum_{s_t=1}^{k} \hat{\xi}_{(s_t|s_{t-j})}(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi})^{-1}$$

$$\cdot \sum_{s_{t+h-j}=1}^{k} \hat{\xi}_{(s_{t+h-j}|s_t)} \exp\left\{\frac{a^2}{2}\sigma^2_{s_{t+h-j}}\right\} \boldsymbol{\iota}'_{s_{t+h-j}}\mathbf{P}^j\boldsymbol{\varphi}$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi})^{-1} \cdot \sum_{s_t=1}^{k} \hat{\xi}_{(s_t|s_{t-j})}(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi})^{-1} \cdot (\boldsymbol{\varphi}' \odot (\boldsymbol{\iota}'_{s_t}\mathbf{P}^{h-j}))\mathbf{P}^j\boldsymbol{\varphi},$$

and for $j \geqslant h$ we get:

$$Cov[\psi^*_{t+h,t}, \psi^*_{t+h-j,t-j}]$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \sum_{s_{t+h-j}=1}^{k} \sum_{s_t=1}^{k} \sum_{s_{t+h}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}\hat{\xi}_{(s_{t+h-j}|s_{t-j})}\hat{\xi}_{(s_t|s_{t+h-j})}\hat{\xi}_{(s_{t+h}|s_t)}$$

$$\cdot \mathrm{E}[\exp\{a\sigma_{s_{t+h}}v_{t+h} + a\sigma_{s_{t+h-j}}v_{t+h-j} - \log(\hat{\boldsymbol{\xi}}'_{t|t}\mathbf{P}^h\boldsymbol{\varphi})$$

$$- \log(\hat{\boldsymbol{\xi}}'_{t-j|t-j}\mathbf{P}^h\boldsymbol{\varphi})\}|S_{t-j} = s_{t-j}, S_{t+h-j} = s_{t+h-j}, S_t = s_t, S_{t+h} = s_{t+h}]$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \sum_{s_{t+h-j}=1}^{k} \sum_{s_t=1}^{k} \sum_{s_{t+h}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}\hat{\xi}_{(s_{t+h-j}|s_{t-j})}\hat{\xi}_{(s_t|s_{t+h-j})}\hat{\xi}_{(s_{t+h}|s_t)}$$

$$\cdot \exp\left\{ -\log(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi}) - \log(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi}) + \frac{a^2}{2}\sigma^2_{s_{t+h}} + \frac{a^2}{2}\sigma^2_{s_{t+h-j}} \right\}$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi})^{-1} \cdot \sum_{s_{t+h-j}=1}^{k} \hat{\xi}_{(s_{t+h-j}|s_{t-j})} \cdot \exp\left\{\frac{a^2}{2}\sigma^2_{s_{t+h-j}}\right\}$$

$$\cdot \sum_{s_t=1}^{k} \hat{\xi}_{(s_t|s_{t+h-j})}(\boldsymbol{\iota}'_{s_t}\mathbf{P}^h\boldsymbol{\varphi})^{-1} \sum_{s_{t+h}=1}^{k} \hat{\xi}_{(s_{t+h}|s_t)} \exp\left\{\frac{a^2}{2}\sigma^2_{s_{t+h}}\right\}$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi})^{-1} \cdot \sum_{s_{t+h-j}=1}^{k} \hat{\xi}_{(s_{t+h-j}|s_{t-j})} \cdot \exp\left\{\frac{a^2}{2}\sigma^2_{s_{t+h-j}}\right\}$$

$$= -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \bar{\bar{\xi}}_{(s_{t-j})}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi})^{-1}(\boldsymbol{\iota}'_{s_{t-j}}\mathbf{P}^h\boldsymbol{\varphi}) = 0.$$

The autocovariance is zero for all lags greater than or equal to the forecast horizon. □

References

Albert, J., Chib, S., 1993. Bayesian analysis via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. Journal of Business and Economic Statistics 11, 1–15.

Batchelor, R., Peel, D.A., 1998. Rationality testing under asymmetric loss. Economics Letters 61, 49–54.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Chesher, A., Irish, M., 1987. Residual analysis in the grouped and censored normal linear model. Journal of Econometrics 34, 33–61.

Chib, S., 1996. Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics 75, 79–97.

Christoffersen, P., Jacobs, K., 2004. The importance of the loss function in option valuation. Journal of Financial Economics 72, 291–318.

Christoffersen, P.F., Diebold, F.X., 1996. Further results on forecasting and model selection under asymmetric loss. Journal of Applied Econometrics 11, 561–572.

Christoffersen, P.F., Diebold, F.X., 1997. Optimal prediction under asymmetric loss. Econometric Theory 13, 808–817.

Corradi, V., Swanson, N.R., 2002. A consistent test for nonlinear out of sample predictive accuracy. Journal of Econometrics 110, 353–381.

Diebold, F.X., 2004. Elements of Forecasting, third ed. Southwestern, Berlin.

Diebold, F.X., Lopez, J., 1996. Forecast evaluation and combination. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics. North-Holland, Amsterdam, pp. 241–268.

Drost, F.C., Nijman, T.E., 1993. Temporal aggregation of GARCH processes. Econometrica 61, 909–928.

Elliott, G., Timmermann, A., 2004. Optimal forecast combinations under general loss functions and forecast error distributions. Journal of Econometrics 122, 47–79.

Elliott, G., Komunjer, I., Timmermann, A., 2005. Estimation and testing of forecast rationality under flexible loss. Review of Economic Studies 72, 1107–1125.

Gourieroux, C., Monfort, A., Renault, E., Trongnon, A., 1987. Generalized residuals. Journal of Econometrics 34, 5–32.

Granger, C.W.J., 1969. Prediction with a generalized cost function. OR 20, 199–207.

Granger, C.W.J., 1999. Outline of forecast theory using generalized cost functions. Spanish Economic Review 1, 161–173.

Granger, C.W.J., Newbold, P., 1986. Forecasting Economic Time Series, second ed. Academic Press, New York.

Granger, C.W.J., Pesaran, M.H., 2000. Economic and statistical measures of forecast accuracy. Journal of Forecasting 19, 537–560.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, New Jersey.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–386.

Hoque, A., Magnus, J.R., Pesaran, B., 1988. The exact multi-period mean-square forecast error for the first-order autoregressive model. Journal of Econometrics 39, 327–346.

James, W., Stein, C., 1961. Estimation with quadratic loss. In: , Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, CA, pp. 361–379.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H., Lee, T.-C., 1985. The Theory and Practice of Econometrics, second ed. Wiley, New York.

Kim, C.-J., Nelson, C.R., 1999. State-space Models with Regime Switching. MIT Press, Cambridge, US.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, second ed. Springer, New York.

Magnus, J.R., Pesaran, B., 1989. The exact multi-period mean-square forecast error for the first-order autoregressive model with an intercept. Journal of Econometrics 42, 157–179.

Magnus, J.R., Pesaran, B., 1991. The bias of forecasts from a first-order autoregression. Econometric Theory 7, 222–235.

Nobay, A.R., Peel, D.A., 2003. Optimal discretionary monetary policy in a model of asymmetric central bank preferences. Economic Journal 113, 657–665.

Patton, A.J., Timmermann, A., 2005. Testing forecast optimality under unknown loss. Journal of the American Statistical Association, forthcoming.

Rosenblatt, M., 1952. Remarks on a multivariate transformation. Annals of Mathematical Statistics 23, 470–472.

Rudin, W., 1964. Principles of Mathematical Analysis, second ed. McGraw-Hill, New York.

Sancetta, A., Satchell, S., 2004. Cost of capital and the regulator's preferences: an investigation into a new method of estimating regulatory beta. Cambridge Working Papers in Economics 2004–0441.

Sentana, E., 2005. Least squares predictions and mean-variance analysis. Journal of Financial Econometrics 3, 56–78.

Skouras, S., 2006. Decisionmetrics: a decision-based approach to Econometric modeling. Journal of Econometrics, forthcoming.

Varian, H.R., 1974. A Bayesian approach to real estate assessment. In: Fienberg, S.E., Zellner, A. (Eds.), Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage. North-Holland, Amsterdam, pp. 195–208.

Whiteman, C.H., 1996. Bayesian prediction under asymmetric linear loss: forecasting state tax revenues in Iowa. In: Johnson, W.O., Lee, J.C., Zellner, A. (Eds.), Forecasting Prediction and Modeling in Statistics and Econometrics. Springer, New York.

Zellner, A., 1986. Bayesian estimation and prediction using asymmetric loss functions. Journal of the American Statistical Association 81, 446–451.

Zellner, A., Chen, B., 2001. Bayesian modeling of economics and data requirements. Macroeconomic Dynamics 5, 673–700.

Zellner, A., Hong, C., 1989. Forecasting international growth rates using Bayesian shrinkage and other procedures. Journal of Econometrics 40, 183–202.