Properties of Optimal Forecasts[*]

Andrew J. Patton
London School of Economics

Allan Timmermann
University of California, San Diego

24 March, 2003

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 should be serially uncorrelated at the single period horizon with increasing variance as the forecast horizon grows. Using analytical results, we show in this paper that all the 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. Our theoretical results suggest that many of the conclusions in the empirical literature concerning suboptimality of forecasts could be premature. We extend the properties that an optimal forecast should have to a more general setting than previously considered in the literature. We also present results on forecast error properties that may be tested when the forecaster's loss function is unknown, and introduce a change of measure, following which the optimum forecast errors for general loss functions have the same properties as optimum errors under MSE loss.

Keywords: forecast evaluation, loss function, rationality, efficient markets.

J.E.L. Codes: C53, C22, C52

[*] The authors would like to thank Jeremy Berkowitz, Carlos Capistran, Graham Elliott, Raffaella Giacomini, Clive Granger, Peter Hansen, Oliver Linton, Mark Machina, Francisco Penaranda, Kevin Sheppard, Hal White and seminar participants at LSE, McGill, UC Riverside and the 2002 EC^2 conference for useful comments. All remaining deficiencies are the responsibility of the authors. Patton gratefully acknowledges financial support from the IAM Programme in Hedge Fund Research at LSE. Patton: London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Email: a.patton@lse.ac.uk. Timmermann: Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla CA 92093-0508, U.S.A. Email: atimmerm@ucsd.edu.

1 Introduction

Knowledge of the properties possessed by an optimal forecast is crucial in many key areas of economics and finance such as in tests of the efficient market hypothesis in foreign exchange, bond and stock markets and tests of the rationality of decision makers in a variety of macroeconomic applications. Almost without any exception empirical work has relied on testing properties that optimal forecasts have under mean squared error (MSE) loss.[1] These properties include unbiasedness of the forecast, lack of serial correlation in one-step-ahead forecast errors, serial correlation of order $h-1$ at the h-period horizon and 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 decision maker's loss function.

Increasingly the assumption of symmetric loss has been questioned in the literature. Christoffersen and Diebold (1997), Diebold (2001), Granger and Newbold (1986), Granger and Pesaran (2000), Pesaran and Skouras (2001), Skouras (2001) and West, Edison and Cho (1993) call for a more decision theoretic approach to forecasting that considers the losses derived from over- and underpredictions. There are often no reason why losses should be symmetric around a zero forecast error (the perfect prediction). For instance, financial analysts' forecasts have been found to be strongly biased[2] and it is easy to understand why. Underprediction of corporate earnings is likely to lead to lower demand for stocks, lower stock prices and a worsened relationship between the analyst and the firm in question. Overpredictions, on the other hand, are likely to be better tolerated.

In this paper we 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 data generating process. While bias of the optimal forecast 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), to our knowledge, failure of the

[1] See, e.g., Brown and Maital (1981), Cargill and Meyer (1980), De Bondt and Bange (1992), Dokko and Edelstein (1989), Figlewski and Wachtel (1981), Keane and Runkle (1990, 1998), Lakonishok (1980), Mishkin (1981), Muth (1961), Pesando (1975) and Schroeter and Smith (1986) and Zarnowitz (1985).

[2] See De Bondt and Thaler (1990) and Abarbanell and Bernard (1992) for example.

remaining optimality properties has not previously been shown[3,4].

We derive closed-form results in the context of a commonly used asymmetric loss function (linear-exponential, or "linex") and a widely used nonlinear data generating process, namely the regime switching model suggested by Hamilton (1989). We find that not only can the optimal forecast be biased, but the 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 next extend the properties that an optimal forecast should have to a more general setting than that previously considered in the literature. Our results suggest that many of the conclusions in the empirical literature concerning suboptimality of forecasts have been premature. We prove that the *expected loss*, rather than the forecast error variance, is a non-decreasing function of the forecast horizon and that a "generalized forecast error" has mean zero and limited serial correlation, and is a martingale difference sequence at the single-period horizon.

We also introduce a transformation from the usual probability measure to a "MSE-loss probability measure", under which the optimal forecasts are unbiased and forecast errors are serially uncorrelated, in spite of the fact that these properties generally fail to hold under the physical measure. These results are analogous to the change of measure from the physical measure to the risk-neutral measure, under which assets may be priced as though investors are risk-neutral.

Finally, we establish some surprising new results that trade off restrictions on the loss function against restrictions on the data generating process. In situations where the conditional higher order moments of the forecast variable are constant, we show that although the optimal forecast may well be biased, the one-step optimal forecast errors are not serially correlated while the h-step forecast errors are at most MA(h-1). This holds irrespective of the shape of the loss function. This offers a new way to test optimality of forecast errors that is robust to the loss function, but requires restrictions on the underlying data generating process. This result will be useful in the common situation where the shape of the loss function is unknown, whereas the restrictions on the data

[3] Under asymmetric loss functions such as lin-lin and linex, Christoffersen and Diebold (1997) establish that the optimal forecast is biased and characterize the optimal bias analytically. Their study does not, however, consider the other properties of optimal forecast errors such as lack of serial correlation and non-decreasing variance.

[4] Hoque, *et al.* (1988), and Magnus and Pesaran (1987 and 1989) discuss violations of the standard properties of optimal forecasts caused by estimation error, rather than by a choice of loss function different from MSE. In this paper we consider the case of zero estimation error, to rule this out as a cause of apparent violations.

generating process can be tested empirically.

The outline of the paper is as follows. Section 2 summarizes the properties of optimal linear predictions under stationarity and squared error loss. Section 3 demonstrates how each of these properties can be violated under asymmetric loss in the context of two nonlinear data generating processes. Section 4 derives testable properties of the forecast errors when restrictions are imposed on the loss function while Section 5 establishes properties of optimal forecasts under general loss and verifies that these are satisfied for the models considered in Section 3. Section 5 also contains the change of measure results. Section 6 concludes. An appendix contains technical details and proofs.

2 Properties of optimal linear predictions under squared error loss

Suppose that a decision maker is interested in forecasting some univariate time series, $Y = \{Y_t; t = 1, 2, ...\}$, h steps ahead given information at time t, Ω_t. At a minimum Ω_t includes the filtration generated by $\{Y_{t-k}; k \geq 0\}$, but it may also be expanded to include other information. Optimality of the forecast must be established with reference to the loss function that the decision maker is trying to minimize. Although the loss may depend on both the outcome, Y_{t+h}, and the prediction, $\hat{Y}_{t+h,t}$, it is very common to assume that the loss function simply depends on the h-step-ahead forecast error

$$e_{t+h,t} = Y_{t+h} - \hat{Y}_{t+h,t} \tag{1}$$

and to impose the following restrictions on the loss function, see, e.g., Granger (1999), Diebold (2001):

Assumption 1: $L(0) = 0$ (minimal loss of zero).

Assumption 2: $L(e_{t+h,t}) \geq 0$ for all $e_{t+h,t}$

Assumption 3: $L(e_{t+h,t})$ is non-decreasing in $|e_{t+h,t}|$:

$$\begin{aligned} L(e^*_{t+h,t}) &\geq L(e^{**}_{t+h,t}) \text{ if } e^*_{t+h,t} > e^{**}_{t+h,t} \geq 0 \\ L(e^*_{t+h,t}) &\geq L(e^{**}_{t+h,t}) \text{ if } e^*_{t+h,t} < e^{**}_{t+h,t} \leq 0 \end{aligned}$$

Below we will make use of a smaller set of assumptions.

While the above properties are quite general, the vast majority of work on optimal forecasts has been derived in the context of linear predictions under mean squared error (MSE) loss:

$$L(e_{t+h,t}) = ae_{t+h,t}^2, \;\; a > 0. \tag{2}$$

This is clearly a special case but given its dominance in applied work it is useful to outline the properties that optimal forecasts have under MSE loss. For this purpose, suppose that Y_t has zero mean and is covariance stationary.[5] Wold's representation theorem then establishes that it can be represented as a linear combination of serially uncorrelated white noise terms:

$$Y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} \tag{3}$$

where $\varepsilon_t = Y_t - P(Y_t|y_{t-1}, y_{t-2}, ...)$ is white noise and $P(Y_t|y_{t-1}, y_{t-2}, ...)$ is the linear least squares projection of Y_t on $y_{t-1}, y_{t-2}, ...$.[6] ε_t satisfies the following conditions

$$\begin{aligned} E[\varepsilon_t] &= 0 \\ E[\varepsilon_t^2] &= \sigma^2 \geq 0 \\ E[\varepsilon_t \varepsilon_s] &= 0, \text{ for all } t \neq s. \end{aligned} \tag{4}$$

These conditions imply that ε_t is serially uncorrelated with constant unconditional variance and zero mean. The weights θ_j are such that $\theta_0 = 1$ and $\sum_{j=0}^{\infty} \theta_j^2 < \infty$. Assuming that $\Omega_t = \sigma(\varepsilon_t, \varepsilon_{t-1}, ..)$, where $\sigma(X)$ is the sigma algebra generated by X, the linear prediction of Y_{t+h} that minimizes MSE loss can easily be derived from this infinite order moving average process:

$$P(Y_{t+h}|\Omega_t) = \sum_{s=0}^{\infty} \theta_{h+s} \varepsilon_{t-s}. \tag{5}$$

The forecast error is

$$e_{t+h,t} = \sum_{s=0}^{h-1} \theta_s \varepsilon_{t+h-s}, \tag{6}$$

so the MSE is

$$MSE\left(P(Y_{t+h}|\Omega_t)\right) = a\sigma^2 \left(\sum_{s=0}^{h-1} \theta_s^2 \right) \tag{7}$$

It follows from these expressions that, under MSE loss, the optimal forecast has the following properties:

[5] A linearly deterministic component can also be added, but this has no consequence for our analysis.

[6] The linear projection of Y_t on $y_{t-1}, y_{t-2}, ...$ can also be expanded as a Volterra series that includes higher order powers such as y_{t-1}^2, y_{t-1}^3, c.f. Granger and Terasvirta (1993).

1. The forecast is unbiased:

$$E[e_{t+h,t}] = 0.$$

2. The forecast error variance is non-decreasing in the forecast horizon, h. This can readily be seen from (7)

$$Var\left(e_{t+h,t}\right) = \sigma^2 \left(\sum_{i=0}^{h-1} \theta_i^2\right) \geq \sigma^2 \left(\sum_{i=0}^{h-2} \theta_i^2\right) = Var\left(e_{t+h-1,t}\right).$$

3. The one-step forecast errors are white noise:

$$e_{t+1,t} = \varepsilon_{t+1}$$

   which, by construction, is serially uncorrelated with mean zero.

4. The $h-$step forecast errors are at most $MA(h-1)$:

$$e_{t+h,t} = \varepsilon_{t+h} + \theta_1 \varepsilon_{t+h-1} + ... + \theta_{h-1} \varepsilon_{t+1}.$$

Notice also that while the conditional forecast of the mean, $P(Y_{t+h}|\Omega_t)$, is time-varying and depends on all shocks $\{\varepsilon_{t-i}\}_{i=0}^{\infty}$ up to time t, the variance of the conditional forecast error is time-invariant and only depends on the time horizon, h.[7]

Properties such as these have been extensively tested in empirical studies of optimality of predictions or rationality of forecasts. However, as we show in the next section, they cease to be valid when the assumption of MSE loss is relaxed.

3 Violation of the Optimality Properties under Asymmetric Loss

In this section we demonstrate how each of the properties established under MSE loss and linear least squares projections may be rejected under more general assumptions about the loss function and the data generating process. We set up a specific example, making reasonable assumptions about the forecaster's loss function and the DGP, and then show that in this example all of the

[7] Although the results were derived under linear least squares projections, they can be demonstrated for more general loss functions when the innovations $\{\varepsilon_t\}$ are Gaussian. For this case the linear projection of Y_t on past shocks is identical to the (optimal) conditional expectation so that the assumption of linearity of the forecast is not restrictive.

standard properties of an optimal forecast are violated. Our example is an idealised case, where in addition to knowing the form of the DGP, the forecaster is assumed to also know the parameters of the DGP, removing estimation error from the problem. The forecasts in this example are thus perfectly optimal. Violations of the standard properties of optimal forecasts caused by estimation error rather than asymmetric loss have been investigated in Hoque, *et al.* (1988), and Magnus and Pesaran (1987, 1989).

3.1 A simple example

We establish our results in the context of the linear-exponential (linex) loss function, which allows for asymmetries:

$$L\left(e_{t+h,t};a\right)=\exp\left\{ae_{t+h,t}\right\}-ae_{t+h,t}-1,\;\; a\neq 0 \tag{8}$$

This loss function has been used extensively to demonstrate the effect of asymmetric loss, c.f. Varian (1974), Zellner (1986) and Christoffersen and Diebold (1997). An optimal forecast is defined by minimising the conditional expected loss:

$$\hat{Y}^{*}_{t+h,t}\equiv\arg\min_{\hat{Y}}E_t\left[L\left(Y_{t+h},\hat{Y}\right)\right]$$

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

$$E_t\left[\frac{\partial L\left(Y_{t+h}-\hat{Y}^{*}_{t+h,t};a\right)}{\partial\hat{Y}_{t+h,t}}\right]=a-aE_t\left[\exp\left\{a\left(Y_{t+h}-\hat{Y}^{*}_{t+h,t}\right)\right\}\right]=0$$

We derive analytical expressions for the optimal forecast and the expected loss using a popular nonlinear data generating process, namely a regime switching model of the type proposed by Hamilton (1989)[8]. Suppose that $\{Y_t\}$ is generated by a simple mixture of normals regime switching model driven by some underlying state process, S_t :

$$\begin{aligned} Y_{t+1} &= \mu_{s_{t+1}}+\sigma_{s_{t+1}}v_{t+1} \\ v_{t+1} &\sim i.i.d.\; N(0,1) \\ s_{t+1} &= 1,...,k. \end{aligned} \tag{9}$$

[8] In a previous version of this paper we also presented results for the case that the DGP was another popular nonlinear process; the GARCH(1,1) model proposed by Bollerslev (1986). Little additional intuition was to be had with this second example and so we do not discuss it in the interests of brevity.

We assume that the state indicator function, S_{t+1}, is independently distributed of all past, current and future values of v_{t+1}. The state-specific means and variances can be collected in $k \times 1$ vectors, $\boldsymbol{\mu} = [\mu_1, ..., \mu_k]'$, $\boldsymbol{\sigma}^2 = \left[\sigma_1^2, ..., \sigma_k^2\right]'$. Conditional on a given realization of the state variable, $S_{t+1} = s_{t+1}$, Y_{t+1} is Gaussian with mean $\mu_{s_{t+1}}$ and variance $\sigma^2_{s_{t+1}}$, but the states are assumed to be unobserved random variables and Y_{t+1} can be strongly non-Gaussian unconditionally.

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 Markov chain with transition probability matrix

$$\mathbf{P}(s_{t+1}|s_t) = \mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & \vdots \\ \vdots & \vdots & ... & p_{k-1k} \\ p_{k1} & \cdots & p_{kk-1} & p_{kk} \end{bmatrix} \tag{10}$$

where each row of $\mathbf{P}$ sums to one. The vector comprising the probability of being in state s_{t+h} at time $t+h$ given Ω_t is denoted by $\hat{\boldsymbol{\pi}}_{s_{t+h},t}$, i.e. $\hat{\boldsymbol{\pi}}_{s_{t+h},t} = (\Pr(S_{t+h} = 1|\Omega_t), ..., \Pr(S_{t+h} = k|\Omega_t))'$, while $\bar{\boldsymbol{\pi}}$ is the vector of unconditional or ergodic state probabilities that solve the equation $\bar{\boldsymbol{\pi}}'\mathbf{P} = \bar{\boldsymbol{\pi}}'$. Note that $\hat{\boldsymbol{\pi}}_{s_t,t}$ will not be a vector of ones and zeros, as the variable S_t is not Ω_t-measurable.

Consider the h-step-ahead forecasting problem. Using the conditional normality of v_{t+h}, the expected loss is

$$\begin{aligned} E_t\left[L\left(e_{t+h,t};a\right)\right] &= E_t\left[\exp\left\{a\left(Y_{t+h} - \hat{Y}_{t+h,t}\right)\right\}\right] - aE_t\left[Y_{t+h}\right] + a\hat{Y}_{t+h,t} - 1 \\ &= \sum_{s_{t+h}=1}^{k} \hat{\pi}_{s_{t+h},t} E_t\left[\exp\left\{a\left(Y_{t+h} - \hat{Y}_{t+h,t}\right)\right\}|S_{t+h} = s_{t+h}\right] \\ &\quad -a\sum_{s_{t+h}=1}^{k} \hat{\pi}_{s_{t+h},t} E_t\left[Y_{t+h}|S_{t+h} = s_{t+h}\right] + a\hat{Y}_{t+h,t} - 1 \\ &= \hat{\boldsymbol{\pi}}'_{s_t,t}\mathbf{P}^h \exp\left\{a\boldsymbol{\mu} - a\hat{Y}_{t+h,t} + \frac{a^2}{2}\boldsymbol{\sigma}^2\right\} - a\hat{\boldsymbol{\pi}}'_{s_t,t}\mathbf{P}^h\boldsymbol{\mu} + a\hat{Y}_{t+h,t} - 1 \end{aligned} \tag{11}$$

where we used $E_t[.]$ as shorthand notation for $E[.|\Omega_t]$, the conditional expectation given Ω_t. Note that in this paper all $\exp\{\cdot\}$ and $\log(\cdot)$ operators are applied element-by-element to vector and matrix arguments. Differentiating with respect to $\hat{Y}_{t+h,t}$ and setting the resulting expression equal to zero gives the first order condition

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

If $\mu_1 = \mu_2 = ... = \mu_k = \mu$, we can solve for $\hat{Y}^*_{t+h,t}$ to get an expression that is easier to interpret:

$$\hat{Y}^*_{t+h,t} = \mu + \frac{1}{a} \log \left(\hat{\boldsymbol{\pi}}'_{s_t,t} \mathbf{P}^h \boldsymbol{\varphi} \right), \tag{12}$$

where $\boldsymbol{\varphi} \equiv \exp \left\{ \frac{a^2}{2} \boldsymbol{\sigma}^2 \right\}$. The associated h-step forecast error is

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

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

Proposition 1 *The unconditional and conditional bias in the optimal forecast error for the Markov switching process (9) is given by:*

$$\begin{aligned} E_t \left[e^*_{t+h,t} \right] &= -\frac{1}{a} \log \left(\hat{\boldsymbol{\pi}}'_{s_t,t} \mathbf{P}^h \boldsymbol{\varphi} \right) && (13) \\ E \left[e^*_{t+h,t} \right] &= -\frac{1}{a} \bar{\boldsymbol{\pi}}' \boldsymbol{\lambda}_h && (14) \\ &\to -\frac{1}{a} \log \left(\bar{\boldsymbol{\pi}}' \boldsymbol{\varphi} \right) \ \textit{as } h \to \infty \end{aligned}$$

where $\boldsymbol{\lambda}_h \equiv \log \left(\mathbf{P}^h \boldsymbol{\varphi} \right)$. *Thus 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:

$$\begin{aligned} \boldsymbol{\mu} &= [0, 0]' \\ \boldsymbol{\sigma} &= [0.5, 2]' \\ P &= \begin{bmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{bmatrix} \text{ so} \\ \bar{\boldsymbol{\pi}} &= \left[\frac{2}{3}, \frac{1}{3} \right]' \end{aligned}$$

The unconditional mean of Y_t is zero, and the unconditional variance is $\bar{\boldsymbol{\pi}}' \boldsymbol{\sigma}^2 = 1.5$. This parameterisation is not dissimilar to the empirical results obtained when this model is estimated on macroeconomic or financial data. For this particular parameterization the optimal bias in $e^*_{t+1,t}$ is -1.17, indicating that it is optimal to over-predict. Figure 1 shows the 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.

[INSERT FIGURE 1 HERE]

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 10.0%. Under symmetric loss the optimal forecast is the mean, and so under symmetric distributions the amount of probability mass either side of the forecast would be 50%. In Figure 2 we plot the optimal forecast bias as a function of the forecast horizon (using the steady-state weights as initial probabilities). The bias for this case is an increasing (in absolute value) function of h and asymptotes to -1.17.

[INSERT FIGURE 2 HERE]

We next demonstrate the violation of property 2. This is best done using some new notation. We let $\odot$ be the Hadamard (element-by-element) product, and $\boldsymbol{\iota}$ be a $k \times 1$ vector of ones. The result is as follows:

Proposition 2 *The variance of the forecast error from the Markov switching process (9) associated with the optimum forecast is given by*

$$Var\left(e^*_{t+h,t}\right) = \bar{\boldsymbol{\pi}}'\boldsymbol{\sigma}^2 + \frac{1}{a^2}\boldsymbol{\lambda}'_h\left(\left(\bar{\boldsymbol{\pi}}\boldsymbol{\iota}'\right)\odot\mathbf{I} - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'\right)\boldsymbol{\lambda}_h \tag{15}$$

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, $\bar{\boldsymbol{\pi}}'\boldsymbol{\sigma}^2$.

Corollary 3 *The mean-square forecast error (MSFE) from the Markov switching process (9) associated with the optimum forecast is given by*

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

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 $\bar{\boldsymbol{\pi}}'\boldsymbol{\sigma}^2 + \left(\frac{1}{a}\log\left(\bar{\boldsymbol{\pi}}'\boldsymbol{\varphi}\right)\right)^2$.

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

Using the numerical example described above the unconditional variance of the optimal forecast error as a function of the forecast horizon is shown in Figure 3.

[INSERT FIGURE 3 HERE]

Thus it is possible that the forecast error at the distant future has a higher variance than at the near future[9]. The reason for this surprising result lies in the mis-match of the forecast objective function, L and the variance of the forecast error, $Var(e_{t+h,t})$, and thus does not occur when using quadratic loss (see next section). Such a mismatch of the objective function and the performance metric is common in economics, c.f. Christoffersen and Jacobs (2002) and Corradi and Swanson (2002).

Using the expression for $Var\left(e^*_{t+h,t}\right)$ in Proposition 2, 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:

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

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:

$$P = \boldsymbol{\iota}\bar{\boldsymbol{\pi}}'.$$

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 $\boldsymbol{\lambda}_h = \boldsymbol{\iota}\log\left(\bar{\boldsymbol{\pi}}'\boldsymbol{\varphi}\right)$ for all h, so:

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

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

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.

Violation of properties 3 and 4: Now consider the autocorrelation function of the optimal forecast errors. In the standard linear, quadratic loss framework an optimal h-step forecast is a MA process of order no greater than $(h-1)$. This implies that all autocovariances beyond the $(h-1)^{th}$ lag are zero. In our setting this need not hold:

Proposition 4 *The h-step-ahead forecast error from the Markov switching process (9) is generally serially correlated with autocovariance given by*

$$Cov\left[e^*_{t+h,t}, e^*_{t+h-j,t-j}\right] = \bar{\boldsymbol{\pi}}'\boldsymbol{\sigma}^2\mathbf{1}_{\{j=0\}} + \frac{1}{a^2}\boldsymbol{\lambda}_h'\left(\left(\bar{\boldsymbol{\pi}}\boldsymbol{\iota}'\right)\odot P^j - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'\right)\boldsymbol{\lambda}_h. \tag{16}$$

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

Using the same parameterization as in the earlier example, the autocorrelation function for various forecast horizons is presented in Figure 4.

[INSERT FIGURE 4 HERE]

Notice that for all forecast horizons there exist positive autocorrelations beyond $h-1$. 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).[10]

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 Nijmann (1993).

[10] We can again consider the two special cases: *iid* Normal ($\sigma_1 = \sigma_2 = \sigma$), and *iid* mixture of normals ($\mathbf{P} = \boldsymbol{\iota}\bar{\boldsymbol{\pi}}'$). 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. We discuss this result more generally in Section 4.

3.2 What drives the Results?

So far we have established that all four of the properties traditionally associated with an optimal forecast may be violated under linex loss for nonlinear data generating processes. However, it is not entirely clear what drives the results, since the interaction between nonlinearity and asymmetric loss can be difficult to disentangle. In this section we therefore investigate the effects of relaxing the assumptions of MSE loss and linear projections one at a time by considering the properties of optimal forecasts under MSE loss and nonlinear data generating process versus under linex loss and a restricted DGP with dynamics only in the conditional mean.

3.2.1 Mean squared error loss and arbitrary data generating process

First suppose that the loss function is of the MSE type whereas we do not impose any restrictions on the DGP. We collect the results in the following proposition.

Proposition 5 *Let the loss function be:*

$$L\left(Y_{t+h}-\hat{Y}_{t+h,t}\right)=\left(Y_{t+h}-\hat{Y}_{t+h,t}\right)^{2}$$

and let the (possibly nonlinear) process Y_t be stationary. Then the following are true:

1. *The optimal forecast of Y_{t+h} is $E_t\left[Y_{t+h}\right]$ for all forecast horizons h,*
2. *The optimal forecast error is conditionally (and unconditionally) unbiased,*
3. *The unconditional variance of the optimal forecast error is non-decreasing as a function of the forecast horizon, and*
4. *The optimal h-step forecast error exhibits zero serial correlation beyond the $(h-1)$ th lag.*

The above proposition shows that the standard properties of optimal forecasts are generated by the assumption of mean squared error loss alone; assumptions on the DGP (beyond stationarity) are not required.

For completeness, we verify the above general results for the regime switching process previously considered. The verification of the first two properties is simple. The third property is verified by:

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

which is constant for all horizons.

The autocovariance properties of the optimal forecast errors under the regime switching process are given by:

$$\begin{aligned} Cov\left(e^*_{t+h,t}, e^*_{t+h-j,t-j}\right) &= E\left[\sigma_{s_{t+h-j}}\sigma_{s_{t+h}}\nu_{t+h-j}\nu_{t+h}\right] \\ &= \sum_{s_{t+h-j=1}}^{k} \sum_{s_{t+h}=1}^{k} \bar{\boldsymbol{\pi}}_{\left(s_{t+h-j}\right)}\pi_{s_{h+t}|t+h-j}\sigma_{s_{t+h-j}}\sigma_{s_{t+h}} \cdot \\ &\quad E\left[\nu_{t+h-j}\nu_{t+h}|S_{t+h-j}=s_{t+h-j}, S_{t+h}=s_{t+h}\right] \\ &= 0 \text{ for } j \neq 0. \end{aligned}$$

Thus the optimal forecast errors are conditionally and unconditionally unbiased, have constant unconditional variance as a function of the forecast horizon, and are serially uncorrelated at all lags.

4 Asymmetric loss and DGPs with dynamics only in the conditional mean

In this section we consider the combination of asymmetric loss functions with a restricted class of DGPs; namely those with dynamics in the conditional mean but no dynamics in the remainder of the conditional distribution. This class of DGPs is still quite broad, and includes ARMA processes and non-linear regressions. Such a random variable may be written as:

$$\begin{aligned} Y_{t+h} &= E\left[Y_{t+h}|\Omega_t\right] + \varepsilon_{t+h}, \text{ where } \varepsilon_{t+h}|\Omega_t \sim D_h \text{ and} \\ E\left[Y_{t+h}|\Omega_t\right] &= g\left(Z_t\right) \end{aligned}$$

where g is some function of $Z_t \in \Omega_t$. The restriction of dynamics only in the conditional mean implies that the innovation term, ε_{t+h}, is drawn from some distribution, D_h, which will generally depend on the forecast horizon, but is *independent* of Ω_t and so is not denoted with a subscript t. Note that this restriction implies that

$$E\left[\phi\left(\varepsilon_{t+h}\right)\cdot Z_t\right]=E\left[\phi\left(\varepsilon_{t+h}\right)\right]E\left[Z_t\right]$$

for all functions ϕ and any vector of elements $Z_t\in\Omega_t$, and that $E_t\left[\varepsilon_{t+h}\right]=0$.

The types of loss functions we consider here are those that depend only upon the forecast error, i.e., $L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)=L\left(Y_{t+h}-\hat{Y}_{t+h,t}\right)=L\left(e_{t+h,t}\right)$. Many common loss functions are of this form, for example lin-lin, quad-quad and linex. However this restriction does rule out certain loss functions, for example those that focus on proportional errors, such as $L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)=\left(Y_{t+h}\cdot\hat{Y}_{t+h,t}^{-1}-1\right)^2$.

As an example, consider the MA data generating process in equation (3), with Gaussian residuals, and a linex loss function. The first order condition for the optimal forecast is

$$E_t\left[\exp\left\{a\left(Y_{t+h}-\hat{Y}^*_{t+h,t}\right)\right\}\right]=1,$$

so that (from (6))

$$\exp\left\{\frac{a^2\sigma^2}{2}\left(\sum_{i=0}^{h-1}\theta_i^2\right)+a\left(\sum_{i=0}^{\infty}\theta_{h+i}\varepsilon_{t-i}\right)-a\hat{Y}^*_{t+h,t}\right\}=1$$

and the optimal forecast is given by

$$\hat{Y}^*_{t+h,t}=\sum_{i=0}^{\infty}\theta_{h+i}\varepsilon_{t-i}+\frac{a\sigma^2}{2}\sum_{i=0}^{h-1}\theta_i^2$$

This is consistent with the result of Granger (1969) and Christoffersen and Diebold (1997), who show that for this combination of loss function and DGP the optimal forecast is of the form:

$$\hat{Y}^*_{t+h,t}=E_t\left[Y_{t+h}\right]+\alpha_h,$$

where α_h is a bias term that depends only on the loss function and the forecast horizon. If the conditional distribution of $Y_{t+h}|\Omega_t$ has dynamics beyond those in the conditional mean, Christoffersen and Diebold (1997) show that the bias term will depend not only on the forecast horizon and the loss function, but also on the higher-order dynamics. This would correspond to a violation of our assumption that D_h is independent of Ω_t.

In the case without higher-order dynamics we obtain the following serial correlation properties of the optimal forecast error.

Proposition 6 *Let Y be any stationary process such that*

$$Y_{t+h} = E_t\left[Y_{t+h}\right] + \varepsilon_{t+h},\ \ \varepsilon_{t+h}|\Omega_t \sim D_h$$

Then $Cov\left(e^*_{t+h,t}, e^*_{t+h-j,t-j}\right) = 0$ *for all* $j \geq h$ *and any* h, *for all loss functions that are dependent only upon the forecast error.*

The above proposition shows that under a somewhat restrictive assumption on the DGP, and only one weak assumption on the loss function, the optimal forecast errors are serially uncorrelated at lags greater than or equal to the forecast horizon, for *any* loss function. This implies that given a sequence of realizations and forecasts, $\left\{\left(Y_{t+h}, \hat{Y}_{t+h,t}\right)\right\}_{t=1}^{T}$, we may test for forecast optimality *without knowledge of the forecaster's loss function* by testing the serial correlation properties of the forecast errors. For financial applications the assumption of constant higher-order conditional moments may be too strong, but in macroeconomic applications the assumption that all dynamics are driven by the conditional mean may be palatable. In this case, tests of forecast optimality need *not* rely on the assumption of MSE loss, as in the papers listed in footnote 1, or on the assumption that the loss function is known up to an unknown parameter vector and that the forecast model is linear, as in Elliott, *et al.* (2002). Instead forecast optimality can be tested with a large degree of robustness to the loss function of the forecaster.

In the linex-ARMA example discussed above, the optimal forecast error is

$$e^*_{t+h,t} = \sum_{i=0}^{h-1} \theta_i \varepsilon_{t+h-i} - \frac{a\sigma^2}{2}\sum_{i=0}^{h-1}\theta_i^2.$$

In this case the h-period forecast error is an $MA(h-1)$ process and the variance of the forecast error is

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

which is non-decreasing in h. The conditions assumed in this section are also sufficient to yield results on the behaviour of the variance of the optimal forecast error as a function of h, as shown below.

Proposition 7 *Let Y be any stationary process such that*

$$Y_{t+h} = E_t\left[Y_{t+h}\right] + \varepsilon_{t+h},\ \ \varepsilon_{t+h}|\Omega_t \sim D_h$$

Then $V\left[e^{*}_{t+h,t}\right]$ *is a weakly increasing function of* h *for all loss functions that are dependent only upon the forecast error.*

Like Proposition 6, the above proposition may be used to test forecast optimality in the absence of information on the forecaster's loss function, under the assumption of mean-only dynamics in the variable of interest. Given a time series of forecasts with a range of horizons, Proposition 7 suggests testing that the variance of the forecast error is weakly increasing with the forecast horizon.

Overall, the results presented in Sections 3.2.1 and 4 demonstrate that it is the *combination* of asymmetric loss and dynamics in the conditional distribution beyond those in the conditional mean that generate the violations reported in Section 3. Under MSE loss and an arbitrary DGP we showed that the standard properties hold. Under a weak assumption on the loss function and the restriction that the conditional density has no dynamics beyond the conditional mean we showed that while the optimal forecast is biased, the optimal forecast errors are serially uncorrelated for lags greater than $(h-1)$ and the unconditional forecast error variance is weakly increasing in h.

5 Properties of Optimal Forecasts under General Conditions

While quadratic loss is commonly used in empirical work, in a more general setting the optimal forecast, $\hat{Y}^{*}_{t+h,t}$, is chosen to minimize the expected loss, where the loss function need not be a function solely of the forecast error:

$$L = L\left(Y_{t+h}, \hat{Y}_{t+h,t}\right)$$

We will make the following assumptions about the loss function and the data generating process for Y_{t+h} :

Assumption 4: The function L is analytic except at a finite number of points.

Assumption 5: The expected loss, $E_t\left[L\left(Y_{t+h}, \hat{Y}_{t+h,t}\right)\right]$, is finite for all values of $\hat{Y}_{t+h,t}$ and for all h.

Assumption 6: The expected marginal loss, $E_t\left[\partial L\left(Y_{t+h}, \hat{Y}_{t+h,t}\right)/\partial \hat{Y}_{t+h,t}\right]$, is finite for all but a finite number of values of $\hat{Y}_{t+h,t}$ and for all h.

As we are interested only in characterising the behaviour of the optimal forecast, without actually finding the optimal forecast, we do not need to assume that the expected loss has a unique minimum, or a unique minimum in a region around some value.

The optimal forecast in general cases is defined as:

$$\hat{Y}^*_{t+h,t} \equiv \underset{\hat{Y}_{t+h,t}}{\arg\min} E_t\left[L\left(Y_{t+h}, \hat{Y}_{t+h,t}\right)\right] \tag{17}$$

$$= \underset{\hat{Y}_{t+h,t}}{\arg\min} \int L\left(y, \hat{Y}_{t+h,t}\right) f_{t+h,t}(y)\, dy \tag{18}$$

where $Y_{t+h}|\Omega_t$ has density $f_{t+h,t}$.

Under assumption 6 the first order condition becomes[11]

$$\begin{aligned} 0 &= \frac{\partial E_t\left[L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)\right]}{\partial \hat{Y}_{t+h,t}} \\ &= E_t\left[\frac{\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}}\right] \\ &= \int \frac{\partial L\left(y, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}} f_{t+h,t}(y)\, dy. \end{aligned} \tag{19}$$

This condition can be rewritten using what Granger (1999) refers to as the (optimal) generalized forecast error[12], $\psi^*_{t+h,t}$,[13]

$$\psi^*_{t+h,t} \equiv \frac{\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}} \tag{20}$$

so that (19) simplifies to

$$E_t[\psi^*_{t+h,t}] = \int \psi^*_{t+h,t} f_{t+h,t}(y)\, dy = 0 \tag{21}$$

Under a broad set of conditions $\psi^*_{t+h,t}$ is therefore a martingale difference sequence with respect to the information set used to compute the forecast, Ω_t.[14]

Often $\psi^*_{t+h,t}$ can be derived explicitly. For the regime switching process/linex loss example the generalized forecast error is:

[11] As the bounds on the integral are defined by the conditional density of Y_{t+h} given Ω_t, they are unaffected by the choice of $\hat{Y}_{t+h,t}$ and so two of the terms in Leibnitz's rule (see Casella and Berger, 1990, for example) drop out.

[12] Granger (1999) only considers loss functions that have the forecast error as an argument, and so defines the generalised forecast error as $\psi^*_{t+h,t} \equiv \partial L\left(e_{t+h,t}\right)/\partial e_{t+h,t}$. Our definition is slightly more general, and in our case the generalised forecast error is the negative of the generalised forecast error in Granger's (1999) case.

[13] While this term is appropriate under prediction-error loss, more generally $\psi^*_{t+h,t}$ can be viewed as the marginal loss associated with a particular prediction, $\hat{Y}_{t+h,t}$.

[14] Notice that we are not simply considering linear projections on information in Ω_t. Only if Y_{t+h} and the variables relevant for forecasting it, $\mathbf{X}_t$, are jointly Gaussian will the two be identical.

$$\psi^*_{t+h,t} = a - a\exp\left\{a\sigma_{s_{t+h}}\nu_{t+h} - \log\left(\hat{\boldsymbol{\pi}}'_{s_t,t}\mathbf{P}^h\boldsymbol{\varphi}\right)\right\}. \tag{22}$$

Under MSE loss, the h-step generalized forecast error is:

$$\begin{aligned} \psi^*_{t+h,t} &= -2\left(Y_{t+h} - \hat{Y}^*_{t+h,t}\right) \\ &= -2e^*_{t+h,t}, \end{aligned} \tag{23}$$

and so the generalized forecast error is simply the negative of twice the standard forecast error. It turns out that the close relation of the standard forecast error and the generalized forecast error in the case of mean squared error loss is the reason for the standard forecast error having such nice properties in that case. As we showed in the previous section, the properties of the standard forecast error do not hold for asymmetric loss and nonlinear processes; they do, however, hold for the generalized forecast error. We now turn our attention to proving properties of the generalized forecast error analogous to those for the standard case.

5.1 Unbiasedness of the generalized forecast error

It is easy to establish that, although the forecast error, $e^*_{t+h,t}$, need not be unbiased, the generalized forecast error, $\psi^*_{t+h,t}$, is unbiased:

Proposition 8 *The generalized forecast error has conditional (and unconditional) mean zero.*

For the regime switching process the conditional mean of the generalized forecast error is

$$\begin{aligned} E_t\left[\psi^*_{t+h,t}\right] &= a - a\left(\hat{\boldsymbol{\pi}}'_{s_t,t}\mathbf{P}^h\boldsymbol{\varphi}\right)^{-1} E_t\left[\exp\left\{a\sigma_{s_{t+h}}\nu_{t+h}\right\}\right] \\ &= a - a\left(\hat{\boldsymbol{\pi}}'_{s_t,t}\mathbf{P}^h\boldsymbol{\varphi}\right)^{-1} \hat{\boldsymbol{\pi}}'_{s_t,t}\mathbf{P}^h \exp\left\{a\boldsymbol{\sigma}^2\right\} \\ &= 0 \end{aligned}$$

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

5.2 Non-decreasing expected loss as a function of the forecast horizon

In the standard framework the optimal forecast is unbiased and the loss function is quadratic. This leads to the equality of the optimal forecast error variance and the expected loss from the optimal forecast:

$$E\left[L\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right]=E\left[e^{*2}_{t+h,t}\right]=Var\left(e^*_{t+h,t}\right). \tag{24}$$

In general this equality will not hold, and indeed the optimal forecast error variance is not necessarily of interest; rather, the quantity of interest is the expected loss from the forecast. For the regime switching process we showed that the variance of the optimal forecast error need not be non-decreasing with the forecast horizon, contrary to results in the standard framework. The reason for this is a mis-match of the forecaster's loss/objective function and variance. Sentana (1998), *inter alia,* also discusses the problem of mis-matched objective functions. Under general loss functions, if we instead look at the unconditional expected loss as a function of the forecast horizon we obtain the following result:

Proposition 9 *Under strict stationarity of Y_t, the unconditional expected loss of an optimal forecast error is a non-decreasing function of the forecast horizon. The conditional expected loss, however, need not be a non-decreasing function of the forecast horizon.*

The unconditional expected loss as a function of the forecast horizon behaves as follows in the regime switching example.

Corollary 10 *The unconditional expected loss in the regime switching example is*

$$E\left[L\left(Y_{t+h},\hat{Y}^*_{t+h,t};a\right)\right]=\bar{\boldsymbol{\pi}}'\log\left(\mathbf{P}^h\boldsymbol{\varphi}\right)$$

and

$$E\left[L\left(Y_{t+h},\hat{Y}^*_{t+h,t};a\right)\right]\rightarrow\log\left(\bar{\boldsymbol{\pi}}'\boldsymbol{\varphi}\right)\ \ \textit{as } h\rightarrow\infty$$

In the numerical example used above, the expected loss as a function of the forecast horizon is:

[INSERT FIGURE 5 HERE]

5.3 Serial correlation in the generalized forecast error

A property of optimal h-step ahead forecast errors under MSE loss is that they are MA processes or order no greater than $h-1$. In a non-linear, non-Gaussian framework an MA process need not completely describe the dependence properties of the generalized forecast error, however the autocorrelation function of the generalized forecast error will match some $MA\left(h-1\right)$ process.

Proposition 11 *The generalized forecast error from an optimal h-step ahead forecast made at time t exhibits zero correlation with any function of any element of the time t information set, Ω_t. In particular, the generalized forecast error will exhibit zero serial correlation for lags greater than $(h-1)$.*

For completeness, we derive the autocorrelation function for the optimal generalized forecast error for our regime switching example.

Corollary 12 *The generalized forecast error from an optimal h-step ahead forecast made at time t in the regime switching example has the following autocovariance function:*

$$Cov\left[\psi^*_{t+h,t}, \psi^*_{t+h-j,t-j}\right] = \begin{cases} V\left[\psi^*_{t+h,t}\right] = -a^2 + a^2 \sum_{s_t=1}^{k} \bar{\pi}_{(s_t)} \left(\pi'_{s_t,t} P^h \varphi\right)^{-2} \left(\pi'_{s_t,t} P^h \varphi^4\right) & j=0 \\ \quad -a^2 + a^2 \sum_{s_{t-j}=1}^{k} \bar{\pi}_{(s_{t-j})} \left(\pi'_{s_{t-j},t-j} P^h \varphi\right)^{-1} \cdot & \\ \sum_{s_t=1}^{k} \pi_{s_t,t-j} \left(\pi'_{s_t,t} P^h \varphi\right)^{-1} \cdot \left(\varphi' \odot \left(\pi'_{s_t,t} P^{h-j}\right)\right) P^j \varphi & 0<j<h \\ 0 & j \geq h \end{cases}$$

where $\varphi^4 \equiv \exp\left\{2a^2\sigma^2_{s_{t+h}}\right\}$.

Using the numerical example above, we present the autocorrelation function for the generalized optimal forecast error in Figure 6.

[INSERT FIGURE 6 HERE]

5.4 Properties of the optimal forecast error under a change of measure

In previous sections we showed that by changing our object of analysis from the usual forecast error to the 'generalised' forecast error we were able to obtain the usual properties of unbiasedness and zero serial correlation. In this section we instead consider changing the probability measure used to compute the properties of the forecast error. This analysis is akin to the use of risk-neutral densities in asset pricing, see Cochrane (2001) for example. In asset pricing one may scale the objective, or physical, probabilities by the stochastic discount factor, or the discounted ratio of marginal utilities, to obtain a risk-neutral probability measure, and then apply risk-neutral pricing methods. Here we will scale the objective probability measure by the ratio of the marginal loss, $\partial L/\partial \hat{y}$, to the forecast error, and then show that under the new probability measure, which we call the "MSE-loss probability measure", the standard properties hold. The following results thus suggest an alternative means of evaluating forecasts made using asymmetric loss functions.

5.4.1 Unbiasedness under a change of measure

Suppose that $\frac{\partial L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)}{\partial \hat{Y}} > 0$ if $Y_{t+h} > \hat{Y}_{t+h,t}$ and $\frac{\partial L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)}{\partial \hat{Y}} < 0$ if $Y_{t+h} < \hat{Y}_{t+h,t}$, and notice that the conditional distribution of the forecast error, $f_{e_{t+h,t}}$, given Ω_t and a forecast $\hat{Y}_{t+h,t}$, satisfies:

$$f_{e_{t+h.t}}\left(e;\hat{Y}_{t+h,t}\right) = f_{t+h,t}\left(\hat{Y}_{t+h,t}+e\right) \text{ for all } \left(e,\hat{Y}_{t+h,t}\right) \in R^2 \tag{25}$$

where $f_{t+h,t}$ is the conditional distribution of Y_{t+h} given Ω_t.

Definition 13 *Assume that*

$$\left| E_t\left[\frac{1}{e_{t+h,t}}\frac{\partial L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)}{\partial \hat{Y}}\right]\right| < \infty \text{ for all } t,h \text{ and } \hat{Y}_{t+h,t}.$$

Then let the univariate "MSE-loss probability measure", $f^*_{e_{t+h,t}}$, be defined by

$$f^*_{e_{t+h,t}}\left(e;\hat{Y}_{t+h,t}\right) = \frac{\frac{1}{e}\cdot\left.\frac{\partial L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)}{\partial \hat{Y}}\right|_{Y_{t+h}=\hat{Y}_{t+h,t}+e}\cdot f_{e_{t+h,t}}\left(e;\hat{Y}_{t+h,t}\right)}{E_t\left[\frac{1}{Y_{t+h}-\hat{Y}_{t+h,t}}\frac{\partial L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)}{\partial \hat{Y}}\right]} \tag{26}$$

$$\equiv \frac{\left(\psi_{t+h,t}\left(\hat{Y}_{t+h,t}\right)/e\right)\cdot f_{e_{t+h,t}}\left(e;\hat{Y}_{t+h,t}\right)}{E_t\left[\frac{1}{Y_{t+h}-\hat{Y}_{t+h,t}}\frac{\partial L\left(Y_{t+h},\hat{Y}_{t+h,t}\right)}{\partial \hat{Y}}\right]} \tag{27}$$

Proposition 14 *The univariate "MSE-loss probability measure", $f^*_{e_{t+h,t}}$, defined above is a proper probability density function.*

Note that by construction the MSE-loss probability measure f^* is absolutely continuous with respect to the usual probability measure, f, (that is, $f^{**} << f$). See White (1994) for a definition of absolute continuity.

Proposition 15 *The optimal forecast error, $e^*_{t+h,t} = Y_{t+h} - \hat{Y}^*_{t+h,t}$ has conditional (and unconditional) mean zero under the MSE-loss probability measure.*

5.4.2 Zero serial correlation under a change of measure

We can further show that under the MSE-loss probability measure the optimal h-step ahead forecast errors exhibit zero serial correlation for all lags greater than $h-1$. In the proof of the following proposition we make reference to the bivariate MSE-loss probability measure, but do not need to explicitly define it in order to obtain the result.

Proposition 16 *The optimal forecast error,* $e^*_{t+h,t} = Y_{t+h} - \hat{Y}^*_{t+h,t}$ *has zero serial unconditional correlation under the MSE-loss probability measure for all lags greater than* $h-1$*.*

5.5 Forecast error variance and expected loss for elliptically distributed random variables

Under construction...

6 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. We demonstrated analytically how they are violated under more general assumptions about the loss function, extending the 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 data generating process may be nonlinear but strictly stationary.

We introduced a change of measure, analogous to the change of measure from objective to risk-neutral commonly employed in asset pricing. Under the new probability measure, which we call the "MSE-loss probability measure", the optimal h-step forecast error for any general loss function has zero conditional mean and zero serial correlation for all lags greater than $h-1$, ie, the same properties as an optimal forecast under MSE loss. This is a novel line of analysis, and one that may lead to new ways of testing forecast optimality.

We have deliberately constrained our analysis in this paper to ignore parameter estimation uncertainty. Our results are all the stronger since we have shown that simply changing the loss function and allowing for nonlinear dynamics can imply that all the standard properties an optimal forecast is usually thought to possess no longer remain valid. Parameter estimation uncertainty is another source that could lead to rejections of tests of forecast optimality in practice, see Hoque, *et al.* (1988) and Magnus and Pesaran (1987, 1989).

Our analysis does not imply that 'anything goes' and that forecast rationality is not testable. Rather, it suggests that researchers have to use economic arguments to establish the underlying loss function as suggested in a recent paper by Elliott, Komunjer and Timmermann (2002) or, alternatively, try to conduct tests that are robust to the shape of the loss function by exploiting (testable) restrictions on the dynamics of the data generating process. Two such results were presented in section 4 of this paper; the first on the autocorrelation structure of optimal forecast errors, and the second on the variance of optimal forecast errors as a function of the forecast horizon. Deriving testable implications of forecast optimality with limited knowledge of the DGP and the forecaster's loss function is an interesting area for future research.

References

[1] Abarbanell, J. S. and V. L. Bernard, 1992, Tests of Analysts' Overreaction/Underreaction to Earnings Information as an Explanation for Anomalous Stock Price Behaviour. Journal of Finance, 46(3), 1181 - 1207.

[2] Bollerslev, T., 1986, Generalized Autoregressive Conditional Heteroskedasticity. Journal of Econometrics 31, 307-327.

[3] Brown, B.Y. and S. Maital, 1981, What do economists know? An empirical study of experts' expectations. Econometrica 49, 491-504.

[4] Cargill, T.F. and R.A. Meyer, 1980, The Term Structure of Inflationary Expectations and Market Efficiency. Journal of Finance 35, 57-70.

[5] Casella, G. and R. L. Berger, 1990, Statistical Inference. Duxbury Press.

[6] Christoffersen, P. and K. Jacobs, 2002, The Importance of the Loss Function in Option Valuation. Working paper, Faculty of Management, McGill University.

[7] Christoffersen, P.F. and F.X. Diebold, 1997, Optimal prediction under asymmetric loss. Econometric Theory 13, 808-817.

[8] Clements, M.P. and D.F. Hendry, 1998, Forecasting Economic Time Series, Cambridge University Press.

[9] Cochrange, John H., 2001, *Asset Pricing*, Princeton University Press, USA.

[10] Corradi, V. and N. R. Swanson, 2002, A Consistent Test for Nonlinear Out of Sample Predictive Accuracy. Journal of Econometrics, 110, 353-381.

[11] De Bondt, W. F. M. and R. H. Thaler, 1990. Do Security Analysts Overreact? American Economic Review, 80(2), 52-57.

[12] De Bondt, W.F.M. and M.M. Bange, 1992, Inflation Forecast Errors and Time Variation in Term Premia. Journal of Financial and Quantitative Analysis 27, 479-496.

[13] Diebold, F.X., 2001, Elements of Forecasting (2nd edition). Southwestern.

[14] Diebold, F.X. and J.A. Lopez, 1996, Forecast Evaluation and Combination. Ch. 8 in G.S. Maddala and C.R. Rao, eds., Handbook of Statistics, Vol. 14.

[15] Dokko, Y. and R. H. Edelstein, 1989, How Well do Economists Forecast Stock Market Prices? A study of the Livingston Surveys. American Economic Review 79, 865-871.

[16] Drost, Feike C., and Nijman, Theo E., 1993, Temporal aggregation of GARCH processes, Econometrica, 61, 909-928.

[17] Elliott, G., I. Komunjer, and A. Timmermann, 2002, Estimating Loss Function Parameters, working paper, Department of Economics, University of California, San Diego.

[18] Engle, R.F., 1982, Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica 50, 987-1007.

[19] Figlewski, S. and P. Wachtel, 1981, The Formation of Inflationary Expectations. Review of Economics and Statistics 63, 1-10.

[20] Granger, C.W.J., 1969, Prediction with a generalized cost function," OR, 20, 199-207.

[21] Granger, C.W.J., 1999, Outline of Forecast Theory Using Generalized Cost Functions. Spanish Economic Review 1, 161-173.

[22] Granger, C.W.J., and P. Newbold, 1986, Forecasting Economic Time Series, Second Edition. Academic Press.

[23] Granger, C.W.J. and M.H. Pesaran, 2000, Economic and Statistical Measures of Forecast Accuracy. Journal of Forecasting 19, 537-560.

[24] Granger, C.W.J. and T. Terasvirta, 1993, Modelling Nonlinear Economic Relationships. Oxford University Press.

[25] Hamilton, J.D., 1989, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57, 357-86.

[26] Hoque, Asraul, Magnus, Jan R., and Pesaran, Bahram, 1988, The Exact Multi-Period Mean-Square Forecast Error for the First-Order Autoregressive Model, *Journal of Econometrics, 39, 327-346.*

[27] Ingersoll, Jonathan E., 1987, *Theory of Financial Decision Making*, Rowman and Littlefield, USA.

[28] Keane, M.P. and D.E. Runkle, 1990, Testing the Rationality of Price Forecasts: New Evidence from Panel Data. American Economic Review 80, 714-735.

[29] Keane, M.P. and D.E. Runkle, 1998, Are Financial Analysts' Forecasts of Corporate Profits Rational? Journal of Political Economy 106(4), 768-805.

[30] Lakonishok, J., 1980, Stock Market Return Expectations: Some General Properties. Journal of Finance 35, 921-931.

[31] Magnus, Jan R., and Pesaran, Bahram, 1989, The Exact Multi-Peiod Mean-Square Forecast Error for the First-Order Autoregressive Model with an Intercept, *Journal of Econometrics, 42, 157-179.*

[32] Magnus, Jan R., and Pesaran, Bahram, 1991, The Bias of Forecasts from a First-Order Autoregression, *Econometric Theory, 7, 222-235.*

[33] Mishkin, F.S., 1981, Are Markets Forecasts Rational? American Economic Review 71, 295-306.

[34] Muth, J. F., 1961, Rational Expectations and the Theory of Price Movements. Econometrica 29(3), 315-335.

[35] Pesando, J.E., 1975, A Note on the Rationality of the Livingston Price Expectations. Journal of Political Economy 83, 849-858.

[36] Pesaran, M.H. and Skouras, S., 2001, Decision-based Methods for Forecast Evaluation. In Clements, M.P. and D.F. Hendry (eds.) Companion to Economic Forecasting. Basil Blackwell.

[37] Schroeter, J.R. and S.L. Smith, 1986, A Reexamination of the Livingston Price Expectations. Journal of Money, Credit and Banking 18, 239-246.

[38] Sentana, Enrique, 1998, Least Squares Predictions and Mean-Variance Analysis, CEMFI Working Paper 9711, Centre for Monetary and Financial Studies.

[39] Skouras, S., 2001, Decisionmetrics: A Decision-based Approach to Econometric Modeling. Mimeo, Santa Fe Institute.

[40] Varian, H. R., 1974, A Bayesian Approach to Real Estate Assessment. In Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage, eds. S.E. Fienberg and A. Zellner, Amsterdam: North Holland, 195-208.

[41] West, K.D., H.J. Edison and D. Cho, 1993, A Utility-based Comparison of Some Models of Exchange Rate Volatility. Journal of International Economics 35, 23-46.

[42] White, Halbert, 1994, *Estimation, Inference and Specification Analysis*, Econometric Society Monographs 22, Cambridge University Press, New York.

[43] Zarnowitz, V., 1985, Rational Expectations and Macroeconomic Forecasts. Journal of Business and Economic Statistics 3, 293-311.

[44] Zellner, A., 1986, Bayesian Estimation and Prediction Using Asymmetric Loss Functions. Journal of the American Statistical Association, 81, 446-451.

Appendix

Proof of Proposition 1. The h-step-ahead forecast error has a conditional expectation of

$$E_t\left[e^*_{t+h,t}\right] = -\frac{1}{a}\log\left(\hat{\boldsymbol{\pi}}'_{s_t,t}\mathbf{P}^h\boldsymbol{\varphi}\right)$$

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

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

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

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

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

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

Proof of Proposition 2. From Proposition 1 we have

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

Here $\bar{\pi}_{(i)}$ is the i^{th} element of the vector $\bar{\boldsymbol{\pi}}$, 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 $h \to \infty$, notice that

$$\boldsymbol{\lambda}_\infty \equiv \lim_{h\to\infty}\boldsymbol{\lambda}_h = \boldsymbol{\iota}\log\left(\bar{\boldsymbol{\pi}}'\boldsymbol{\varphi}\right).$$

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

$$\boldsymbol{\iota}'\left(\left(\bar{\boldsymbol{\pi}}\boldsymbol{\iota}'\right)\odot I - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'\right)\boldsymbol{\iota} = \boldsymbol{\iota}'\left(\left(\bar{\boldsymbol{\pi}}\boldsymbol{\iota}'\right)\odot I\right)\boldsymbol{\iota} - \boldsymbol{\iota}'\bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'\boldsymbol{\iota} = \bar{\boldsymbol{\pi}}'\boldsymbol{\iota} - \left(\bar{\boldsymbol{\pi}}'\boldsymbol{\iota}\right)'\left(\bar{\boldsymbol{\pi}}'\boldsymbol{\iota}\right) = 0.$$

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

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

■

Proof of Corollary 3. Follows directly from the proof of Proposition 2. ■

Proof of Proposition 4. The autocovariance function for an h-step forecast is:

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

For fixed h, as $j \to \infty$, $Cov\left[e^*_{t+h,t}, e^*_{t+h-j,t-j}\right] \to \frac{1}{a^2}\boldsymbol{\lambda}'_h\left(\left(\bar{\boldsymbol{\pi}}\boldsymbol{\iota}'\odot\boldsymbol{\iota}\bar{\boldsymbol{\pi}}'\right) - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'\right)\boldsymbol{\lambda}_h = 0$. ■

Proof of Proposition 5. The first order condition implies that

$$\begin{aligned}
L\left(Y_{t+h},\hat{Y}_{t+h}\right) &\equiv \left(Y_{t+h}-\hat{Y}_{t+h}\right)^2 \\
\frac{\partial E_t\left[L\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right]}{\partial \hat{Y}_{t+h,t}} &= -2\left(E_t\left[Y_{t+h}\right]-\hat{Y}^*_{t+h,t}\right)=0, \text{ so} \\
\hat{Y}^*_{t+h,t} &= E_t\left[Y_{t+h}\right], \text{ and} \\
e^*_{t+h,t} &= Y_{t+h}-E_t\left[Y_{t+h}\right]
\end{aligned}$$

Thus the optimal forecast under MSE is conditionally and unconditionally unbiased for all forecast horizons, for all DGPs.

The remainder of the proof follows directly from the proofs of Propositions 9 and 11, presented below, when one observes the relation between the forecast error and the generalized forecast error (defined in Section 5), $\psi^*_{t+h,t}$, for the mean squared loss case: $e^*_{t+h,t}=-\frac{1}{2}\psi^*_{t+h,t}$. ■

Proof of Proposition 6. Under the conditions given Christoffersen and Diebold (1997) show that the optimal forecast may be written as

$$\hat{Y}^*_{t+h,t}=E_t\left[Y_{t+h}\right]+\alpha_h$$

and so the optimal forecast error is $e^*_{t+h,t}=Y_{t+h}-\hat{Y}^*_{t+h,t}=\varepsilon_{t+h}-\alpha_h$. Since α_h is constant for fixed h,

$$\begin{aligned}
Cov\left[e^*_{t+h,t},e^*_{t+h-j,t-j}\right] &= Cov\left[\varepsilon_{t+h},\varepsilon_{t+h-j}\right] \\
&= E\left[\varepsilon_{t+h}\cdot\varepsilon_{t+h-j}\right] \\
&= E\left[E_t\left[\varepsilon_{t+h}\right]\cdot\varepsilon_{t+h-j}\right] \text{ for } \forall j\geq h \\
&= 0
\end{aligned}$$

■

Proof of Proposition 7. Consider $h>0$ and $j>0$. Let

$$\begin{aligned}
Y_{t+h+j} &= E_t\left[Y_{t+h+j}\right]+\eta_{t+h+j},\ \eta_{t+h+j}|\Omega_t\sim D_{h+j} \\
Y_{t+h+j} &= E_{t+j}\left[Y_{t+h+j}\right]+\varepsilon_{t+h+j},\ \varepsilon_{t+h+j}|\Omega_{t+j}\sim D_h
\end{aligned}$$

From Christoffersen and Diebold (1997) we know that under the above assumptions $\hat{Y}^*_{t+h,t} = E_t\left[Y_{t+h}\right] + \alpha_h$, so

$$\begin{aligned} e^*_{t+h+j,t} &= \eta_{t+h+j} - \alpha_{h+j} \\ e^*_{t+h+j,t+j} &= \varepsilon_{t+h+j} - \alpha_h \end{aligned}$$

where α_h and α_{h+j} are constants. Thus $V_t\left[e^*_{t+h+j,t}\right] = V_t\left[\eta_{t+h+j}\right] \equiv \sigma^2_{h+j}$, and $V_t\left[e^*_{t+h+j,t+j}\right] \equiv \sigma^2_h$. Note also that $V\left[e^*_{t+h+j,t}\right] = E\left[E_t\left[\eta^2_{t+h+j}\right]\right] = \sigma^2_{h+j}$, and similarly $V\left[e^*_{t+h+j,t+j}\right] = \sigma^2_h$. Now we seek to show that $\sigma^2_{h+j} \geq \sigma^2_h$.

$$\begin{aligned} V\left[e^*_{t+h+j,t}\right] &= V_t\left[Y_{t+h+j} - E_t\left[Y_{t+h+j}\right]\right] \\ &= V_t\left[\varepsilon_{t+h+j} + \left(E_{t+j}\left[Y_{t+h+j}\right] - E_t\left[Y_{t+h+j}\right]\right)\right] \\ &= \sigma^2_h + V_t\left[E_{t+j}\left[Y_{t+h+j}\right]\right] + 2Cov_t\left[\varepsilon_{t+h+j}, E_{t+j}\left[Y_{t+h+j}\right] - E_t\left[Y_{t+h+j}\right]\right] \\ &\geq \sigma^2_h \\ &= V\left[e^*_{t+h,t}\right] \end{aligned}$$

where the first equality follows from the equality of the conditional and unconditional variance of the forecast error in this scenario; the third equality follows from the fact that $E_t\left[Y_{t+h+j}\right]$ is constant given Ω_t; the weak inequality follows from the non-negativity of $V_t\left[E_{t+1}\left[Y_{t+2}\right]\right]$ and that $E_{t+j}\left[\varepsilon_{t+h+j} \cdot \phi\left(Z^{t+j}\right)\right] = 0$; the final equality follows from the fact that D_h does not change with t. The cases where $h = 0$ and/or $j = 0$ are trivial. Thus $V\left[e^*_{t+h+j,t}\right] \geq V\left[e^*_{t+h,t}\right]$ $\forall\, h,\, j \geq 0$. ■

Proof of Proposition 8.

$$E_t\left[\psi^*_{t+h,t}\right] = E_t\left[\frac{\partial L\left(y_{t+h}, \hat{y}^*_{t+h,t}\right)}{\partial \hat{y}_{t+h,t}}\right] = 0,$$

by the first-order condition for the optimality of $\hat{Y}^*_{t+h,t}$, and $E\left[\psi^*_{t+h,t}\right] = 0$ by the law of iterated expectations. ■

Proof of Proposition 9. By strict stationarity of $\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)$ for all h and j we have

$$E\left[E_t\left[L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)\right]\right] = E\left[E_{t-j}\left[L\left(Y_{t+h-j}, \hat{Y}^*_{t+h-j,t-j}\right)\right]\right]$$

and so the unconditional expected loss only depends on the forecast horizon, and not on the period when the forecast was made.

By the optimality of the forecast $\hat{Y}^*_{t+h,t}$ we also have, for $\forall j \geq 0$,

$$\begin{aligned} E_t\left[L\left(Y_{t+h}, \hat{Y}^*_{t+h,t-j}\right)\right] &\geq E_t\left[L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)\right] \\ E\left[L\left(Y_{t+h}, \hat{Y}^*_{t+h,t-j}\right)\right] &\geq E\left[L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)\right] \\ E\left[L\left(Y_{t+h+j}, \hat{Y}^*_{t+h+j,t}\right)\right] &\geq E\left[L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)\right] \end{aligned}$$

where the second line follows using the law of iterated expectations and the third line follows from strict stationarity. Hence the unconditional expected loss is a non-decreasing function of the forecast horizon.

To show that the conditional expected loss may be an increasing or a decreasing function of the forecast horizon we need only construct an example. We will use the 2-state regime switching/linex loss example from Section 3. Assume that $\hat{\boldsymbol{\pi}}_{s_t,t} = [0.95, 0.05]'$. Then from equations (11) and (12) we know that optimum forecasts and resulting conditional expected losses are: $\hat{Y}^*_{t+1,t} = 0.5376$, $\hat{Y}^*_{t+2,t} = 0.6616$, $E_t\left[L\left(Y_{t+1}, \hat{Y}^*_{t+1,t}\right)\right] = 3.1685$ and $E_t\left[L\left(Y_{t+2}, \hat{Y}^*_{t+2,t}\right)\right] = 3.7390$. If, on the other hand, $\hat{\boldsymbol{\pi}}_{s_t,t} = [0.05, 0.95]'$ then the optimal forecasts and resulting conditional expected losses are: $\hat{Y}^*_{t+1,t} = 1.8714$, $\hat{Y}^*_{t+2,t} = 1.7927$, $E_t\left[L\left(Y_{t+1}, \hat{Y}^*_{t+1,t}\right)\right] = 8.1050$ and $E_t\left[L\left(Y_{t+2}, \hat{Y}^*_{t+2,t}\right)\right] = 7.9995$. So if we start from a point where there is a high probability of being in the low volatility state, then the conditional expected loss is increasing with h. But if we start from a point where there is a high probability of being in the high volatility state, then the conditional expected loss is decreasing with h. ■

Proof of Corollary 10. Follows using similar steps as in the proofs of Propositions 2 and 4. Available from authors upon request. ■

Proof of Proposition 11. Since $\sigma\left(Y_t, Y_{t-1}, ...\right) \subseteq \Omega_t$ by assumption we know that $\psi^*_{t+h-j,t-j} = \partial L\left(Y_{t+h-j}, \hat{Y}^*_{t+h-j,t-j}\right)/\partial\hat{y}$ is an element of Ω_t for all $j \geq h$. From the first-order condition for the optimality of $\hat{Y}^*_{t+h,t}$ we have:

$$E\left[\psi^*_{t+h,t}|\Omega_t\right] = E\left[\left.\frac{\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}}\right|\Omega_t\right] = 0,$$

which implies $E\left[\psi^*_{t+h,t} \cdot \gamma\left(X_t\right)\right] = 0$ for all $X_t \in \Omega_t$ and all functions γ. Thus $\psi^*_{t+h,t}$ is uncorrelated with any function of any element of Ω_t. This implies that

$$E\left[\psi^*_{t+h,t} \cdot \psi^*_{t+h-j,t-j}\right] = 0 \text{ for all } j \geq h$$

and so $\psi^*_{t+h,t}$ is uncorrelated with $\psi^*_{t+h-j,t-j}$. ■

Proof of Corollary 12. Follows using similar steps as in the proofs of Propositions 2 and 4. Available from authors upon request. ■

Proof of Proposition 14. We need to show that $f^*_{e_{t+h}} \geq 0$ for all possible values of e, and that $\int f^*_{e_{t+h,t}}\left(u; \hat{Y}_{t+h,t}\right) du = 1$. By the assumption that $\frac{\partial L(Y_{t+h}, \hat{Y}_{t+h,t})}{\partial \hat{Y}} > 0$ if $Y_{t+h} > \hat{Y}_{t+h,t}$ and $\frac{\partial L(Y_{t+h}, \hat{Y}_{t+h,t})}{\partial \hat{Y}} < 0$ if $Y_{t+h} < \hat{Y}_{t+h,t}$ we have that

$$\frac{1}{e} \cdot \left.\frac{\partial L\left(Y_{t+h}, \hat{Y}_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}}\right|_{Y_{t+h}=\hat{Y}_{t+h,t}+e} > 0 \text{ for all } e \neq 0$$

Thus both the numerator and denominator in the definition of $f^*_{e_{t+h,t}}$ are non-negative, so $f^*_{e_{t+h,t}}\left(e; \hat{Y}_{t+h,t}\right) \geq 0$, if $f_{e_{t+h,t}}\left(e; \hat{Y}_{t+h,t}\right) \geq 0$. By the construction of $f^*_{e_{t+h,t}}$ it is clear that it integrates to 1. ■

Proof of Proposition 15.

$$\begin{aligned} E^*_t\left[e^*_{t+h,t}\right] &\equiv \int e f^*_{e_{t+h,t}}\left(e; \hat{Y}_{t+h,t}\right) de \\ &= E_t\left[\frac{\psi^*_{t+h,t}}{e^*_{t+h,t}}\right]^{-1} \cdot \int \left.\frac{\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}}\right|_{Y_{t+h}=\hat{Y}^*_{t+h,t}+e} \cdot f_{e_{t+h,t}}\left(e; \hat{Y}_{t+h,t}\right) de \\ &= 0 \end{aligned}$$

as the second part of the second line equals zero by the first-order condition for an optimal forecast. The unconditional mean is also zero by the law of iterated expectations. ■

Proof of Proposition 16. Since $E^*\left[e^*_{t+h,t}\right] = 0$ we need only show that $E^*\left[e^*_{t+h,t} \cdot e^*_{t+h+j,t+j}\right] = 0$ for $j \geq h$.

$$\begin{aligned} E^*\left[e^*_{t+h,t} \cdot e^*_{t+h+j,t+j}\right] &= E^*\left[e^*_{t+h,t} \cdot E^*_{t+j}\left[e^*_{t+h+j,t+j}\right]\right] \text{ for } j \geq h \text{ by the LIE} \\ &= 0 \end{aligned}$$

by Proposition 15. ■

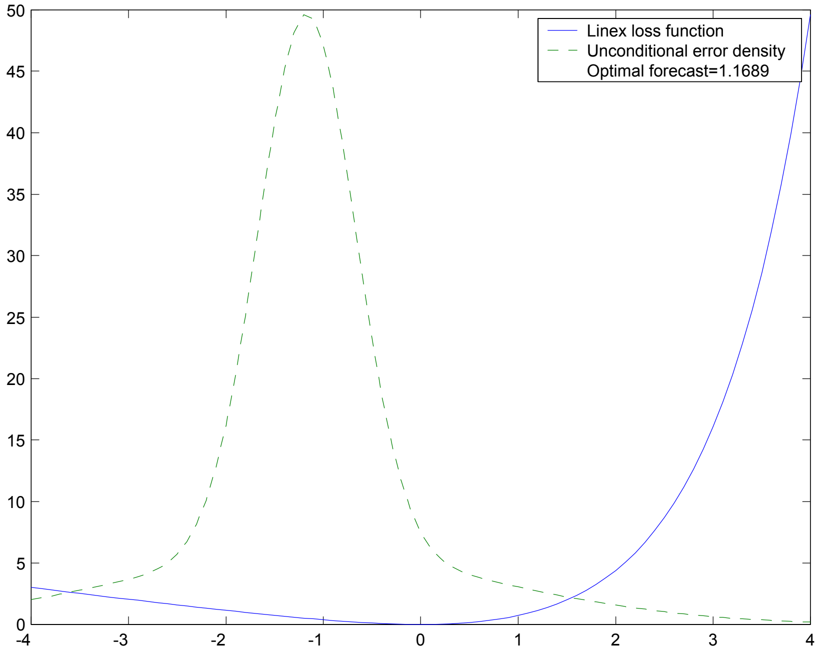

Figure 1: *Linear-exponential loss function and unconditional optimal forecast error density, two-state regime switching example.*

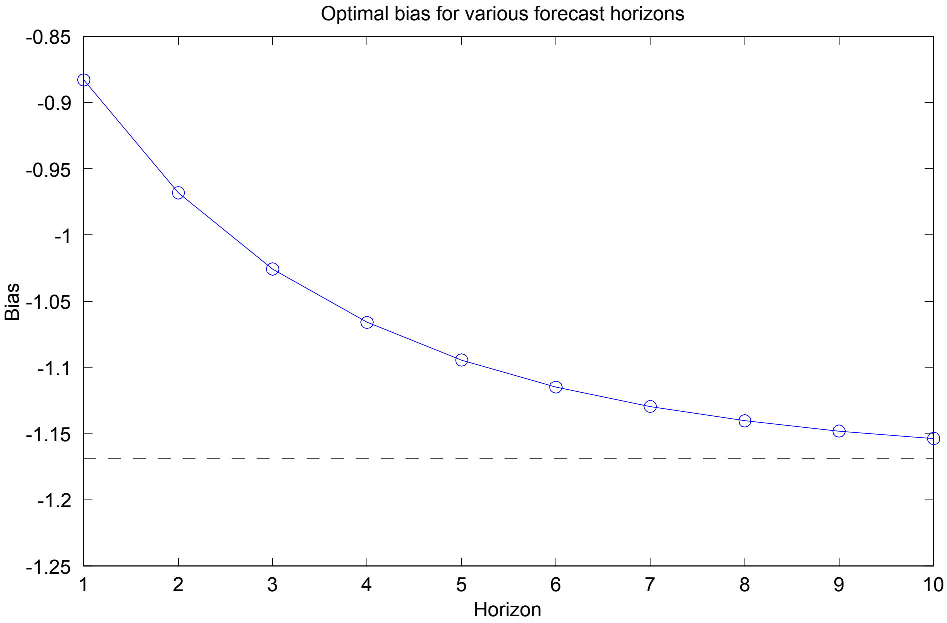

Figure 2: *Bias in the optimal forecast for various forecast horizons, two-state regime switching example.*

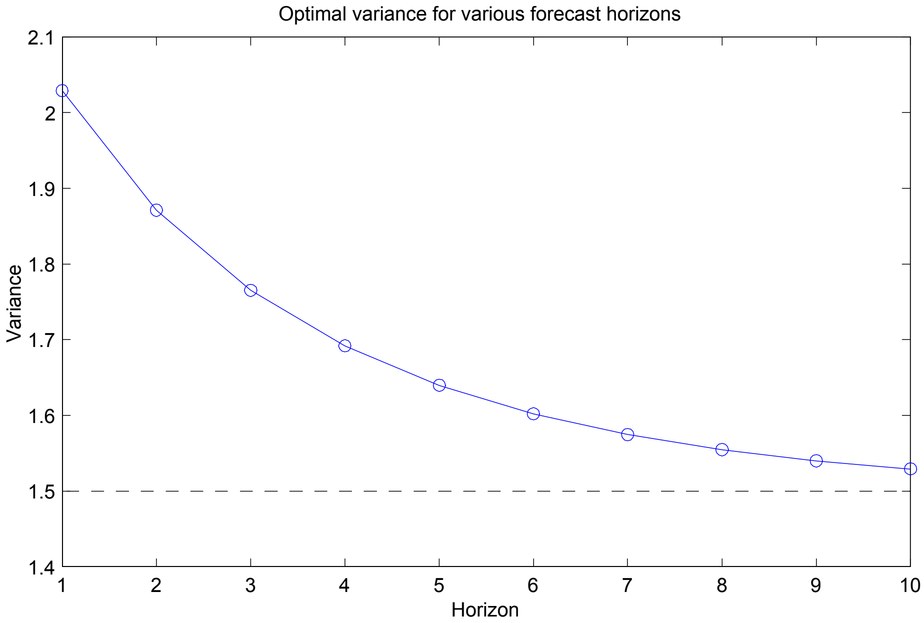

Figure 3: *Variance of the optimal h-step forecast error for various forecast horizons, two-state regime switching example.*

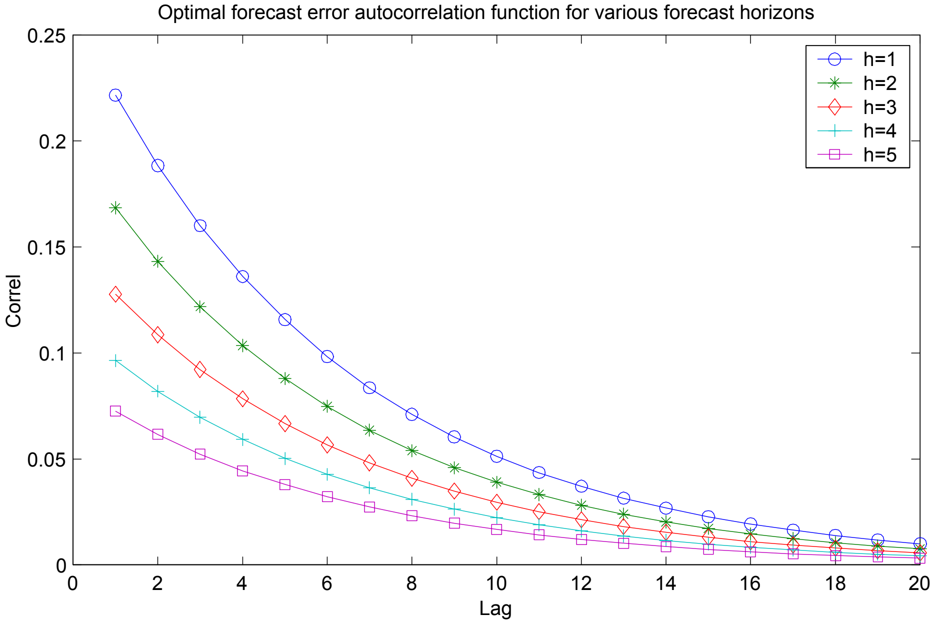

Figure 4: *Autocorrelation in the optimal h-step forecast error for various forecast horizons, two-state regime switching example.*

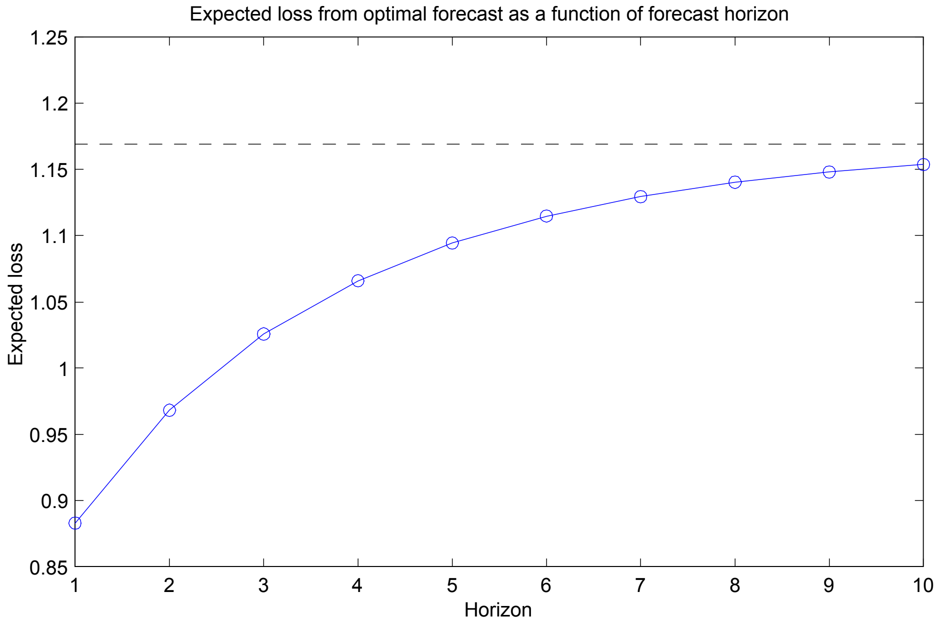

Figure 5: *Expected loss from the optimal forecast for various forecast horizons, two-state regime switching example.*

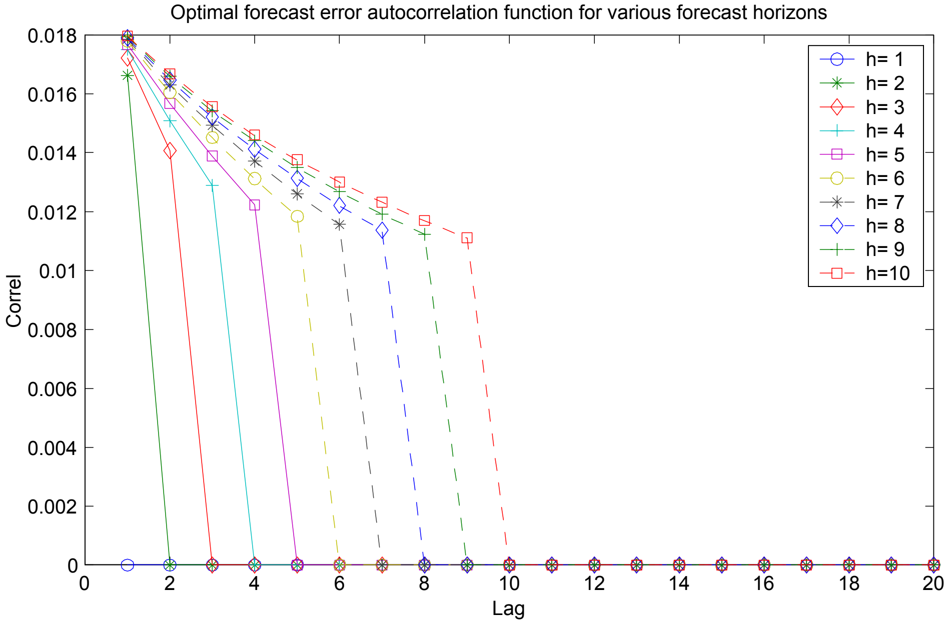

Figure 6: *Autocorrelation in the generalised optimal forecast error for various forecast horizons, two-state regime switching example.*