Testable Implications of Forecast Optimality*

by

Andrew J Patton
London School of Economics and Political Science
and
Allan Timmermann
University of California, San Diego

Contents:

Discussion Paper
No.EM/05/485
February 2005

The Suntory Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020 7955 6679

---

* The authors would like to thank Jeremy Berkowitz, Sean Campbell, Carlos Capistran, Peter Christoffersen, Valentina Corradi, Frank Diebold, Graham Elliott, Raffaella Giacomini, Clive Granger, Peter Hansen, Oliver Linton, Mark Machina, Francisco Penaranda, Jim Powell, Paul Ruud, Kevin Sheppard, Hal White, Stanley Zin and seminar participants at Berkeley, Brown, CORE, Erasmus University Rotterdam, LSE, McGill, Queen Mary, UC Riverside, UCLA, UCSD, Stanford, Warwick Business School and the 2002 EC^2 conference for useful comments. All remaining deficiencies are the responsibility of the authors.

Abstract

Evaluation of forecast optimality in economics and finance has almost exclusively been conducted on the assumption of mean squared error loss under which forecasts should be unbiased and forecast errors serially uncorrelated at the single period horizon with increasing variance as the forecast horizon grows. This paper considers properties of optimal forecasts under general loss functions and establishes new testable implications of forecast optimality. These hold when the forecaster's loss function is unknown but testable restrictions can be imposed on the data generating process, trading off conditions on the data generating process against conditions on the loss function. Finally, we propose flexible parametric estimation of the forecaster's loss function, and obtain a test of forecast optimality via a test of over-identifying restrictions.

Keywords: forecast evaluation; loss function; rationality tests.

JEL No.: C53, C22, C52.

Contact details:
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 areas of economics and finance and is used, inter alia, in tests of the efficient market hypothesis in financial markets and in tests of the rationality of decision makers in a variety of macroeconomic applications. Almost without 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 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.

Since the forecaster's loss function is unknown in most applications, the key question is whether we can derive testable implications of optimal forecasts under more general families of loss functions. Irrespective of the loss function and data generating process, a generalized orthogonality principle must hold provided information is efficiently embedded in the forecast. Implications of this principle will, however, vary significantly with assumptions about the underlying loss function and data generating process (DGP). Most notably, none of the standard properties established in the linear-quadratic framework survives to a more general setting, c.f. Patton and Timmermann (2004). This has important implications for empirical work and means that earlier conclusions concerning the suboptimality of forecasts may have been premature and largely driven by the assumption of quadratic loss.

The contribution of this paper is to establish some surprising new results that trade off restrictions on the forecaster's loss function against restrictions on the DGP. For example, 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 at most display serial dependence of order $h-1$. This holds irrespective of the shape of the loss function and offers a new way to test optimality of forecast errors that is robust to the loss function, but requires restrictions on the underlying DGP. This result will be useful in the common situation where the shape of the loss function is

[1] For references to numerous papers on forecast rationality see www.Phil.frb.org/econ/spf/spfbib.html. For examples of forecast evaluation under objectives other than MSE see West et al. (1993) and Pesaran and Skouras (2001).

unknown, whereas the restrictions on the DGP can be tested empirically. We present similar results that hold when the DGP exhibits heteroskedasticity of a general unknown form, using a new family of quantile-based tests. We also present a method to test forecast optimality based on a flexible model of the loss function, via a test of over-identifying restrictions. Finally, we introduce a transformation from the usual probability measure to an "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 (or "objective") measure.

The outline of the paper is as follows. Section 2 briefly summarizes the properties of optimal forecasts under squared error loss, establishes properties of optimal forecasts under general *known* loss functions and contains the change of measure result. Section 3 derives some testable properties of optimal forecasts when the loss function is unknown but testable restrictions can be imposed on the DGP. Empirical applications to survey forecasts of inflation and output growth are presented in Section 4. Section 5 concludes. An appendix contains technical details and proofs.

2 Testable Implications under General Known Loss Functions

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, $\mathcal{F}_t$. We assume that $Y \equiv \{Y_t : \Omega \to \mathbb{R},\ t = 1, 2, ...\}$ is a stochastic process on a complete probability space $(\Omega, \mathcal{F}, P)$, where $\Omega = \mathbb{R}^{m\infty} \equiv \times_{t=1}^{\infty}\mathbb{R}^m$, and $\mathcal{F} = \mathcal{B}^{m\infty} \equiv \mathcal{B}(\mathbb{R}^{m\infty})$, the Borel σ-field generated by $\mathbb{R}^{m\infty}$. Y_t is thus adapted to the information set available at time t, denoted $\mathcal{F}_t$. At a minimum $\mathcal{F}_t$ includes the filtration generated by $\{Y_{t-k}; k \geq 0\}$, but it may also be expanded to include a set of instruments $Z_t \in \mathcal{F}_t$, which have support $\mathcal{Z}_t$.[2] Let $Z = \{Z_t; t = 1, 2, ...\}$,and denote the conditional distribution of Y_{t+h} given $\mathcal{F}_t$ as $F_{t+h,t}$, i.e. $Y_{t+h}|\mathcal{F}_t \sim F_{t+h,t}$, and the conditional density, if it exists, as $f_{t+h,t}$. Point forecasts conditional on $\mathcal{F}_t$ are denoted by $\hat{Y}_{t+h,t}$ and belong to $\mathcal{Y}$, a compact subset of $\mathbb{R}$, while forecast errors are given by $e_{t+h,t} = Y_{t+h} - \hat{Y}_{t+h,t}$. In general the objective of the forecast is to minimize the expected value of some loss function, $L(Y_{t+h}, \hat{Y}_{t+h,t})$, which is a mapping from

[2] The assumption that Y_t is adapted to $\mathcal{F}_t$ rules out the direct application of the results in this paper to, e.g., volatility forecast evaluation. In such a scenario the object of interest, conditional variance, is not adapted to $\mathcal{F}_t$. Using imperfect proxies for the object of interest can cause difficulties, as pointed out by Hansen and Lunde (2003).

realizations and forecasts to the real line, $L : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$.[3] That is, in general

$$\hat{Y}^*_{t+h,t} \equiv \underset{\hat{y} \in \mathcal{Y}}{\arg\min}\ E_t \left[L \left(Y_{t+h}, \hat{y} \right) \right]. \tag{1}$$

$E_t[.]$ is shorthand notation for $E[.|\mathcal{F}_t]$, the conditional expectation given $\mathcal{F}_t$. We also define the conditional variance, $V_t = E[(Y - E[Y|\mathcal{F}_t])^2|\mathcal{F}_t]$ and the unconditional equivalents, $E[.]$ and $V(.)$.

2.1 Properties under MSE Loss

The vast majority of work on optimal forecasts assumes a squared error loss function:

$$L(Y_{t+h}, \hat{Y}_{t+h,t}) = \theta \left(Y_{t+h} - \hat{Y}_{t+h,t} \right)^2,\ \theta > 0. \tag{2}$$

Under this loss function, optimal forecasts have the following standard properties:

Proposition 1 *Let the loss function be*

$$L \left(Y_{t+h}, \hat{Y}_{t+h,t} \right) = \theta \left(Y_{t+h} - \hat{Y}_{t+h,t} \right)^2, \quad \theta > 0,$$

and assume that $|E_t [Y_{t+h}]| < \infty$ *and* $E_t \left[Y^2_{t+h} \right] < \infty$ *for all* t *and* h. *Then*

1. *The optimal forecast of* Y_{t+h} *is* $E_t [Y_{t+h}]$ *for all forecast horizons* h;
2. *The optimal forecast error is conditionally (and unconditionally) unbiased;*
3. *The optimal* h*-step forecast error exhibits zero serial covariance beyond lag* $(h-1)$*;*

   If we further assume that Y *is covariance stationary, then we obtain:*

4. *The unconditional variance of the optimal forecast error is non-decreasing as a function of the forecast horizon.*

[3] The general decision problem underlying a forecast is to maximize the expected value of some utility function, $U(Y_{t+h}, \mathcal{A}(\hat{Y}_{t+h,t}))$, that depends on the outcome of Y_{t+h} as well as on the decision maker's actions, $\mathcal{A}(\hat{Y}_{t+h,t})$, which in general depend on the full distribution forecast of $Y_{t+h}, F_{t+h,t}$. Our focus on properties of point forecasts is largely dictated by availability of data and need not be very restrictive. For example, Machina and Granger (2004) show that, under conditions on the second derivatives of $U(.)$, there exists a unique point-forecast equivalent which leads to the same decision as if a full distribution forecast had been available.

All proofs are given in the appendix. The proposition shows that the standard properties of optimal forecasts are generated by the assumption of mean squared error loss alone; in particular, assumptions on the DGP (beyond covariance stationarity and finite first and second moments) are not required. Properties such as these have been extensively tested in empirical studies of optimality of predictions or rationality of forecasts, e.g. by testing that the intercept is zero ($\alpha = 0$) and the slope is unity ($\beta = 1$) in the Minzer-Zarnowitz (1969) regression

$$Y_{t+h} = \alpha + \beta \hat{Y}_{t+h,t} + \varepsilon_{t+h}. \tag{3}$$

2.2 Properties under General Loss Functions

Under general loss the first order condition for the optimal forecast is[4]

$$0 = E_t \left[\frac{\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}} \right] = \int \frac{\partial L\left(y, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}} dF_{t+h,t}\left(y\right). \tag{4}$$

This condition can be rewritten using what Granger (1999) refers to as the (optimal) generalized forecast error, $\psi^*_{t+h,t} = \partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right) / \partial \hat{Y}_{t+h,t}$,[5] so that (4) simplifies to

$$E_t[\psi^*_{t+h,t}] = \int \psi^*_{t+h,t} dF_{t+h,t}\left(y\right) = 0. \tag{5}$$

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, $\mathcal{F}_t$. The generalized forecast error is closely related to the "generalized residual" often used in the analysis of discrete, censored or grouped variables, see Gourieroux, *et al.* (1987) and Chesher and Irish (1987) for example. Both the generalized forecast error and the generalized residual are based on first-order (or 'score') conditions.

2.2.1 Technical Assumptions

We next turn our attention to proving properties of the generalized forecast error analogous to those for the standard case. We will sometimes, though not generally, make use of the following assumption on the DGP for $\{Y, Z\}$:

[4] This result relies on the ability to interchange the expectation and differentiation operators. Assumptions L1-L3 given below are sufficient conditions for this to hold.

[5] Granger (1999) considers loss functions that have the forecast error as an argument, and so defines the generalised forecast error as $\psi^*_{t+h,t} \equiv \partial L\left(e_{t+h,t}\right) / \partial e_{t+h,t}$. $\psi^*_{t+h,t}$ can be viewed as the marginal loss associated with a particular prediction, $\hat{Y}_{t+h,t}$.

Assumption D1: *The data generating process for* $\{Y, Z\}$ *is strictly stationary.*

The following properties of the loss function are assumed at various points of the analysis:

Assumption L1: *The loss function is (at least) once differentiable with respect to its second argument, except on a set of* $F_{t+h,t}$*-measure zero, for all* t and h.

Assumption L2: $\int L(y, \hat{y})\, dF_{t+h,t}(y|Z_t) < \infty$ *for some* $\hat{y} \in \mathcal{Y}$*, for all* h *and all* $Z_t \in \mathcal{Z}_t$*, where* $\mathcal{Y}$ *is a compact subset of* $\mathbb{R}$.

Note that assumption L2 implies that $E[L(Y_{t+h}, \hat{y})] < \infty$ for all $\hat{y} \in \mathcal{Y}$:

$$\begin{aligned} E[L(Y_{t+h}, \hat{y})] &= \int_{\mathcal{Z}_t} \int_{\mathbb{R}} L(y, \hat{y})\, dF_{t+h,t}(y|Z_t)\, dF(Z_t) \\ &= \int_{\mathcal{Z}_t} E[L(Y_{t+h}, \hat{y})|Z_t]\, dF(Z_t) \\ &\leq \int_{\mathcal{Z}_t} \sup_{z_t} (E[L(Y_{t+h}, \hat{y})|z_t])\, dF(Z_t) \\ &= \sup_{z_t} E[L(Y_{t+h}, \hat{y})|z_t] \\ &< \infty. \end{aligned}$$

Assumption L2': *An interior optimum of the problem*

$$\min_{\hat{y} \in \mathcal{Y}} \int L(y, \hat{y})\, dF_{t+h,t}(y)$$

exists for all t *and* h.

Assumption L3: $\left| \int (\partial L(y, \hat{y})/\partial \hat{y})\, dF_{t+h,t}(y) \right| < \infty$ *for some* $\hat{y} \in \mathcal{Y}$*, for all* t*,* h.

Assumption L4: $\partial L(y, \hat{y})/\partial \hat{y} \leq (\geq) 0$ *for* $y \geq (\leq)\ \hat{y}$.

Assumption L5: *The loss function is solely a function of the forecast error.*

Assumption L2 simply ensures that the conditional expected loss from a forecast is finite, for some finite forecast. Assumptions L1 and L2' allow us to use the first-order condition of the minimization problem to study the optimal forecast. One set of sufficient conditions for Assumption L2' to hold are Assumption L2 and:

Assumption L5': *The loss function is a non-monotone, convex function solely of the forecast error.*

We do not require that L is everywhere differentiable with respect to its second argument, nor do we need to assume a unique optimum (though this is obtained if we impose Assumption L5', with the convexity of the loss function being strict). Assumption L3 is required to interchange expectation and differentiation: $\partial E_t[L(Y_{t+h}, \hat{y})]/\partial \hat{y} = E_t[\partial L(Y_{t+h}, \hat{y})/\partial \hat{y}]$. The bounds on the

integral on the left-hand side of this expression are unaffected by the choice of $\hat{y}$, and so two of the terms in Leibnitz's rule drop out, meaning we need only assume that the term on the right-hand side is finite. Assumption L4 simply imposes that the loss function is non-decreasing as the forecast moves further away (in either direction) from the true value, which is a reasonable assumption. It is common to impose that $L(\hat{y}, \hat{y}) = 0$, i.e., the loss from a perfect forecast is zero, but this is obviously just a normalization and is not required here.

It is possible to show that under general (non-MSE) loss the properties of the optimal forecast error listed in Proposition 1 can all be violated; see Patton and Timmermann (2004) for an example using a regime switching model and the "linex" loss function of Varian (1974). However, corresponding properties of the generalized forecast error, $\psi^*_{t+h,t}$, can be shown to hold in general:

Proposition 2 *1. Let assumptions L1, L2' and L3 hold. Then the generalized forecast error, $\psi^*_{t+h,t}$, has conditional (and unconditional) mean zero.*

2. Let assumptions L1, L2' and L3 hold. Then the generalized forecast error from an optimal h-step forecast made at time t exhibits zero correlation with any function of any element of the time t information set, $\mathcal{F}_t$, for which second moments exist. In particular, the generalized forecast error will exhibit zero serial correlation for lags greater than $(h-1)$.[6]

3. Let assumptions D1 and L2 hold. Then 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.

This result is useful when the loss function is known since $\psi^*_{t+h,t}$ can then be calculated directly and gives rise to generalized efficiency tests that project $\psi^*_{t+h,t}$ on period-t instruments and test the martingale difference property ($\alpha = \beta = 0$ for all $Z_t \in \mathcal{F}_t$):

$$\psi_{t+h,t} = \alpha + \beta' Z_t + u_{t+h}. \tag{6}$$

The above test will not generally be consistent against all departures from forecast optimality. A consistent test of forecast optimality based on the generalized forecast errors could be constructed using the methods of Bierens (1990), de Jong (1996) and Bierens and Ploberger (1997).

[6] Optimal h-step forecast errors under MSE loss are MA processes of order no greater than $h-1$. In a non-linear 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 $MA(h-1)$ process.

If the same forecaster reported forecasts for multiple horizons we can conduct a joint test of forecast optimality across all horizons. Note that we do not require that the loss function is the same across all horizons, i.e., the one-step ahead forecasting problem may involve a different loss function to the two-step ahead forecasting problem, even for the same forecaster. A joint test of optimality across all horizons may be conducted as:

$$\begin{bmatrix} \psi_{t+1,t} \\ \psi_{t+2,t} \\ \vdots \\ \psi_{t+H,t} \end{bmatrix} = A + BZ_t + u_{t,H} \tag{7}$$

and then testing $H_0 : A = B = 0$ vs. $H_a : A \neq 0 \cup B \neq 0$. More concretely, one possibility is to estimate a vector autoregression (VAR) for the generalized forecast errors:

$$\begin{bmatrix} \psi_{t+1,t} \\ \psi_{t+2,t} \\ \vdots \\ \psi_{t+H,t} \end{bmatrix} = A + B_1 \begin{bmatrix} \psi_{t,t-1} \\ \psi_{t+1,t-1} \\ \vdots \\ \psi_{t+H-1,t-1} \end{bmatrix} + ... + B_J \begin{bmatrix} \psi_{t+1-J,t-J} \\ \psi_{t+2-J,t-J} \\ \vdots \\ \psi_{t+H-J,t-J} \end{bmatrix} + u_{t,H}. \tag{8}$$

Since the dependent variable above is only $\mathcal{F}_{t+H}$-measurable, the appropriate restriction is not that $B_j = 0$ for all j; rather one should test that $A = B_j = 0$ for all $j \geq H$, and that the first j columns of B_j are equal to zero for $1 \leq j < h$.

2.3 Properties under a Change of Measure

In the previous section we showed that by changing our object of analysis from the forecast error to the "generalized forecast error" we can obtain the usual properties of unbiasedness and zero serial correlation. We next consider instead 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, c.f. Harrison and Kreps (1979). 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 the standard properties hold; i.e., under the new measure $\left(e^*_{t+h,t}, \mathcal{F}_t\right)$ is a martingale difference sequence, when $\hat{y} = \hat{Y}^*_{t+h,t}$. We call

the new measure the "MSE-loss probability measure". The resulting method thus suggests an alternative means of evaluating forecasts made using asymmetric loss functions.

The conditional distribution of the forecast error, $F_{e_{t+h,t}}$, given $\mathcal{F}_t$ and a forecast $\hat{y}$, satisfies

$$F_{e_{t+h},t}\left(e;\hat{y}\right)=F_{t+h,t}\left(\hat{y}+e\right), \tag{9}$$

for all $\left(e,\hat{Y}_{t+h,t}\right)\in\mathbb{R}^2$ where $F_{t+h,t}$ is the conditional distribution of Y_{t+h} given $\mathcal{F}_t$. In defining the MSE-loss probability measure we need to make the following assumption:

Assumption L6: $0<E\left[\left(Y_{t+h}-\hat{y}\right)^{-1}\partial L\left(Y_{t+h},\hat{y}\right)/\partial\hat{y}\middle|Z_t\right]<\infty$ *for all* h, *all* $\hat{y}\in\mathcal{Y}$, *and all* $Z_t\in\mathcal{Z}_t$.

Definition 1 *Let assumptions L4 and L6 hold and let*

$$\Lambda\left(e,\hat{y}\right)\equiv\frac{1}{e}\cdot\left.\frac{\partial L\left(y,\hat{y}\right)}{\partial\hat{y}}\right|_{y=\hat{y}+e} \tag{10}$$

Then the "MSE-loss probability measure", $d\tilde{F}_{e_{t+h,t}}\left(\cdot|\hat{y}\right)$, *is defined by*

$$d\tilde{F}_{e_{t+h,t}}\left(e;\hat{y}\right)=\frac{\Lambda\left(e,\hat{y}\right)}{E_t\left[\Lambda\left(Y_{t+h}-\hat{y},\hat{y}\right)\right]}\cdot dF_{e_{t+h,t}}\left(e;\hat{y}\right) \tag{11}$$

By construction the MSE-loss probability measure $\tilde{F}\left(\cdot|\hat{y}\right)$ is absolutely continuous with respect to the usual probability measure, $F\left(\cdot|\hat{y}\right)$, (that is, $\tilde{F}\left(\cdot|\hat{y}\right)<<F\left(\cdot|\hat{y}\right)$). The function

$$\tilde{\Lambda}_{t+h,t}\left(e,\hat{y}\right)\equiv\frac{\Lambda\left(e,\hat{y}\right)}{E_t\left[\Lambda\left(Y_{t+h}-\hat{y},\hat{y}\right)\right]} \tag{12}$$

is the Radon-Nikodým derivative $d\tilde{F}_{e_{t+h},t}\left(\cdot|\hat{y}\right)/dF_{e_{t+h},t}\left(\cdot|\hat{y}\right)$. If we let $u=e^{-1}$, then Assumption L6 requires that $\left.\partial L\left(y,\hat{y}\right)/\partial\hat{y}\right|_{y=\hat{y}+1/u}=O\left(u^{-1}\right)$. Note that $\Lambda\left(e,\hat{y}\right)$ is well-defined at $e=0$ for some common loss functions. For example,

$$\begin{aligned}
MSE &: \lim_{e\to0}\Lambda\left(e,\hat{y}\right)=-2\\
Linex &: \lim_{e\to0}\Lambda\left(e,\hat{y}\right)=-a^2\\
PropMSE &: \lim_{e\to0}\Lambda\left(e,\hat{y}\right)=-2/\hat{y}^2
\end{aligned}$$

where the *Linex* and *PropMSE* loss functions are $L\left(y,\hat{y}\right)=\exp(ae)-ae+1$ $(a\neq0)$ and $L\left(y,\hat{y}\right)=\left(y/\hat{y}-1\right)^2$, respectively. For mean absolute error loss, $L\left(y,\hat{y}\right)=|e|$, the limits from both directions diverge to $-\infty$, meaning that there is no MSE-loss density under MAE in general. However, if the variable of interest is conditionally symmetrically distributed at all points in time, then the optimal

forecast under MAE coincides with the optimal forecast under MSE, as the conditional mean is equal to the conditional median, and so the appropriate Radon-Nikodým derivative is equal to one.

We now show that under the MSE-loss probability measure the optimal h-step ahead forecast errors exhibit the properties that we would expect from optimal forecasts under MSE loss:

Proposition 3 *1. Let assumptions L1, L4 and L6 hold. Then the "MSE-loss probability measure", $\tilde{F}_{e_{t+h,t}}(\cdot|\hat{y})$, defined in equation (11) is a proper probability distribution function for all $y \in \mathcal{Y}$.*

2. If we further let assumption L2' hold, then the optimal forecast error, $e^*_{t+h,t} = Y_{t+h} - \hat{Y}^*_{t+h,t}$ has conditional mean zero under the MSE-loss probability measure $\tilde{F}_{e_{t+h,t}}\left(\cdot|\hat{Y}^*_{t+h,t}\right)$.

3. The optimal forecast error is serially uncorrelated under the MSE-loss probability measure, $\tilde{F}_{e_{t+h,t}}\left(\cdot|\hat{Y}^*_{t+h,t}\right)$, for all lags greater than $h-1$.

4. $\tilde{V}\left[e^*_{t+h,t}\right]$, the variance of $e^*_{t+h,t}$ under $\tilde{F}_{e_{t+h,t}}$ evaluated at $\hat{Y}^*_{t+h,t}$, is non-decreasing as a function of the forecast horizon.

Notice that $e^*_{t+h,t}$ is a martingale difference sequence, with respect to $\mathcal{F}_t$, under $\tilde{F}_{t+h,t}$. Furthermore, although the MSE loss probability measure operates on forecast errors, the result holds for general loss functions having $Y_{t+h}, \hat{Y}^*_{t+h,t}$ as separate arguments.

It is worth emphasizing that the MSE-loss probability measure is a *conditional* distribution, and so obtaining an estimate of it from data is not as simple as it would be if it was an unconditional distribution. If we assume that the density $f_{e_{t+h},t}$ exists then it is possible, under some conditions, to obtain a consistent estimate of $f_{e_{t+h},t}$ via semi-nonparametric density estimation, see Gallant and Nychka (1987). If L is known then Λ is, of course, also known.[7] With consistent estimates of $f_{e_{t+h},t}$ and Λ it is simple to construct an estimator of $\tilde{f}_{e_{t+h},t}$.

To illustrate how the MSE-loss error density differs from the objective error density, we present a simple example in Figure 1 that assumes linex loss with parameter $a = 1$, and that the DGP is a simple two-state regime switching process, with $Y_{t+1}|\mathcal{F}_t$ being $N\left(0, \sigma^2_{s_{t+1}}\right)$ in state s_{t+1}, where $\sigma_1 = 0.5$ and $\sigma_2 = 2$, and with $\Pr[S_{t+1} = 1|S_t = 1] = 0.95$ and $\Pr[S_{t+1} = 2|S_t = 2] = 0.90$. This is a special case of the widely-used regime switching model proposed by Hamilton (1989). The parameters selected for this example are not dissimilar to empirical results frequently obtained when

[7] If L is unknown, a nonparametric estimate of Λ may be obtained via sieve estimation methods, for example, see Andrews (1991) or Chen and Shen (1998).

this model is estimated on financial data, c.f. Patton and Timmermann (2004) for example. Figure 1 shows the objective and MSE-loss error densities, evaluated at $\hat{Y}^*_{t+h,t}$, for various values of the state probability vector, $\hat{\pi}_{s_t,t} \equiv [\Pr[S_t = 1|\mathcal{F}_t], \Pr[S_t = 2|\mathcal{F}_t]]'$. The shape of the transformation from f to $\tilde{f}$ differs depending on the state probabilities, but in all cases probability mass is shifted to the right in order to remove the (optimal) negative bias that is present under the objective probability distribution for e due to the high cost associated with positive forecast errors.

3 Testable Implications under Unknown Loss Functions

The results in Proposition 2 can be used to test the optimality of a sequence of forecasts if the forecaster's loss function is known, and certain basic conditions hold. In this section we establish some properties of optimal forecasts that may be tested when the loss function of the forecaster is *unknown*. To obtain some of these results we consider a restricted class of DGPs, namely those with dynamics in the conditional mean and conditional variance, but no dynamics in the remainder of the conditional distribution. This class of DGPs is quite broad, and includes ARMA processes and non-linear regressions, possibly with GARCH or stochastic volatility in the conditional variance process. Throughout, we will use the following notation: $\mu_{t+h,t} \equiv E_t[Y_{t+h}]$ and $\sigma^2_{t+h,t} \equiv V_t[Y_{t+h}]$.

3.1 Conditional mean dynamics only

Consider the class of DGPs that satisfy the following condition:

Assumption D2: *The DGP is such that* $Y_{t+h} = \mu_{t+h,t} + \varepsilon_{t+h}$, $\varepsilon_{t+h}|\mathcal{F}_t \sim F_{\varepsilon,h}\left(0, \sigma^2_{\varepsilon,h}\right)$, *where* $F_{\varepsilon,h}\left(0, \sigma^2_{\varepsilon,h}\right)$ *is some distribution, with mean zero and variance* $\sigma^2_{\varepsilon,h}$ *(which may be infinite), that may depend on* h*, but does not depend on* $\mathcal{F}_t$.

The restriction of dynamics only in the conditional mean implies that the innovation term, ε_{t+h}, is drawn from some distribution, $F_{\varepsilon,h}$, which will generally depend on the forecast horizon, but is *independent* of $\mathcal{F}_t$ and so is not denoted with a subscript t. We shall concentrate on loss functions that satisfy assumption L5, i.e. $L(y, \hat{y}) = L(y - \hat{y}) = L(e)$, $\forall\ (y,\ \hat{y}) \in \mathbb{R} \times \mathcal{Y}$. Although this restriction rules out certain loss functions, many common loss functions are of this form, for example lin-lin and linex. With these two assumptions we obtain the following serial correlation property of the optimal forecast error:

Proposition 4 *Let the DGP and loss function satisfy assumptions D2, L2 and L5. Then* $e^*_{t+h,t}$ *is independent of all* $Z_t \in \mathcal{F}_t$. *In particular,* $Cov\left(e^*_{t+h,t}, e^*_{t+h-j,t-j}\right) = 0$ *for all* $j \geq h$ *and any* $h > 0$.

This proposition shows that under a testable 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* error-based loss function. This implies that given a sequence of realizations and forecasts, $\left\{Y_{t+h}, \hat{Y}_{t+h,t}\right\}_{t=1}^{T}$, we may test for forecast optimality *without* knowledge of the forecaster's loss function by testing the serial correlation properties of the forecast errors.[8] For financial applications the assumption of constant higher-order conditional moments is clearly too strong, but in some 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, 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, e.g. by testing $\beta = 0$ in regressions such as

$$e_{t+h,t} = \alpha + \beta' Z_t + u_{t+h}. \tag{13}$$

If forecasts with various horizons are available from the same forecaster we may again "stack" the forecast errors into a vector, and test the optimality of all forecasts jointly:

$$\begin{bmatrix} e_{t+1,t} \\ e_{t+2,t} \\ \vdots \\ e_{t+H,t} \end{bmatrix} = A + BZ_t + u_{t,H}$$

and then test $H_0 : B = 0$. Alternatively, a VAR structure similar to equation (8) could be employed.

In Patton and Timmermann (2004) it was shown that although the unconditional expected loss is always a non-decreasing function of the forecast horizon, the unconditional forecast error variance may or may not be a non-decreasing function of the forecast horizon. We next establish conditions under which the unconditional forecast error variance is non-decreasing.

[8] We focus on the testing of the serial correlation property though of course more generally we could test for complete independence between $e^*_{t+h,t}$ and any $Z_t \in \mathcal{F}_t$.

Proposition 5 *Let the loss function satisfy assumptions L2, L4 and L5, and the DGP satisfy D2 with* $\sigma^2_{\varepsilon,h} < \infty$*, for some* $h \geq 1$*. Then* $V\left[e^*_{t+h,t}\right]$ *is a weakly increasing function of* h*.*

Like Proposition 4, this 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 5 makes it clear that optimality tests based on the variance of the forecast error being weakly increasing in the forecast horizon require restrictive assumptions either on the loss function (i.e., MSE loss in particular) or on the DGP (i.e., dynamics in the conditional mean only). This suggests that the equivalence between unconditional expected loss and unconditional error variance is peculiar to the MSE loss function, and will not generally be true for other loss functions.

3.2 Conditional mean and conditional variance dynamics

The assumption of constant conditional variance is too restrictive in many applications. We next provide results for a more general class of conditional scale-location DGPs, that satisfy the following assumption:

Assumption D3: *The DGP is such that* $Y_{t+h} = \mu_{t+h,t} + \sigma_{t+h,t}\eta_{t+h}$, $\eta_{t+h}|\mathcal{F}_t \sim F_{\eta,h}(0,1)$*, where* $F_{\eta,h}(0,1)$ *is some distribution with mean zero and variance one that may depend on* h*, but does not depend on* $\mathcal{F}_t$*.*

This class of DGPs is very broad and includes most common volatility processes, e.g. ARCH and stochastic volatility, see Engle (1982) and Shephard (2004). It nests those of Assumption D2, at the cost that we must be more restrictive on the class of loss functions that we consider.[9] Specifically, we need to make the following assumption:

Assumption L5": *The loss function is a homogeneous function solely of the forecast error.*

This assumption implies that $L(ae) = g(a)\,L(e)$ for some positive function g. Commonly used loss functions such as lin-lin and asymmetric quadratic loss are of this form while the linex loss function is excluded. With these two assumptions we obtain the following testable implications of forecast optimality:

[9] It would be possible to extend our analysis to allow η_{t+h} to have no finite moments. In such a case $\mu_{t+h,t}$ and $\sigma_{t+h,t}$ would no longer be interpretable as the conditional mean and standard deviation of $Y_{t+h}|\mathcal{F}_t$, and would instead simply be $\mathcal{F}_t$-measurable location and scale shifters. We do not consider this extension here.

Proposition 6 *Let the DGP and loss function satisfy assumptions D3 and L2, and define the standardized optimal forecast error as* $d^*_{t+h,t} = e^*_{t+h,t}/\sigma_{t+h,t}$. *Then:*

1. If L5" also holds, the optimal forecast takes the following form:

$$\hat{Y}^*_{t+h,t} = \mu_{t+h,t} + \sigma_{t+h,t} \cdot \gamma^*_h \tag{14}$$

where γ_h *is a constant, depending only upon the forecast horizon,* h*, and the loss function.*

2. If L5" also holds, $d^*_{t+h,t}$ *is independent of any element* $Z_t \in \mathcal{F}_t$. *In particular,* $Cov\left(d^{*r}_{t+h,t}, d^{*s}_{t+h-j,t-j}\right) = 0$ *for all* $j \geq h$ *and any* $h > 0$ *and all* r, s *for which the covariance exists. Further,* $V_t\left[d^*_{t+h,t}\right] = 1$ *for all* $h > 0$.

3. If L5" does not hold but L5 holds, then

$$\hat{Y}^*_{t+h,t} = \mu_{t+h,t} + \gamma^*\left(\sigma_{t+h,t}, L, h\right)$$

where $\gamma^*_{t+h,t} \equiv \gamma^*\left(\sigma_{t+h,t}, L, h\right)$ *is, in general, time-varying and depends on* $\mathcal{F}_t$, h *and* L.

The second part of Proposition 6 spells out the empirical implications of the result under homogenous loss. Although the optimal forecast error will not, in general, be unbiased, serially uncorrelated or homoskedastic, an optimal forecast error scaled by the conditional standard deviation will be independent of any $Z_t \in \mathcal{F}_t$. This implies that $d^*_{t+h,t}$ will be serially uncorrelated and homoskedastic. Forecast optimality could therefore be tested, for instance, by estimating

$$\begin{aligned} d_{t+h,t} &= \alpha_0 + \sum_{i=0}^{p} \beta_i d_{t-i,t-i-h} + u_{t+h} \\ u_{t+h} &= \sigma^2_{u,t+h} v_{t+h},\ v_{t+h} \sim (0,1) \\ \sigma^2_{u,t+h} &= \omega_0 + \sum_{i=1}^{q} \omega_i u^2_{t+1-i} \end{aligned} \tag{15}$$

and testing that $H_0 : \omega_0 = 1 \cap \beta_i = \omega_j = 0\ \forall\ i = 0, 1, 2, ..., p$ and $j = 1, 2, ..., q$. As in previous cases, if forecasts with different horizons are available from the same forecaster we may stack the standardized forecast errors into a vector and test the optimality of all forecasts jointly.

The above test is easily computed although it requires that an estimate of $\sigma^2_{t+h,t}$ is available. Under certain conditions, a consistent estimate of this conditional variance can be obtained from the observed Y_t process either by means of parametric methods (e.g. using a GARCH-type model) or by non-parametric methods, using a realized volatility estimator (see Andersen, *et al.*, 2003) for example. In contrast, and importantly for empirical work, no estimate of $\mu_{t+h,t}$ is required.

If the loss function is not homogenous, part 3 of Proposition 6 shows that the optimal forecast will, in general, be a function of the time-varying conditional variance, $\sigma_t^2(Y_{t+h})$, and the loss function, L. Since $\gamma_{t+h,t}$ is time-varying and depends on the unknown loss function this makes it difficult to design tests of forecast optimality under time-varying first and second moments and unknown (non-homogeneous) loss.

Researchers will not always have a reliable estimate of $\sigma_{t+h,t}^2$ available, and so it would be particularly useful to establish under which conditions an optimality test can be based only on such observables. If we restrict the first and second moment dynamics to be linked then such an implication may be obtained. This is a case of particular interest in financial applications - i.e. when the target variable is returns and expected returns are proportional to the level of risk as measured by the conditional standard deviation, see Engle, *et al.* (1987).

Assumption D3': *The DGP is such that* $Y_{t+h} = \beta\sigma_{t+h,t} + \sigma_{t+h,t}\eta_{t+h}$, $\eta_{t+h}|\mathcal{F}_t \sim F_{\eta,h}(0,1)$, *where* $\beta \in \mathbb{R}$ and $F_{\eta,h}$ *is some distribution with mean zero and variance one that may depend on* h, *but does not depend on* $\mathcal{F}_t$.

Proposition 7 *Let the DGP and loss function satisfy assumptions D3', L2 and L5", and assume that* $\beta \neq -\gamma_h$. *Define* $d_{t+h,t}^* \equiv \left(Y_{t+h} - \hat{Y}_{t+h,t}^*\right)/\hat{Y}_{t+h,t}^*$. *Then* $d_{t+h,t}^*$ *is independent of any element* $Z_t \in \mathcal{F}_t$. *In particular,* $Cov\left(d_{t+h,t}^{*r}, d_{t+h-j,t-j}^{*s}\right) = 0$ *for all* $j \geq h$ *and any* $h > 0$ *and all* r, s *for which the covariance exists.*

Note that the assumption that $\beta \neq -\gamma_h$ is easily checked: if this is true then the optimal forecast is identically zero for all t. Also, note that the unit variance property of the standardized optimal forecast error in Proposition 6 no longer holds. Under the conditions of Proposition 7 we may test forecast optimality without specific knowledge of the loss function or any of the moments of the DGP, by testing that there is no serial correlation beyond lag $h-1$ in the $d_{t+h,t}^*$ series, and/or that the $d_{t+h,t}^*$ series is homoskedastic. This could be done simply via a Mincer-Zarnowitz regression of powers of $d_{t+h,t}^*$ on a constant and lags of various powers of $d_{t+h,t}^*$ of order greater than or equal to h, as in equation (15).

Under certain conditions, it is possible to show that we can express the optimal forecast as a conditional quantile of the variable of interest. The usefulness of this result lies in the surprising finding that the optimal forecast is the *same* quantile at all points in time, though the quantile may

change with the forecast horizon and is typically unknown as it depends on the loss function. With such a representation we can obtain an alternative test of forecast optimality using tests of quantile forecasts without the need to estimate of the conditional variance of the variable of interest.

Proposition 8 *Let the DGP and loss function satisfy assumptions D2, L2 and L5, or assumptions D3, L2 and L5" Then:*

1. *The optimal forecast is such that, for all t,*

$$F_{t+h,t}\left(\hat{Y}^*_{t+h,t}\right) = q^*_h, \tag{16}$$

where $q^*_h \in (0,1)$ depends only upon the forecast horizon, h, and the loss function, L. If $F_{t+h,t}$ is continuous and strictly increasing, then we can alternatively express this as:

$$\hat{Y}^*_{t+h,t} = F^{-1}_{t+h,t}\left(q^*_h\right) \tag{17}$$

2. *Let*

$$I^*_{t+h,t} \equiv \mathbf{1}\left(Y_{t+h} \leq \hat{Y}^*_{t+h,t}\right) \tag{18}$$

Then $I^*_{t+h,t}$ is independent of all $Z_t \in \mathcal{F}_t$. In particular, $I^*_{t+h,t} - q_h$ is a martingale difference sequence with respect to $\mathcal{F}_t$.

Notice how assumptions on the loss function can be traded off against assumptions on the DGP. This result gives rise to a new test that is applicable even though q^*_h is unknown. The test simply projects the indicator function on elements in $\mathcal{F}_t$ and an intercept and tests that $\beta = 0$:

$$I^*_{t+h,t} = \alpha + \beta Z_t + u_{t+h}. \tag{19}$$

A logit model could instead be used, to better reflect the binary nature of the dependent variable. Alternatively, the LR test of independence of Christoffersen (1998) could be employed to test for serial dependence in $I^*_{t+h,t}$. If q^*_h is known, it can further be tested that $\alpha = q^*_h$. We may again stack the indicator variables into a vector and test multiple horizons jointly, if forecasts for multiple horizons are available.

When D3 fails to hold it is, in general, difficult to obtain results that are easy to test even if restrictions are imposed on the loss function. To see this, consider the following more general DGP

Assumption D4: *The DGP is such that* $Y_{t+h} = \mu_{t+h,t} + \sigma_{t+h,t}\eta_{t+h}$, $\eta_{t+h}|\mathcal{F}_t \sim F_{\eta,t+h,t}(0,1)$, *where* $F_{\eta,t+h,t}(0,1)$ *is some time-varying distribution with mean zero and variance one that depends on* $\mathcal{F}_t$ *and possibly on* h.

This class of DGPs nests those of Assumption D3, as we allow for a time-varying conditional mean, conditional variance and other properties of the distribution (e.g., time-varying conditional skew or kurtosis.) If the loss function is assumed to be homogeneous in the forecast error we get the following result:

Proposition 9 *Let the DGP and loss function satisfy assumptions D4, L2 and L5". Then*

$$\hat{Y}^*_{t+h,t} \equiv \arg\min_{\hat{y}} \; E_t\left[L\left(Y_{t+h}, \hat{y}\right)\right] = \mu_{t+h,t} + \sigma_{t+h,t} \cdot \gamma^*_{t+h,t}$$

where $\gamma^*_{t+h,t}$ *is scalar that depends on the loss function, the forecast horizon, and the time-varying properties of* $F_{\eta,t+h,t}$ *beyond the conditional mean and variance. The standardized optimal forecast error,* $d^*_{t+h,t} = e^*_{t+h,t}/\sigma_{t+h,t} = \left(\eta_{t+h} - \gamma^*_{t+h,t}\right)$, *and the indicator variable* $I^*_{t+h,t} \equiv \mathbf{1}\left(Y_{t+h} \leq \hat{Y}^*_{t+h,t}\right)$, *will both in general be serially correlated beyond lag* $h-1$.

Thus when dynamics in the conditional distribution beyond those in the conditional mean and variance are considered we cannot in general obtain a testable restriction on the optimal forecast error, even if $\sigma_{t+h,t}$ is known (or, more generally, even if $F_{t+h,t}$ is known) and we restrict the loss function to satisfy assumption L5".

3.3 General DGPs and flexible parametric loss functions

In cases such as those covered by the previous Proposition, the variable of interest may have dynamics beyond the conditional mean and variance, or the class of loss functions to be considered will not satisfy the restrictions in assumptions L5 or L5". In this case it may still be possible to construct a test based on a flexible parametric estimate of the first-derivative of the loss function with respect to $\hat{y}$. Recall the first-order condition: $0 = E_t\left[\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)/\partial\hat{y}\right]$, which implies $0 = E\left[\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)/\partial\hat{y} \cdot Z_t\right]$ for any $Z_t \in \mathcal{F}_t$.

For notational simplicity, let

$$\lambda\left(y, \hat{y}\right) \equiv \frac{\partial L\left(y, \hat{y}\right)}{\partial \hat{y}}. \tag{20}$$

We may obtain a flexible parametric estimate of $\lambda\left(y, \hat{y}\right)$, denoted $\lambda\left(y, \hat{y}; \theta\right)$, based on a linear spline model, for example. To see how a linear spline could be employed to approximate the function

$\lambda(y, \hat{y})$, assume that $\lambda = \lambda(e)$, let $(\zeta_1, ..., \zeta_K)$ be the nodes and impose that one of the nodes is zero. We impose that the spline is continuous, though not necessarily differentiable, except possibly at zero. We could allow further discontinuities in λ at the cost of introducing more parameters to estimate.

With just a few nodes this class of loss functions is very flexible, nesting both MSE and MAE as special cases, as well as the "quad-quad", "lin-lin", and the symmetric, non-convex loss function of Granger (1969). If we further impose that the spline is continuous at zero, then the MSE loss function is nested without the boundary of the parameter space being hit, at the cost of the MAE and "lin-lin" loss functions not being nested. In this case the resulting estimated loss function is a quadratic spline, and is continuous and (once) differentiable everywhere:

$$\frac{\partial \lambda(e;\theta)}{\partial e} = \begin{cases} \gamma_1, & \text{for } e \leq \zeta_1 \\ \gamma_i, & \text{for } \zeta_{i-1} < e \leq \zeta_i,\ i = 2, ..., K \\ \gamma_{K+1}, & \text{for } e > \zeta_K \end{cases} . \tag{21}$$

where $\theta = \left[\gamma_1, \gamma_2, ..., \gamma_{K+1}\right]'$. $\lambda(e;\theta)$ and $L(e;\theta)$ are constructed from the above specification by imposing that $\lambda(0;\theta) = L(0;\theta) = 0$ and that both $\lambda(e;\theta)$ and $L(e;\theta)$ are continuous in e.[10] Since $\lambda(e;\theta)$ is only identified up to a multiplicative constant, some normalization is required to identify the parameters. Further, it is important to impose constraints on θ so that the resulting estimate of λsatisfies the assumptions required for it to be the first derivative of some valid loss function; for example, assumption L4 requires $\lambda(y, \hat{y}) \leq (\geq)\ 0$ for $y \geq (\leq)\ \hat{y}$.

In applications where we have reason to assume that the loss from a forecast is solely a function of the forecast error−i.e. assumption L5 is satisfied−the problem simplifies to approximating the function $\lambda(y - \hat{y}) = \lambda(e)$. In other cases, such a restriction may not be well-founded and so no such simplification is available, c.f. Machina and Granger (2004). In this case we must employ a more flexible specification to approximate the function $\lambda(e, y)$. Treating $\lambda(y, \hat{y})$ rather than $\lambda(e, y)$ makes it more difficult to impose the conditions necessary for λ to be admissible. We propose the following specification, which is centred around MSE loss, so that $\theta = 0$ implies MSE loss.

[10] If the number of parameters in the spline grows with the sample size, and the model gets sufficiently flexible as to be able to approximate the unknown function arbitrarily well, then the resulting estimate has an interpretation as a nonparametric estimate, see Andrews (1991) or Chen and Shen (1998), though we do not employ this interpretation here.

$$\frac{\partial \lambda(e,y;\theta)}{\partial e} = \begin{cases} \gamma_1 \equiv \Gamma(\varphi_{01} + \varphi_{11} y - \ln K), & \text{for } e \leq \zeta_1 \\ \gamma_i \equiv \left(1 - \sum_{j=1}^{i-1} \gamma_j\right) \cdot \Gamma(\varphi_{0i} + \varphi_{1i} y - \ln K), & \text{for } \zeta_{i-1} < e \leq \zeta_i,\ i = 2, ..., K \\ \gamma_{K+1} = 1 - \sum_{j=1}^{K} \gamma_j, & \text{for } e > \zeta_K \end{cases} \tag{22}$$

where $\Gamma(x) \equiv (1 + e^{-x})^{-1}$ is the logistic transformation. This specification allows y to affect the slopes of λ, guarantees that all slopes are weakly positive, and that the sum of the slopes equals one. Under standard regularity conditions the parameter vector of the approximating function can be estimated via the generalized method of moments (GMM), in a similar fashion to Elliott, *et al.* (2002):

$$\begin{aligned} \hat{\theta}_T &\equiv \arg\min_{\boldsymbol{\theta} \in \Theta} \; g_T(\theta)' W g_T(\theta) \\ g_T(\theta) &\equiv \frac{1}{T} \sum_{t=1}^{T} \lambda\left(e_{t+h,t}, \hat{Y}_{t+h,t}; \theta\right) \cdot Z_t, \end{aligned} \tag{23}$$

where W is a weighting matrix and Θ is a compact set. A test of forecast optimality can be obtained from a test of over-identifying restrictions if we ensure we have more moment restrictions, k, than parameters, p:

$$T g_T\left(\hat{\theta}_n\right)' \hat{W}_T^* g_T\left(\hat{\theta}_T\right) \Rightarrow \chi^2_{k-p}, \text{ as } T \to \infty \tag{24}$$

where $\hat{W}_T^*$ is a consistent estimate of the optimal weight matrix, c.f. Newey and McFadden (1994). An important condition for identification is that the data are not *iid*. If the data are *iid* then we have only a single moment condition from the first-order condition for forecast optimality, and so we cannot estimate more than a single parameter, and we have no over-identifying restrictions available to test. Of course, this condition is likely to be satisfied in most time series applications.

This test of forecast optimality does not rely on any restrictions on the DGP, other than standard conditions required for GMM estimators to be consistent and asymptotically normal.[11] It does, however, rely on the linear spline being sufficiently flexible to approximate the unknown loss function. Thus, a rejection of forecast optimality may be due either to a true failure of forecast optimality or to a failure of the approximation of the forecaster's loss function.

[11] One important GMM regularity condition is that the optimal value of the parameter θ lies in the interior of the set of possible values, Θ. A Taylor series approximation of order greater than one will nest MSE loss, but the MSE loss case lies on the boundary of Θ. If the true loss function is MSE this invalidates the use of standard asymptotic theory, though methods of dealing with this have been proposed, see Andrews (2001). This issue does not arise with our linear spline approximation.

4 Empirical tests of forecast optimality

We next apply the proposed tests to data sets comprised of forecasts and realizations of two key macroeconomic variables: inflation and output growth. We shall use the results of Section 3 which are applicable when the loss function is unknown to the econometrician. In both cases the data set comes from the Survey of Professional Forecasters and is maintained by the Federal Reserve Bank of Philadelphia. To obtain a sufficient number of observations we test the optimality of the consensus forecasts of these variables, defined as the median values of all forecasts available each quarter. By testing the consensus forecast, which of course need not equal the forecast of any individual forecaster, we are testing whether there exists a loss function that would make the forecasts of a "representative forecaster" optimal.

4.1 Inflation Forecasts

We first study quarterly forecasts and realizations of inflation over the period 1983 to 2003, giving 84 observations. Data prior to 1983 was not used due to the well-known change in monetary policy and inflation dynamics over the period 1979-1982.

Under MSE loss, a simple test of optimality would be to test whether the unconditional mean of the forecast errors is zero. The unconditional mean (t-statistic) over this period is -0.35 (-4.73), which is significantly different from zero at the 1% level.[12] Thus we have evidence against optimality of this forecast under MSE loss. However, the presence of asymmetry in the loss function may make the presence of such bias optimal. Engle's (1982) LM test for ARCH in the forecast errors for this series, using four lags, found no evidence of heteroskedasticity in the forecast errors (p-value of 0.80), and thus assumption D2 may apply to this series. Further, simple tests of serial correlation in higher powers of the forecast error, in the spirit of Diebold, *et al.* (1998), support assumption D2 for this series, see Table 1.

In this case, we know that if we can further impose that the loss function is only a function of the forecast error (i.e., assumption L5) then an optimal forecast will generate forecast errors with zero serial correlation. A Ljung-Box test indicated no significant serial correlation in the forecast

[12] This is true whether we control for serial correlation in the residuals or not. Under the null of forecast optimality under MSE loss the forecast errors are serially uncorrelated, though they may be heteroscedastic. We control for heteroscedasticity using White's (1980) variance estimator; the p-value is unchanged if we instead use the Newey-West (1987) variance estimator which also controls for serial correlation.

errors for any lag up to lag 12. Thus we conclude that while we have evidence that the consensus inflation forecasts are not optimal under MSE loss, there is no evidence against optimality under some other, unknown, loss function.

We can also apply two quantile-based tests of forecast optimality presented in Proposition 8. In the first test we regress the indicator variable $\mathbf{1}\left(Y_{t+1} \leq \hat{Y}_{t+1}\right)$ on a constant, the forecast, and the lagged value of inflation and the indicator variable. A test that all parameters but the constant are equal to zero yields a p-value of 0.44, and thus we fail to reject forecast optimality. For the second test we apply the test of Christoffersen (1998), who proposes testing the null that the indicator variable is *iid Bernoulli* with some constant success probability q_h^*, against an alternative that the indicator variable is first-order Markov. This test yields a p-value of 0.79, and so we again fail to reject the optimality of the inflation forecasts.

Rather than imposing assumption D2 we may instead test optimality using the flexible loss function models presented in Section 3.3. We use a linear spline model for $\partial L/\partial \hat{y}$, initially imposing assumption L5. We employ three nodes, $[-0.5, 0, 0.5]$, which correspond to the 0.44, 0.67 and 0.90 quantiles of the empirical distribution of inflation forecast errors. Assumption L4 can be satisfied by noting that with just three nodes we must have $\gamma_i \geq 0$ for all i for λ to be admissible (with more nodes some γ_i may be negative). We normalize the function by imposing $\sum_{i=1}^{K+1} \gamma_i = 1$. As instruments for the moment conditions we use a constant, the contemporaneous value of the forecast, and two lags each of the forecast error and actual inflation. Thus we have six moment conditions and three free parameters. The resulting estimated loss function and derivative are presented in the top panel of Figure 2, along with what would be obtained under MSE loss. The test of forecast optimality with respect to *some* loss function in the class spanned by our approximation in equation (21) is conducted as a test of over-identifying restrictions and yielded a test statistic (p-value) of 0.99 (0.80) . Thus we cannot reject the null that all moment conditions are satisfied at the optimal parameter. This is consistent with forecast optimality under the estimated loss function.

To examine whether the level of inflation (or, equivalently, the forecast of inflation) affects $\partial L/\partial \hat{y}$, we also estimate the more general specification in equation (22). This specification involves three more parameters than the benchmark model, and so we included three more instruments: one additional lag of the level of inflation and two additional lags of the forecast error. Not surprisingly, the test of forecast optimality based on this more general model also fails to reject the null of forecast

optimality (test statistic and p-value of 2.92 and 0.40). More interesting, though, is the resulting estimated loss function, which is displayed in the second and third panels of Figure 2. Only when inflation is low by historic standards (when it is equal to its 0.10 quantile) is the best-fitting loss function approximately symmetric. For inflation equal to its unconditional median or 0.90 quantile the best-fitting loss function is asymmetric, penalizing positive forecast errors (under-predictions) more heavily than negative forecast errors (over-predictions).

4.2 Output Growth Forecasts

We next turn to the forecasts of quarterly nominal output growth over the period 1968 to 2001, yielding 133 observations. The unconditional mean (t-statistic) of the forecast errors over this period is 0.17 (3.20), which is again significant at the 1% level, though of the opposite sign to the previous example. This suggests that these forecasts are not optimal under MSE loss. Engle's LM test for ARCH in the forecast errors, using four lags, yielded a p-value of less than 0.01, i.e. strong evidence of heteroskedasticity. Thus assumption D2 does not apply, but assumption D3, which allows for dynamic conditional means and variances but rules out any dynamics in higher-order moments, may be reasonable. If we further assume that the loss function satisfies assumption L5" then we know that the optimal forecast error will be of the following form:

$$\begin{aligned} e^*_{t+1,t} &= \gamma_1^* \sigma_{t+1,t} + \sigma_{t+1,t}\eta_{t+1} \\ \eta_{t+1}|\mathcal{F}_t &\sim F_\eta(0,1) \end{aligned} \tag{25}$$

If we knew $\sigma_{t+1,t}$ we could construct $d_{t+1,t} \equiv e_{t+1,t}/\sigma_{t+1,t}$ which we showed in Proposition 6 to be serially uncorrelated and homoskedastic. In order to implement a test we propose imposing assumption D3 and modelling $\sigma^2_{t+1,t}$ as a simple GARCH(1,1) process, allowing for the GARCH-in-mean effects implied by Proposition 6.[13] This makes this test of forecast optimality a *joint* test of forecast optimality and correct volatility model specification. Furthermore, the estimation error

[13] For some variables, such as returns on some stocks or exchange rates, a nonparametric measure of the volatility may be available via the "realized variance" estimator, see Andersen *et al.* (2003) for example. Unfortunately, the data required for this estimator is not available for most macroeconomic time series, including the two under analysis here. Further, general nonparametric estimation of a conditional volatility model for this time series is not feasible given the short sample size available. We use instead a simple, parsimonious, GARCH model, which has been shown to work well in numerous other studies of macroeconomic and financial time series. Of course, the possibility remains that this model is mis-specified and that this affects our conclusions.

in the volatility model must be taken into account when conducting inference. The GARCH-in-mean term was significant (with a p-value of less than 0.01). As this term is $\mathcal{F}_t$-measurable this violates the zero predictability property of optimal forecast errors under MSE loss. Simple tests for serial correlation in higher powers of the standardized forecast error, presented in Table 1, support assumption D3 for this series.

Given the short sample size available for this data set and relative complexity of the model being estimated, we used simulations rather than asymptotic theory to obtain a test of forecast optimality with estimated conditional variance. We estimate the small sample distributions of $\widehat{Corr}\left[d^*_{t+1,t}, d^*_{t,t-1}\right]$, $\widehat{Corr}\left[d^{*2}_{t+1,t}, d^{*2}_{t,t-1}\right]$ and $\widehat{Corr}\left[d^*_{t+1,t}, d^*_{t,t-1}\right]^2 + \widehat{Corr}\left[d^{*2}_{t+1,t}, d^{*2}_{t,t-1}\right]^2$ under the null that the forecast error satisfies equation (25) and that the conditional volatility follows a GARCH(1,1) process with normal innovations. We test the optimality of the forecasts of output growth by comparing the observed values for these three statistics with their simulated distributions under the null of forecast optimality. We generate the simulated distributions by simulating a data set of the same length as the original one using the estimated parameters, re-estimating the GARCH-in-mean model on the simulated data set, and computing $\widehat{Corr}\left[d^*_{t+1,t}, d^*_{t,t-1}\right]$, $\widehat{Corr}\left[d^{*2}_{t+1,t}, d^{*2}_{t,t-1}\right]$ and $\widehat{Corr}\left[d^*_{t+1,t}, d^*_{t,t-1}\right]^2 + \widehat{Corr}\left[d^{*2}_{t+1,t}, d^{*2}_{t,t-1}\right]^2$. We repeat the simulation 10,000 times. The observed values of all three of these statistics lie within their 95% confidence bounds (with p-values of 0.47, 0.85, and 0.68 respectively) and we conclude that while these forecasts are not optimal under MSE loss, we have no evidence that they are not optimal for some other unknown loss function satisfying assumption L5".[14]

The two quantile-based tests of forecast optimality presented in Proposition 8 are also applied. In the first test we regress the indicator variable $\mathbf{1}\left(Y_{t+1} \leq \hat{Y}_{t+1}\right)$ on a constant, the forecast, and the lagged value of GDP growth and the indicator variable. The test that all parameters but the constant are equal to zero yields a p-value of 0.12, so we fail to reject forecast optimality. The test of Christoffersen (1998) yields a p-value of 0.98, and so we again fail to reject the optimality of the GDP growth forecasts.

We again use a continuous linear spline model for $\partial L/\partial \hat{y}$, with nodes equal to $[-0.5, 0, 0.5]$, which correspond to the 0.14, 0.42 and 0.73 quantiles of the empirical distribution of GDP forecast

[14] We also estimated the small sample distribution of these statistics assuming a standardised Student's t distribution for the innovations, with degrees of freedom set to 6 and 10. The p-values on the test statistics changed by less than 0.03 in all cases.

errors. As instruments for the moment conditions we use a constant, the contemporaneous value of the forecast, and two lags each of the forecast error, actual GDP growth and the generalized forecast error, so we have eight moment conditions and three free parameters. The estimated loss function for GDP growth, and its derivative, are presented in the top panel of Figure 3. The test of forecast optimality yields a test statistic (p-value) of 7.05 (0.22). Thus we cannot reject the null that the GDP growth forecasts are optimal under the estimated loss function. A test that the estimated loss function is equal to MSE loss yields a p-value of 0.01, indicating a rejection of this restriction at the 5% level.

We finally estimated the more general loss function model, (22), using as additional instruments one additional lag of the level of GDP growth, the forecast error and the generalized forecast error. Again, the test of forecast optimality based on this more general model failed to reject the null of forecast optimality (test statistic and p-value of 5.69 and 0.34). Further, a test of the restrictions that this more general loss function is equal to the simpler loss function presented in the top panel of Figure 3, or equal to MSE loss, leads to p-values of less than 0.01 in both cases, indicating that this more general loss function is significantly different from both. The estimated loss function and an alternative view are presented in the middle and lower panels of Figure 3. This figure shows that the type of asymmetry exhibited by the loss function changes depending on the level of GDP growth and the size of the forecast error. For small forecast errors (those less than about 0.75 in absolute value) the lower panel of Figure 3 shows that negative forecast errors (i.e., over-predictions) are more heavily penalized than positive forecast errors (under-predictions). However, for large forecast errors (those greater than 0.75) the relative penalty applied to positive or negative forecast errors depends on the realized level of GDP growth: for GDP growth equal to its median the loss function is almost symmetric, penalizing negative forecast errors just slightly more than positive forecast errors. For high GDP growth, the loss function penalizes over-prediction more strongly than under-prediction. For low GDP growth the opposite is true. Assuming that forecast users tend to expand (reduce) economic activity when faced with high (low) forecasts of output growth, a possible explanation for this finding is that adjustment costs are asymmetric: when GDP growth runs high it is particularly costly to over-predict and increase economic activity (e.g., due to over-heating and capacity constraints), while when GDP growth is low, reducing activity is similarly costly (due to costs of layoffs etc.).

5 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 provided intuitive results on properties of a "generalized forecast error" that may be tested when the forecaster's loss function is known, and derived testable implications of forecast optimality under general (but unknown) loss function and restrictions on the data generating process. We also proposed a test of optimality based on a flexible parametric estimate of the unknown loss function.

Finally, 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 $h-1$, i.e., 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, although considerable progress has been made on evaluating forecasts in the presence of such effects, c.f. West (1996), McCracken (2000) and Corradi and Swanson (2002). We have also ignored decision theoretical issues. In general decision problems the forecasting and decision problem cannot be separated and an examination of the decision maker's action rule and full density forecast is required to test rationality, c.f. Diebold, Gunther and Tay (1998). Thus, if the object of the analysis is to derive a decision maker's optimal actions, in general the entire forecast density matters rather than simply the point forecast. However, there are cases where certainty equivalence can still be established for risk-averse or risk-seeking decision makers, c.f. Whittle's (1983) risk-sensitive optimal control method. Alternatively, one can think of situations where the forecast is the decision, as in the case of information services that provide a single forecast for multiple users (e.g., financial analysts or international organizations such as the IMF or OECD).

In our empirical study of the Survey of Professional Forecasters data, consistent with earlier studies, we found strong and significant evidence against forecast optimality under MSE loss. However, when allowing for the possibility of non-MSE loss we found no such evidence. This serves as a concrete example of our claim that previous reports of irrationality or sub-optimality may have relied heavily on the strong assumption that the forecaster's loss function is of the MSE type.

Appendix

Proof of Proposition 1. This proof follows directly from the proof of Proposition 2 below, when one observes the relation between the forecast error and the generalized forecast error, $\psi^*_{t+h,t}$, for the mean squared loss case: $e^*_{t+h,t} = -\frac{1}{2}\psi^*_{t+h,t}$, and noting that the MSE loss function satisfies assumptions L1, L3 and L5' which implies a unique interior optimum. ■

Proof of Proposition 2. 1. Assumptions L1 and L2' allow us to analyze the first-order condition for the optimal forecast, and assumption L3 permits the exchange of differentiation and expectation in the first-order condition, giving us, by the optimality of $\hat{Y}^*_{t+h,t}$,

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

$E\left[\psi^*_{t+h,t}\right] = 0$ follows from the law of iterated expectations.

To prove point 2, since $(Y_t, Y_{t-1}, ...) \subseteq \mathcal{F}_t$ 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 $\mathcal{F}_t$ for all $j \geq h$. Assumptions L1 and L2' again allow us to analyze the first-order condition for the optimal forecast, and assumption L3 permits the exchange of differentiation and expectation in the first-order condition. We thus have

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

which implies $E\left[\psi^*_{t+h,t} \cdot \phi\left(Z_t\right)\right] = 0$ for all $Z_t \in \mathcal{F}_t$ and all functions ϕ for which this moment exists. Thus $\psi^*_{t+h,t}$ is uncorrelated with any function of any element of $\mathcal{F}_t$. This implies that $E\left[\psi^*_{t+h,t} \cdot \psi^*_{t+h-j,t-j}\right] = 0$, for all $j \geq h$, and so $\psi^*_{t+h,t}$ is uncorrelated with $\psi^*_{t+h-j,t-j}$.

To prove point 3, note that assumption (D1) of strict stationarity for $\{Y_{t+h}, Z_t\}$ yields the strict stationarity of $\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)$ since $\hat{Y}^*_{t+h,t}$ is a time-invariant function of Z_t. Thus 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, h, and not on the period when the forecast was made, t. By the optimality of the forecast $\hat{Y}^*_{t+h,t}$ we also have, $\forall j \geq 0$,

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

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. Consider the DGP and loss function used to generate Figure 1: if $\hat{\pi}_{s_t,t} = [0.95, 0.05]'$ it can be shown that the conditional expected loss for $h = 1$ is less than that for $h = 2$; if $\hat{\pi}_{s_t,t} = [0.05, 0.95]'$ the reverse is true. ■

To prove Proposition 3 we prove the following lemma, for the "$\tilde{L}$-loss probability measure", which nests the MSE-loss probability measure as a special case. We will require the following generalization of assumption L6:

Assumption L6': Given two loss functions, L and $\tilde{L}$, $0 < E\left[\left.\frac{\partial L(Y_{t+h},\hat{y})/\partial\hat{y}}{\partial\tilde{L}(Y_{t+h},\hat{y})/\partial\hat{y}}\right| Z^t\right] < \infty$ for all h, all $\hat{y} \in \mathcal{Y}$, and all $Z^t \in \mathcal{Z}^t$.

Lemma 1 *Let L and $\tilde{L}$ be two loss functions, and let $\hat{Y}^*_{t+h,t}$ and $\tilde{Y}^*_{t+h,t}$ be the optimal forecasts of Y_{t+h} at time t under L and $\tilde{L}$ respectively.*

1. Let assumptions L1, L4 and L6' hold for L and $\tilde{L}$. Then the univariate "$\tilde{L}$-loss probability measure", $\tilde{F}_{e_{t+h,t}}$, defined below is a proper probability distribution function.

$$\begin{aligned} d\tilde{F}_{e_{t+h,t}}(e;\hat{y}) &= \frac{\Lambda(e,\hat{y})}{E_t\left[\Lambda\left(Y_{t+h}-\hat{y},\hat{y}\right)\right]} \cdot dF_{e_{t+h,t}}(e;\hat{y}) \\ \text{where } \Lambda(e,\hat{y}) &\equiv \frac{\partial L(y,\hat{y})/\partial\hat{y}|_{y=\hat{y}+e}}{\partial\tilde{L}(y,\hat{y})/\partial\hat{y}\Big|_{y=\hat{y}+e}} \equiv \frac{\psi(\hat{y}+e,\hat{y})}{\tilde{\psi}(\hat{y}+e,\hat{y})} \end{aligned}$$

2. If we further let assumption L2' hold, then the generalized forecast error under $\tilde{L}$ evaluated at $\hat{Y}^*_{t+h,t}$, $\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right) = \partial\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)/\partial\hat{y}$, has conditional mean zero under the $\tilde{L}$-loss probability measure.

3. The generalized forecast error under $\tilde{L}$, evaluated at $\hat{Y}^*_{t+h,t}$, is serially uncorrelated under the $\tilde{L}$-loss probability measure for all lags greater than $h-1$.

4. $\tilde{E}\left[\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right]$ is non-decreasing as a function of the forecast horizon when evaluated at $\hat{Y}^*_{t+h,t}$.

Proof of Lemma 1. We first need to show that $d\tilde{F}_{e_{t+h}} \geq 0$ for all possible values of e, and that $\int d\tilde{F}_{e_{t+h},t}(u;\hat{y})\,du = 1$. By assumption L4 we have $\Lambda(e,\hat{y}) > 0$ for all e where $\Lambda(e,\hat{y})$ exists. (Note that assumption L6' implies that $\Lambda \cdot dF_{e_{t+h},t}$ exists for all e and all $\hat{y}$.) Thus $\Lambda \cdot dF_{e_{t+h},t}$ is non-negative, and

$E_t[\Lambda]$ is positive (and finite by assumption L6'), so $d\tilde{F}_{e_{t+h,t}}\left(e;\hat{Y}_{t+h,t}\right) \geq 0$, if $dF_{e_{t+h,t}}\left(e;\hat{Y}_{t+h,t}\right) \geq 0$. By the construction of $d\tilde{F}_{e_{t+h,t}}$ it is clear that it integrates to 1.

To prove part 2, note that, from the optimality of $\hat{Y}^*_{t+h,t}$ under L,

$$\begin{aligned}\tilde{E}_t\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right] &\propto \int \tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\Lambda\left(e,\hat{Y}^*_{t+h,t}\right)\cdot dF_{e_{t+h,t}}\left(e;\hat{Y}^*_{t+h,t}\right)\\ &= \int \psi\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\cdot dF_{e_{t+h,t}}\left(e;\hat{Y}^*_{t+h,t}\right)\\ &= 0.\end{aligned}$$

The unconditional mean of $\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)$ is also zero by the law of iterated expectations.

In the proof of part 3 we make reference to the bivariate $\tilde{L}$-loss probability measure, but do not need to explicitly define it in order to obtain the result. Since $\tilde{E}\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right]=0$, we need only show that $\tilde{E}\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\cdot\tilde{\psi}\left(Y_{t+h+j},\hat{Y}^*_{t+h+j,t+j}\right)\right]=0$ for $j\geq h$. By part 2,

$$\begin{aligned}&\tilde{E}_t\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\cdot\tilde{\psi}\left(Y_{t+h+j},\hat{Y}^*_{t+h+j,t+j}\right)\right]\\ =\ &\tilde{E}_t\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\cdot E_{t+j}\left[\tilde{\psi}\left(Y_{t+h+j},\hat{Y}^*_{t+h+j,t+j}\right)\right]\right]\ \text{ for } j\geq h\\ =\ &0.\end{aligned}$$

$\tilde{E}\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\cdot\tilde{\psi}\left(Y_{t+h+j},\hat{Y}^*_{t+h+j,t+j}\right)\right]=0$ follows by the law of iterated expectations.

For part 4 note that $\tilde{E}_t\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right]=0$ is the first-order condition of $\min_{\hat{y}}\tilde{E}_t\left[\tilde{L}\left(Y_{t+h},\hat{y}\right)\right]$, so $\tilde{E}_t\left[\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right] \leq \tilde{E}_t\left[\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t-j}\right)\right]\ \forall\ j\geq 0$, and so $\tilde{E}\left[\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right] \leq \tilde{E}\left[\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t-j}\right)\right]=\tilde{E}\left[\tilde{L}\left(Y_{t+h+j},\hat{Y}^*_{t+h+j,t}\right)\right]$ by the law of iterated expectations and the assumption of strict stationarity. Note that the assumption of strict stationarity for $\{Y_{t+h},Z_t\}$ suffices here since $\hat{Y}^*_{t+h,t}$ and the change of measure, $\tilde{\Lambda}_{t+h,t}\left(e,\hat{Y}^*_{t+h,t}\right)$, are time-invariant functions of Z_t. ■

Proof of Proposition 3. Follows from the proof of Lemma 1 setting $\tilde{L}(y,\hat{y})=(y-\hat{y})^2$ and noting that assumption L6 satisfies L6' for this loss function. ■

Proof of Proposition 4. Under assumptions L2 and L5 the optimal forecast may be written as $\hat{Y}^*_{t+h,t}=\mu_{t+h,t}+\alpha^*_{t+h,t}$, so that the optimal forecast error is $e^*_{t+h,t}=Y_{t+h}-\hat{Y}^*_{t+h,t}=\varepsilon_{t+h}-\alpha^*_{t+h,t}$, where $\alpha^*_{t+h,t}$ solves $\min_{\alpha_{t+h,t}}\int L(\varepsilon_{t+h}-\alpha_{t+h,t})dF_{\varepsilon,h}$. By assumption D2, $F_{\varepsilon,h}$ only depends on h and not on t, so we have $\alpha^*_{t+h,t}=\alpha^*_h$, c.f. Granger (1969) and Christoffersen and Diebold (1997). Since α_h is constant for fixed h, we thus have $e^*_{t+h,t}$ is independent of all $Z_t\in\mathcal{F}_t$. ■

Proof of Proposition 5. Consider $h > 0$ and $j > 0$. Let

$$\begin{aligned} Y_{t+h+j} &= E_t\left[Y_{t+h+j}\right] + \eta_{t+h+j},\ \eta_{t+h+j}|\mathcal{F}_t \sim F_{\varepsilon,h+j}\left(0, \sigma^2_{\varepsilon,h+j}\right) \\ Y_{t+h+j} &= E_{t+j}\left[Y_{t+h+j}\right] + \varepsilon_{t+h+j},\ \varepsilon_{t+h+j}|\mathcal{F}_t \sim F_{\varepsilon,h}\left(0, \sigma^2_{\varepsilon,h}\right) \end{aligned}$$

Let $\sigma^2_{\varepsilon,h} < \infty$ and further assume that $\sigma^2_{\varepsilon,h+j} < \infty$. Using again that $\hat{Y}^*_{t+h,t} = E_t\left[Y_{t+h}\right] + \alpha_h$, so $e^*_{t+h+j,t} = \eta_{t+h+j} - \alpha_{h+j}$, and $e^*_{t+h+j,t+j} = \varepsilon_{t+h+j} - \alpha_h$,where α_h and α_{h+j} are constants. Thus $V_t\left[e^*_{t+h+j,t}\right] = V_t\left[\eta_{t+h+j}\right] = \sigma^2_{\varepsilon,h+j}$, and $V_t\left[e^*_{t+h+j,t+j}\right] = \sigma^2_{\varepsilon,h}$, where these moments are independent of t by assumption D2. Note also that $V\left[e^*_{t+h+j,t}\right] = E\left[E_t\left[\eta^2_{t+h+j}\right]\right] = \sigma^2_{\varepsilon,h+j}$, and similarly $V\left[e^*_{t+h+j,t+j}\right] = \sigma^2_{\varepsilon,h}$. Now we seek to show that $\sigma^2_{\varepsilon,h+j} \geq \sigma^2_{\varepsilon,h}$.

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

The first equality follows from the equality of the conditional and unconditional variance of the forecast error under D2; the third equality follows from the fact that $E_t\left[Y_{t+h+j}\right]$ is constant given $\mathcal{F}_t$; the weak inequality follows from the non-negativity of $V_t\left[E_{t+j}\left[Y_{t+h+j}\right]\right]$ and $E_{t+j}\left[\varepsilon_{t+h+j} \cdot \phi\left(Z_{t+j}\right)\right] = 0$; the final equality follows from the fact that $F_{\varepsilon,h}$ does not change with t. The cases where $h = 0$ and/or $j = 0$ are trivial. Thus $V\left[e^*_{t+h+j,t}\right] \geq V\left[e^*_{t+h,t}\right]\ \forall\, h,\ j \geq 0$ if $V\left[e^*_{t+h+j,t}\right] < \infty$. If $\sigma^2_{\varepsilon,h} < \infty$ but $\sigma^2_{\varepsilon,h+j}$ is infinite then the proposition holds trivially. ■

Proof of Proposition 6. Part 1: By homogeneity,

$$\begin{aligned} \hat{Y}^*_{t+h,t} &\equiv \arg\min_{\hat{y}} \int L\left(y - \hat{y}\right) dF_{t+h,t}\left(y\right) \\ &= \arg\min_{\hat{y}} \int g\left(\frac{1}{\sigma_{t+h,t}}\right) L\left(\frac{1}{\sigma_{t+h,t}}\left(y - \hat{y}\right)\right) dF_{t+h,t}\left(y\right) \\ &= \arg\min_{\hat{y}} \int L\left(\frac{1}{\sigma_{t+h,t}}\left(y - \hat{y}\right)\right) dF_{t+h,t}\left(y\right) \\ &= \arg\min_{\hat{y}} \int L\left(\frac{1}{\sigma_{t+h,t}}\left(\mu_{t+h,t} + \sigma_{t+h,t}\eta_{t+h} - \hat{y}\right)\right) dF_{\eta,h}\left(\eta\right) \end{aligned}$$

Let us represent a forecast as $\hat{Y}_{t+h,t} = \mu_{t+h,t} + \sigma_{t+h,t} \cdot \hat{\gamma}_{t+h,t}$, so

$$\begin{aligned}
\hat{Y}^*_{t+h,t} &= \mu_{t+h,t} + \sigma_{t+h,t} \cdot \arg\min_{\hat{\gamma}} \int L\left(\frac{1}{\sigma_{t+h,t}}\left(\mu_{t+h,t} + \sigma_{t+h,t}\eta_{t+h} - \mu_{t+h,t} - \sigma_{t+h,t}\hat{\gamma}\right)\right) dF_{\eta,h}(\eta) \\
&= \mu_{t+h,t} + \sigma_{t+h,t} \cdot \arg\min_{\hat{\gamma}} \int L\left(\eta_{t+h} - \hat{\gamma}\right) dF_{\eta,h}(\eta) \\
&= \mu_{t+h,t} + \sigma_{t+h,t} \cdot \gamma^*_h,
\end{aligned}$$

where the last line follows from the fact that $F_{\eta,h}$ is time-invariant under assumption D3.

Part 2: This result follows from noting that $d^*_{t+h,t} = \eta_{t+h} - \gamma^*_h$, where γ^*_h is a constant and, by assumption D3, η_{t+h} is independent of all elements in $\mathcal{F}_t$ and has unit variance.

Part 3: Following the steps in the proof of Part 1, we find

$$\begin{aligned}
\hat{Y}^*_{t+h,t} &= \mu_{t+h,t} + \arg\min_{\hat{\gamma}} \int L\left(\sigma_{t+h,t}\left(\eta_{t+h} - \hat{\gamma}\right)\right) dF_{\eta,h}(\eta) \\
&\equiv \mu_{t+h,t} + \gamma^*\left(\sigma^2_{t+h,t}, L, h\right).
\end{aligned}$$

That is, $\gamma^*_{t+h,t} \equiv \gamma^*\left(\sigma^2_{t+h}, L, h\right)$. ■

Proof of Proposition 7. Under assumptions D3', L2 and L5" we have from Proposition 6 that $\hat{Y}^*_{t+h,t} = \sigma_{t+h,t}(\beta + \gamma^*_h)$. Thus $d^*_{t+h,t} \equiv e^*_{t+h,t}/\hat{Y}^*_{t+h,t} = \left(\eta_{t+h} - \gamma^*_h\right)/(\beta + \gamma^*_h)$, i.e., an affine transformation of η_{t+h}. The result follows by noting that η_{t+h} is independent of all $Z_t \in \mathcal{F}_t$. ■

Proof of Proposition 8. 1. Under assumptions D2, L2 and L5, or assumptions D3, L2 and L5", we know from above that

$$Y^*_{t+h,t} = \mu_{t+h,t} + \sigma_{t+h,t} \cdot \gamma^*_h$$

with $\sigma_{t+h,t}$ constant under assumption D2. γ^*_h depends only upon the loss function and the forecast horizon. Now notice that $F_{t+h,t}\left(\hat{Y}^*_{t+h,t}\right) \equiv \Pr\left[Y_{t+h} \leq \hat{Y}^*_{t+h,t}|\mathcal{F}_t\right] = \Pr\left[\mu_{t+h,t} + \sigma_{t+h,t}\eta_{t+h} \leq \mu_{t+h,t} + \sigma_{t+h,t} \cdot \gamma_h|\mathcal{F}_t\right] = \Pr\left[\eta_{t+h} \leq \gamma_h|\mathcal{F}_t\right] \equiv q^*_h \ \forall\ t$. Thus $\hat{Y}^*_{t+h,t}$ is the q^*_h conditional quantile of $Y_{t+h}|\mathcal{F}_t \ \forall\ t$. Note that q^*_h is only a function of the loss function and the forecast horizon.

2. Since $I^*_{t+h,t}$ is a binary random variable and $\Pr\left[I^*_{t+h,t} = 1|\mathcal{F}_t\right] = \Pr\left[Y_{t+h} \leq \hat{Y}^*_{t+h,t}|\mathcal{F}_t\right] = q^*_h \ \forall\ t$, we thus have that $I^*_{t+h,t}$ is independent of all $Z_t \in \mathcal{F}_t$. ■

Proof of Proposition 9. Following the steps in the proof of Proposition 6 we find:

$$\begin{aligned}
\hat{Y}^*_{t+h,t} &= \mu_{t+h,t} + \sigma_{t+h,t} \cdot \arg\min_{\hat{\gamma}} \int L\left(\eta_{t+h} - \hat{\gamma}\right) dF_{\eta,t+h,t}(\eta) \\
&= \mu_{t+h,t} + \sigma_{t+h,t} \cdot \gamma^*_{t+h,t}.
\end{aligned}$$

Hence $\gamma^*_{t+h,t}$ will be a function of the loss function and $F_{\eta,t+h,t}$, the latter depending on time-varying properties of the conditional distribution of $Y_{t+h}|\mathcal{F}_t$ beyond the conditional mean and variance. ■

References

[1] Andersen, T.G., T. Bollerslev, F.X. Diebold and P. Labys, 2003, Modeling and Forecasting Realized Volatility, Econometrica, 71, 579-625.

[2] Andrews, D.W.K., 1991, Asymptotic Normality of Series Estimators for Nonparametric and Semi-parametric Regression Models, Econometrica, 59, 307-346.

[3] Andrews, D.W.K., 2001, Testing When a Parameter is on the Boundary of the Maintained Hypothesis, Econometrica, 69, 683-734.

[4] Bierens, H.J., 1990, A Consistent Conditional Moment Test of Functional Form, Econometrica, 58, 1443-1458.

[5] Bierens, H.J., and Ploberger, W., 1997, Asymptotic Theory of Integrated Conditional Moment Tests, Econometrica, 65, 1129-1151.

[6] Chen, X. and X. Shen, 1998, Sieve Extremum Estimates for Weakly Dependent Data, Econometrica, 66, 289-314.

[7] Chesher, A. and M. Irish, 1987, Residual Analysis in the Grouped and Censored Normal Linear Model, Journal of Econometrics, 34, 33-61.

[8] Christoffersen, P.F., 1998, Evaluating Interval Forecasts, International Economic Review, 39, 841-862.

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

[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 Jong, R.M., 1996, The Bierens Test Under Data Dependence, Journal of Econometrics, 72, 1-32.

[12] Diebold, F.X., T. Gunther and A. Tay, 1998, Evaluating Density Forecasts, with Applications to Financial Risk Management. International Economic Review 39, 863-883.

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

[14] Engle, R.F., 1982, Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation, Econometrica, 50, 987-1008.

[15] Engle, R.F., D.M. Lilien and R.P. Robins, 1987, Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model, Econometrica, 55, 391-407.

[16] Gallant, A.R., and D.W. Nychka, 1987, Semi-Nonparametric Maximum Likelihood Estimation, Econometrica, 55, 363-390.

[17] Gourieroux, C., A. Monfort, E. Renault and A. Trongnon, 1987, Generalized Residuals, Journal of Econometrics, 34, 5-32.

[18] Granger, C.W.J., 1969, Prediction with a Generalized Cost Function. OR 20, 199-207.

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

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

[21] Hansen, P.R. and A. Lunde, 2003, Consistent Ranking of Volatility Models, Journal of Econometrics, forthcoming.

[22] Harrison, J.M. and D.M. Kreps, 1979, Martingales and Arbitrage in Multiperiod Securities Markets, Journal of Economic Theory, 20, 381-408.

[23] Machina, M.J., and C.W.J. Granger, 2004, Decision-Based Loss Functions in Forecasting and Estimation. Mimeo, University of California, San Diego.

[24] McCracken, M., 2000, Robust Out of Sample Inference, Journal of Econometrics 99, 195-223.

[25] Mincer, J., and V. Zarnowitz, 1969, The Evaluation of Economic Forecasts, in J. Mincer (ed.) Economic Forecasts and Expectations, National Bureau of Economic Research, New York.

[26] Newey, W.K. and K.D. West, 1987, A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, Econometrica, 55, 703-708.

[27] Newey, W.K., and D. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, in R.F. Engle and D. McFadden, (eds.) Handbook of Econometrics, Vol IV, 2111-2245, North-Holland, Amsterdam.

[28] Patton, A. J., and A. Timmermann, 2004, Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity, working paper, Financial Markets Group, London School of Economics.

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

[30] Shephard, N., 2004, Stochastic Volatility: Selected Readings, Oxford University Press, forthcoming.

[31] 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.

[32] West, K.D., 1996, Asymptotic Inference about Predictive Ability. Econometrica 64, 1067-84.

[33] 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.

[34] White, H.L., 1980, A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, Econometrica, 48, 817-838.

[35] Whittle, P., 1983, Prediction and Regulation by Linear Least Square Methods. Second Edition, Revised. University of Minnesota Press, Minneapolis, MN.

.

Table 1
Tests for higher-order dynamics

	Inflation forecast errors	
Power	1 lag	4 lags
2	0.674	0.971
3	0.781	0.962
4	0.660	0.951
	Standardized output growth forecast errors	
Power	1 lag	4 lags
3	0.506	0.748
4	0.745	0.945

Notes: this table presents p-values on Ljung-Box tests of the hypothesis of zero serial correlation up to k lags, where $k = 1$ or 4. The dependent variables are the second through fourth powers of inflation forecast errors, and the third and fourth powers of standardized output growth forecast errors.

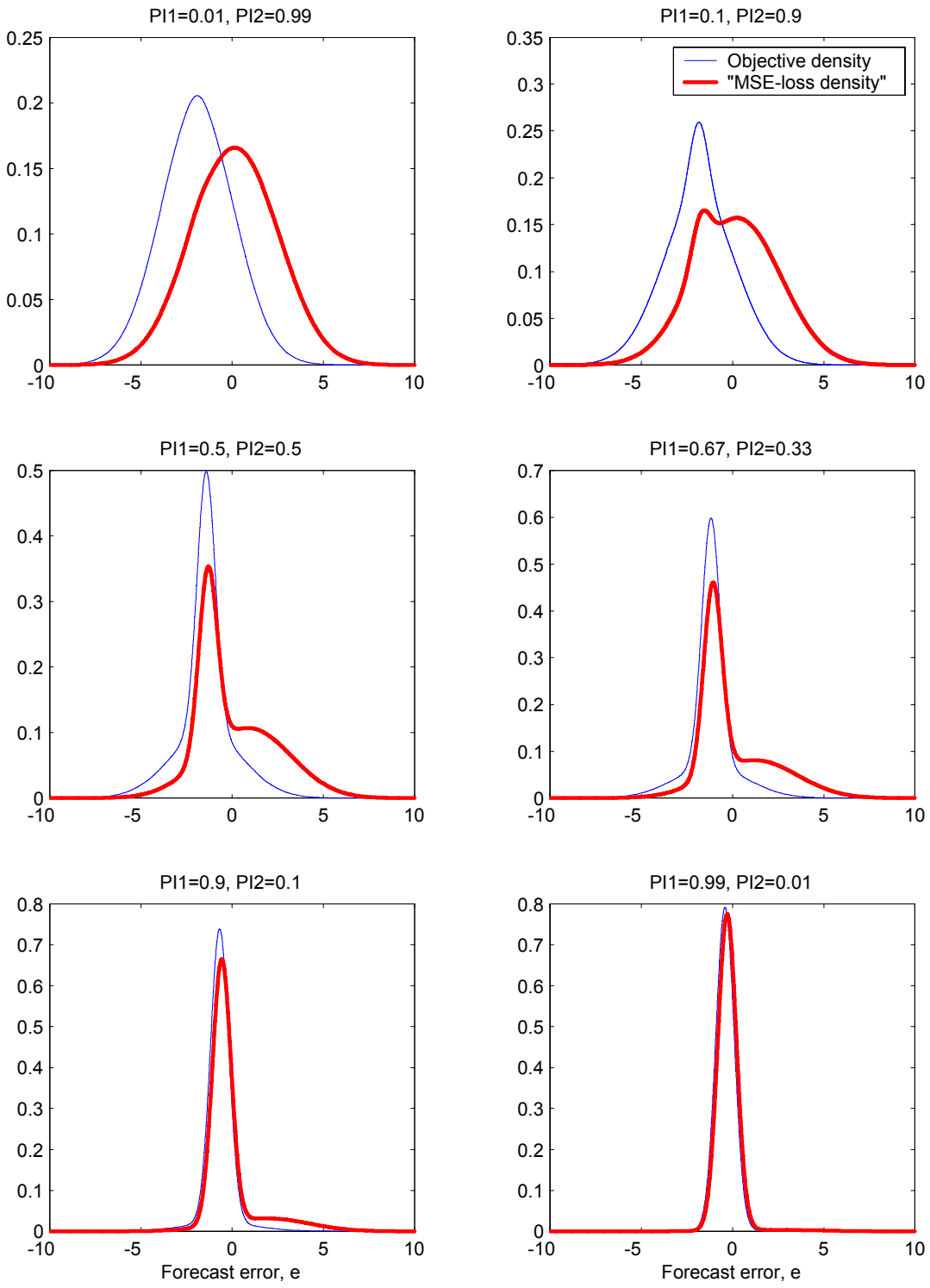

Figure 1: *Objective and "MSE-loss" error densities for a regime switching process, one-step forecast horizon, for various values of the state probability vector,* $\hat{\boldsymbol{\pi}}_{s_t,t}$.

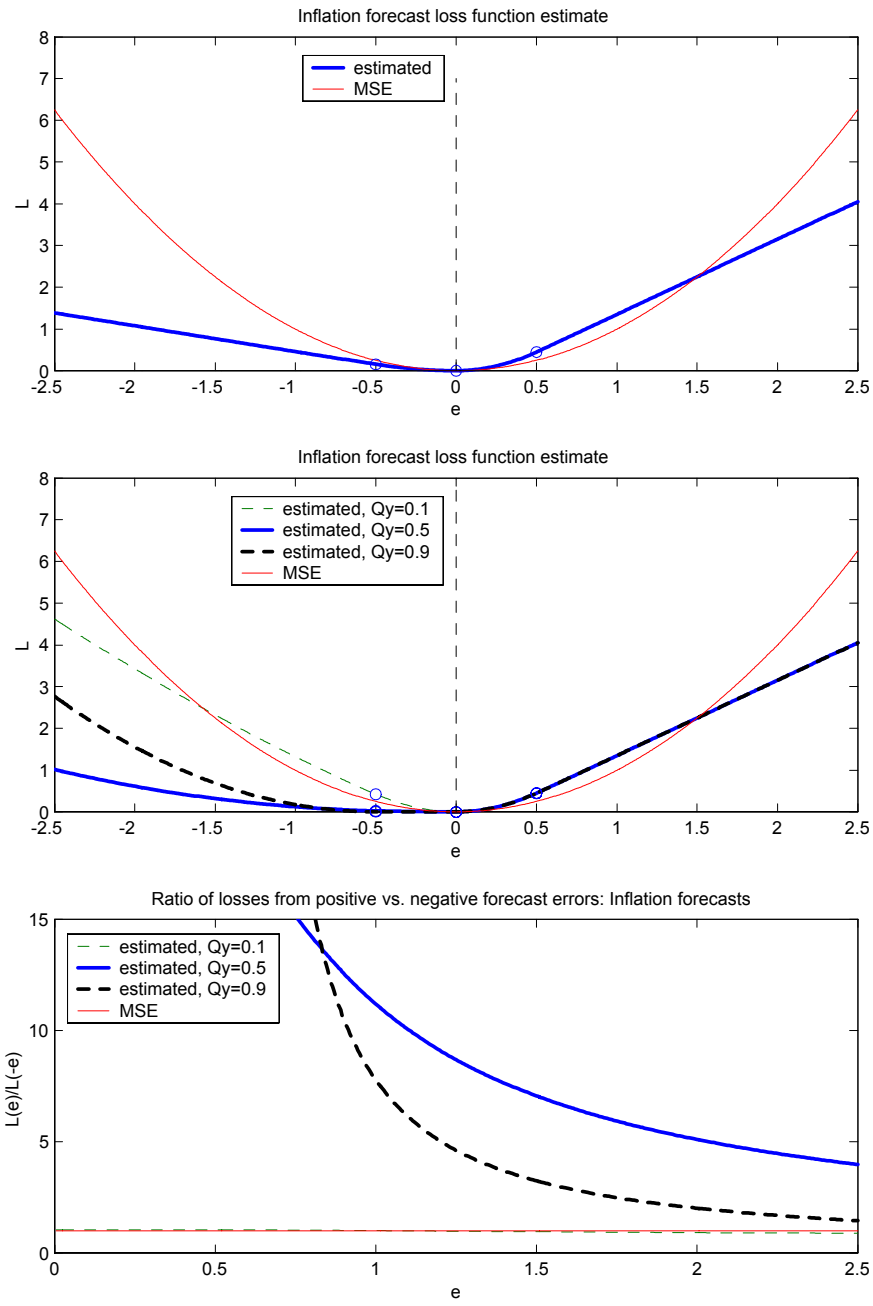

Figure 2: *Estimates of the loss function for inflation based on linear splines with nodes* $[0.5, 0, 0.5]$. *The top panel is the estimate obtained under assumption L5. The second panel is the estimate obtained when allowing the level of inflation to also affect the loss function; the estimated loss function is evaluated for inflation equal to its 0.1, 0.5 and 0.9 quantiles. The lower panel is the ratio of the estimated loss function evaluated at* e *and* $-e$, *for inflation equal to its 0.1, 0.5 and 0.9 quantiles.*

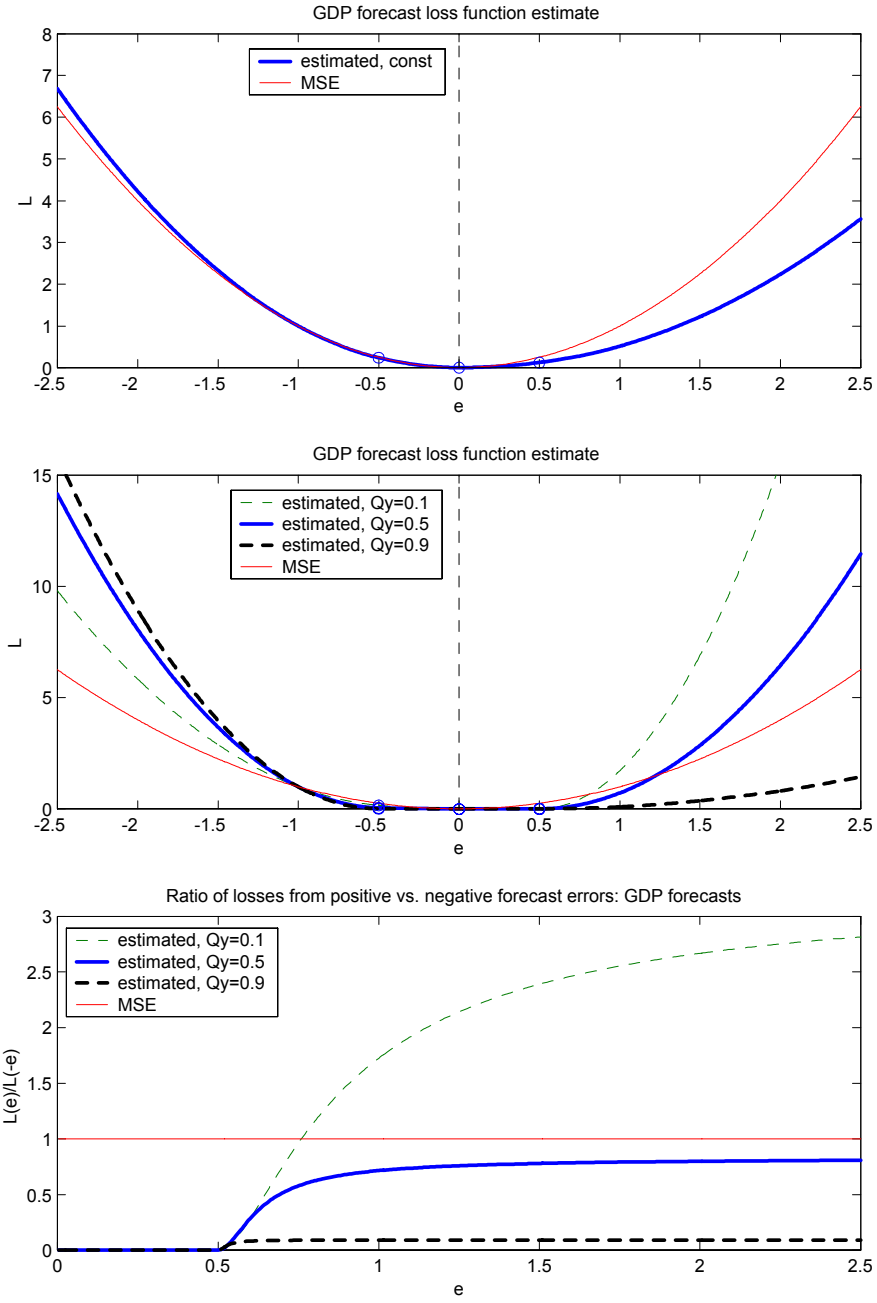

Figure 3: *Estimates of the loss function for GDP growth based on linear splines with nodes* $[0.5, 0, 0.5]$. *The top panel is the estimate obtained under assumption L5. The second panel is the estimate obtained when allowing the level of GDP growth to also affect the loss function; the estimated loss function is evaluated for GDP growth equal to its 0.1, 0.5 and 0.9 quantiles. The lower panel is the ratio of the estimated loss function evaluated at* e *and* $-e$, *for GDP growth equal to its 0.1, 0.5 and 0.9 quantiles.*