Generalized Forecast Errors, A Change of Measure, and Forecast Optimality*

Andrew J. Patton
University of Oxford

Allan Timmermann
University of California, San Diego and CREATES

17 July 2008

This paper establishes properties of optimal forecasts under general loss functions, extending existing results obtained under specific functional forms and data generating processes. We propose a new method that changes the probability measure under which the well-known properties of optimal forecasts under mean squared error loss can be recovered. We illustrate the proposed methods through an empirical application to U.S. inflation forecasting.

Keywords: forecast evaluation, loss function, rationality tests.
J.E.L. Codes: C53, C22, C52

*The authors would like to thank seminar participants at the Festschrift Conference in Honor of Robert F. Engle in San Diego, June 2007, and Graham Elliott, Raffaella Giacomini, Clive Granger, Oliver Linton, Mark Machina, Francisco Penaranda, Kevin Sheppard, Mark Watson, Hal White, Stanley Zin and an anonymous referee for useful comments. All remaining deficiencies are the responsibility of the authors. The second author acknowledges support from CREATES, funded by the Danish National Research Foundation. Patton: Department of Economics and Oxford-Man Institute of Quantitative Finance, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom. Email: andrew.patton@economics.ox.ac.uk. Timmermann: Rady School of Managment and Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla CA 92093-0553, U.S.A. Email: atimmerm@ucsd.edu.

1 Introduction

In a world with constant volatility, concerns about the possibility of asymmetric or non-quadratic loss functions in economic forecasting would (almost) vanish: Granger (1969) showed that in such an environment optimal forecasts will generally equal the conditional mean of the variable of interest, plus a simple constant (an optimal bias term). However, the pioneering and pervasive work of Rob Engle provides overwhelming evidence of time-varying volatility in many macroeconomic and financial time series.[1] In a world with time-varying volatility, asymmetric loss has important implications for forecasting, see Christoffersen and Diebold (1997), Granger (1999) and Patton and Timmermann (2007a).

The traditional assumption of a quadratic and symmetric loss function underlying most of the work on testing forecast optimality is increasingly coming under critical scrutiny, and evaluation of forecast efficiency under asymmetric loss functions has recently gained considerable attention in the applied econometrics literature.[2] Progress has also been made on establishing theoretical properties of optimal forecasts for particular families of loss functions (Christoffersen and Diebold (1997), Elliott, et al. (2005, 2008), Patton and Timmermann (2007b)). However, while some results have been derived for certain classes of loss functions, a more complete set of results has not been established.

Our paper fills this lacuna in the literature by deriving properties of an optimal forecast that hold for general classes of loss functions and general data-generating processes. Working out these properties under general loss is important since none of the standard properties established in the linear-quadratic framework survives to a more general setting in the presence of conditional heteroskedasticity, cf. Patton and Timmermann (2007a). Irrespective of the loss function and data generating process, a generalized orthogonality principle must, however, hold provided information is efficiently embedded in the forecast. Implications of this principle will, of course, vary significantly with assumptions about the loss function and data generating process (DGP). Our results suggest two approaches: transforming the forecast error for a given loss function, or transforming the

[1] See, amongst many others, Engle (1982, 2004), Bollerslev (1986), Engle, *et al.* (1990), the special issue of the *Journal of Econometrics* edited by Engle and Rothschild (1992), as well as surveys by Bollerslev, *et al.* (1994) and Andersen, *et al.* (2006).

[2] See, for example, Christoffersen and Diebold (1996), Pesaran and Skouras (2001), Christoffersen and Jacobs (2004) and Granger and Machina (2006).

density under which the forecast error is being evaluated.

The first approach provides tests that generalize the widely-used Mincer-Zarnowitz (1969) regressions, established under mean squared error (MSE) loss, to hold for arbitrary loss functions. We propose a seemingly unrelated regression (SUR)-based method for testing multiple forecast horizons simultaneously which may yield power improvements when forecasts for multiple horizons are available. This is relevant for survey data such as those provided by the Survey of Professional Forecasters (Philadelphia Federal Reserve) or Consensus Economics as well as for individual forecasts such as those reported by the IMF in the World Economic Outlook.

Our second approach introduces a new line of analysis based on a transformation from the usual probability measure to an "MSE-loss probability measure". Under this new measure, optimal forecasts, from any loss function, 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. This transformation has its roots in asset pricing and "risk neutral" probabilities, see Harrison and Kreps (1979) for example, but to our knowledge has not previously been considered in the context of forecasting.

Relative to existing work, our contributions are as follows. Using the first line of research, we establish population properties for the so-called generalized forecast error which is similar to the score function known from estimation problems. These results build on, extend and formalize results in Granger (1999) as well as in our earlier work (Patton and Timmermann (2007a,b)) and apply to quite general classes of loss functions and data generating processes. Patton and Timmermann (2007b) establish testable implications of simple forecast errors (defined as the outcome minus the predicted value) under forecast optimality, while Patton and Timmermann (2007a) consider the generalized forecast errors but only for more specialized cases such as linex loss with normally distributed innovations. Unlike Elliott et al. (2005), we do not deal with the issue of identification and estimation of the parameters of the forecaster's loss function. The density forecasting results are, to our knowledge, new in the context of the forecast evaluation literature.

The outline of the paper is as follows. Section 2 establishes properties of optimal forecasts under general known loss functions. Section 3 contains the change of measure result, and Section 4 presents empirical illustrations of the results of this paper. Section 5 concludes. An appendix contains technical details and proofs.

2 Testable Implications under General Loss Functions

Suppose that a decision maker is interested in forecasting some univariate time series, $Y \equiv \{Y_t; t = 1, 2, ...\}$, h steps ahead given information at time t, $\mathcal{F}_t$. We assume that $X_t = \left[Y_t, \tilde{Z}_t'\right]'$, where $\tilde{Z}_t$ is a $(m \times 1)$ vector of predictor variables used by the decision maker, and $X \equiv \left\{X_t : \Omega \to \mathbb{R}^{m+1},\ m \in \mathbb{N},\ t = 1, 2, ...\right\}$ is a stochastic process on a complete probability space $(\Omega, \mathcal{F}, P)$, where $\Omega = \mathbb{R}^{(m+1)\infty} \equiv \times_{t=1}^{\infty} \mathbb{R}^{m+1}$, $\mathcal{F} = \mathcal{B}^{(m+1)\infty} \equiv \mathcal{B}\left(\mathbb{R}^{(m+1)\infty}\right)$, the Borel σ-field generated by $\mathbb{R}^{(m+1)\infty}$, and $\mathcal{F}_t$ is the σ-field $\{X_{t-k}; k \geq 0\}$. Y_t is thus adapted to the information set available at time t.[3] We will denote a generic sub-vector of $\tilde{Z}_t$ as Z_t, 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 $e_{t+h,t} = Y_{t+h} - \hat{Y}_{t+h,t}$.[4] 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 realizations and forecasts to the real line, $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$. That is, in general

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

$E_t[.]$ 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 $V(.)$.

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, $\mathcal{A}$, which in general depend on the full distribution forecast of $Y_{t+h}, F_{t+h,t}$. Here we assume that $\mathcal{A}$ depends only on the forecast $\hat{Y}_{t+h,t}$ and we write this as $\mathcal{A}(\hat{Y}_{t+h,t})$. Granger and Machina (2006) show that under certain conditions on the utility function there exists a unique point forecast which leads to the same decision as if a full distribution forecast had been available.

[3] 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 in forecast optimality tests can cause difficulties, as pointed out by Hansen and Lunde (2006) and further studied in Patton (2006).

[4] We focus on point forecasts below, and leave the interesting extension to interval and density forecasting for future research.

2.1 Properties under General Loss Functions

Under general loss the first order condition for the optimal forecast is[5]

$$0 = E_t\left[\frac{\partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}}\right] = \int \frac{\partial L\left(y, \hat{Y}^*_{t+h,t}\right)}{\partial \hat{Y}_{t+h,t}} dF_{t+h,t}(y). \tag{2}$$

This condition can be rewritten using what Granger (1999) refers to as the (optimal) generalized forecast error, $\psi^*_{t+h,t} \equiv \partial L\left(Y_{t+h}, \hat{Y}^*_{t+h,t}\right)/\partial \hat{Y}_{t+h,t}$,[6] so that equation (2) simplifies to

$$E_t[\psi^*_{t+h,t}] = \int \psi^*_{t+h,t} dF_{t+h,t}(y) = 0. \tag{3}$$

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.

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 $X_t \equiv [Y_t, \tilde{Z}_t']'$:

Assumption D1: $\{X_t\}$ *is a strictly stationary stochastic process.*

Note that we do not assume that X_t is continuously distributed and so the results below may apply to forecasts of discrete random variables, such as direction-of-change forecasts or default forecasts. The following properties of the loss function are assumed at various points of the analysis, but not all will be required everywhere.

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: $E_t\left[L\left(Y_{t+h}, \hat{y}\right)\right] < \infty$ *for some* $\hat{y} \in \mathcal{Y}$ *and all* t*, almost surely.*

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

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

[5] 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.

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

exists for all t and h.

Assumption L3: $|E_t\left[\partial L\left(Y_{t+h},\hat{y}\right)/\partial\hat{y}\right]|<\infty$ *for some* $\hat{y}\in\mathcal{Y}$ *and all* t*, almost surely.*

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 L4: *The loss function is a non-monotonic, 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 L4, with the convexity of the loss function being strict). Assumption L3 is required to interchange expectation and differentiation: $\partial E_t\left[L\left(Y_{t+h},\hat{y}\right)\right]/\partial\hat{y}=E_t\left[\partial L\left(Y_{t+h},\hat{y}\right)/\partial\hat{y}\right]$. 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.

The following proposition establishes properties of the generalized forecast error, $\psi^*_{t+h,t}$:

Proposition 1 *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 $(h-1)$.[7]

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.

All proofs are given in the appendix. The above result is useful when the loss function is known, since $\psi^*_{t+h,t}$ can then be calculated directly and employed in generalized efficiency tests that project $\psi^*_{t+h,t}$ on period-t instruments. For example, the martingale difference property of $\psi^*_{t+h,t}$ can be

[7] 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 $MA\left(h-1\right)$ process.

tested by testing $\alpha = \beta = 0$ for all $Z_t \in \mathcal{F}_t$ in the following regression:

$$\psi_{t+h,t} = \alpha + \beta' Z_t + u_{t+h}. \tag{4}$$

The above simple 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). Tests based on generalized forecast errors obtained from a model with estimated parameters can also be conducted, using the methods in West (1996, 2006).

If the same forecaster reported forecasts for multiple horizons we can conduct a joint test of forecast optimality across all horizons. This can be done without requiring that the forecaster's loss function is the same across all horizons, i.e., we allow the one-step ahead forecasting problem to 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:

$$\begin{bmatrix} \psi_{t+1,t} \\ \psi_{t+2,t} \\ \vdots \\ \psi_{t+H,t} \end{bmatrix} = A + BZ_t + u_{t,H} \tag{5}$$

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 seemingly unrelated regressions (SUR) system for the generalized forecast errors:

$$\begin{bmatrix} \psi_{t+1,t} \\ \psi_{t+2,t} \\ \vdots \\ \psi_{t+H,t} \end{bmatrix} = A + B_1 \begin{bmatrix} \psi_{t,t-1} \\ \psi_{t,t-2} \\ \vdots \\ \psi_{t,t-H} \end{bmatrix} + ... + B_J \begin{bmatrix} \psi_{t-J+1,t-J} \\ \psi_{t-J+1,t-J-1} \\ \vdots \\ \psi_{t-J+1,t-J-H+1} \end{bmatrix} + u_{t,H}, \tag{6}$$

and then test $H_0 : A = B = 0$ vs. $H_a : A \neq 0 \cup B \neq 0$.

2.2 Properties under MSE Loss

In the special case of a squared error loss function:

$$L(Y_{t+h}, \hat{Y}_{t+h,t}) = \theta \left(Y_{t+h} - \hat{Y}_{t+h,t} \right)^2, \ \theta > 0, \tag{7}$$

optimal forecasts can be shown to have the standard properties, using the results from Proposition 1. For reference we list these below:

Corollary 1 *Let the loss function be*

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

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

1. The optimal forecast of Y_{t+h} *is* $E_t\left[Y_{t+h}\right]$ *for all forecast horizons* h;

2. The forecast error associated with the optimal forecast has conditional (and unconditional) mean zero;

3. The h*-step forecast error associated with the optimal forecast exhibits zero serial covariance beyond lag* $(h-1)$*;*

Moreover, if we further assume that Y *is covariance stationary, we obtain:*

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

This corollary 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 Mincer-Zarnowitz (1969) regression

$$Y_{t+h} = \alpha + \beta \hat{Y}_{t+h,t} + \varepsilon_{t+h} \tag{8}$$

or equivalently in a regression of forecast errors on current instruments,

$$e_{t+h,t} = \alpha + \beta' Z_t + u_{t+h}. \tag{9}$$

Elliott, Komunjer and Timmermann (2008) show that the estimates of β will be biased when the loss function used to generate the forecasts is of the asymmetric squared loss variety. Moreover, the bias in that case depends on the correlation between the absolute forecast error and the instruments used in the test. It is possible to show that under general (non-MSE) loss the properties of the optimal forecast error listed in Corollary 1 can all be violated; see Patton and Timmermann (2007a) for an example using a regime switching model and the "linex" loss function of Varian (1974).

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. As an alternative approach, 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, cf. 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, $\left(Y_{t+h} - \hat{Y}_{t+h,t}, \mathcal{F}_t\right)$ is a martingale difference sequence when $\hat{Y}_{t+h,t} = \hat{Y}^*_{t+h,t}$, where $\hat{Y}^*_{t+h,t}$ is defined in equation (1). We call the new measure the "MSE-loss probability measure". The resulting method thus suggests an alternative means of evaluating forecasts made using general loss functions.

Note that the conditional distribution of the forecast error, $F_{e_{t+h,t}}$, given $\mathcal{F}_t$ and any forecast $\hat{y} \in \mathcal{Y}$, satisfies

$$F_{e_{t+h},t}\left(e;\hat{y}\right) = F_{t+h,t}\left(\hat{y}+e\right), \tag{10}$$

for all $\left(e, \hat{Y}_{t+h,t}\right) \in \mathbb{R} \times \mathcal{Y}$ where $F_{t+h,t}$ is the conditional distribution of Y_{t+h} given $\mathcal{F}_t$.

To facilitate the change of measure, we make use of the following assumption:

Assumption L5: $\partial L\left(y,\hat{y}\right)/\partial\hat{y} \leq (\geq) \, 0$ *for* $y \geq (\leq)\ \hat{y}$.

Assumption L5 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 $L\left(\hat{y},\hat{y}\right) = 0$, i.e., the loss from a perfect forecast is zero, but this is obviously just a normalization and is not required here.

The sign of $\left(y-\hat{y}\right)^{-1}\partial L\left(y,\hat{y}\right)/\partial\hat{y}$ is negative under assumption L5, and in defining the MSE-loss probability measure we need to further assume that it is bounded and non-zero:

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

Definition 1 *Let assumptions L5 and L6 hold and let*

$$\Lambda(e,\hat{y}) \equiv -\frac{1}{e}\cdot\left.\frac{\partial L(y,\hat{y})}{\partial\hat{y}}\right|_{y=\hat{y}+e} \tag{11}$$

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

$$d\tilde{F}_{e_{t+h,t}}(e;\hat{y}) = \frac{\Lambda(e,\hat{y})}{E_t\left[\Lambda\left(Y_{t+h}-\hat{y},\hat{y}\right)\right]}\cdot dF_{e_{t+h,t}}(e;\hat{y}) \tag{12}$$

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

$$\tilde{\Lambda}_{t+h,t}(e,\hat{y}) \equiv \frac{\Lambda(e,\hat{y})}{E_t\left[\Lambda\left(Y_{t+h}-\hat{y},\hat{y}\right)\right]} \tag{13}$$

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

$$\begin{aligned} MSE &: \lim_{e\to 0}\ \Lambda(e,\hat{y}) = 2 \\ Linex &: \lim_{e\to 0}\ \Lambda(e,\hat{y}) = a^2 \\ PropMSE &: \lim_{e\to 0}\ \Lambda(e,\hat{y}) = 2/\hat{y}^2 \end{aligned}$$

where the *Linex* and *PropMSE* loss functions are defined as $L(y,\hat{y}) = \exp\{ae\} - ae - 1$ and $L(y,\hat{y}) = (y/\hat{y}-1)^2$, respectively. For mean absolute error loss, $L(y,\hat{y}) = |e|$, the limits from both directions diverge, 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 2 *1. Let assumptions L1, L5 and L6 hold. Then the "MSE-loss probability measure", $\tilde{F}_{e_{t+h,t}}(\cdot|\hat{y})$, defined in equation (12) is a proper probability distribution function for all $\hat{y}\in\mathcal{Y}$.*

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 a non-decreasing 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.[8] With consistent estimates of $f_{e_{t+h},t}$ and Λ it is simple to construct an estimator of $\tilde{f}_{e_{t+h},t}$. In recent work, Chernov and Mueller (2007) specify a flexible parametric model for $\tilde{f}_t$ and Λ_t in order to estimate the underlying objective conditional density, f, of forecasters from a variety of macroeconomic surveys. From this density estimate, they are then able to both "bias-correct" the individual forecasts, and compute combination forecasts.

4 Numerical Example and an Application to U.S. Inflation

To illustrate how the MSE-loss error density differs from the objective error density, consider the following simple example. Consider the following AR(1)-GARCH(1,1) data generating process:

$$\begin{aligned} Y_t &= \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t \\ \varepsilon_t &= h_t^{1/2}\nu_t \\ h_t &= \omega + \beta h_{t-1} + \alpha\varepsilon_{t-1}^2 \\ \nu_t|\mathcal{F}_{t-1} &\sim N(0,1). \end{aligned} \tag{14}$$

[8] 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).

Next, consider the simple and analytically tractable "linex" loss function of Varian (1974), scaled by $2/a^2$:

$$L\left(y,\hat{y};a\right)=\frac{2}{a^2}\left(\exp\left\{a\left(y-\hat{y}\right)\right\}-a\left(y-\hat{y}\right)-1\right). \tag{15}$$

The scaling term $2/a^2$ does not affect the optimal forecast, but ensures that this function limits to the MSE loss function as $a \to 0$. When $a > 0$, under-predictions ($y > \hat{y}$, or $e > 0$) carry an approximately exponential penalty, while over-predictions ($y < \hat{y}$, or $e < 0$) carry an approximately linear penalty. When $a < 0$ the penalty for over-predictions is approximately exponential while the penalty for under-predictions is approximately linear. In Figure 1 we present the linex loss function for $a = 3$.

[INSERT FIGURE 1 ABOUT HERE]

Under linex loss, the optimal one-step-ahead forecast and the associated forecast error are (see Varian (1974), Zellner (1986) and Christoffersen and Diebold (1997))

$$\begin{aligned} \hat{Y}_t^* &= E_{t-1}\left[Y_t\right]+\frac{a}{2}V_{t-1}\left[Y_t\right] \\ e_t^* &= -\frac{a}{2}V_{t-1}\left[Y_t\right]+\varepsilon_t \\ &= -\frac{a}{2}h_t+h_t^{1/2}\nu_t \\ \text{so}\quad e_t^*|\mathcal{F}_{t-1} &\sim N\left(-\frac{a}{2}h_t,h_t\right) \end{aligned} \tag{16}$$

and so we see that the process for the conditional mean (an AR(1) process above) does not affect the properties of the optimal forecast error. Notice that the forecast error follows an ARCH-in-mean process of the type analyzed by Engle, Lilien and Robbins (1987).

The generalized forecast error for this example is as follows, and has a log-normal distribution when suitably centered and standardized:

$$\begin{aligned} \psi_t &\equiv \frac{\partial L\left(Y_t,\hat{Y}_t\right)}{\partial\hat{y}}=\frac{2}{a}\left(1-\exp\left\{a\left(Y_t-\hat{Y}_t\right)\right\}\right) \\ \text{so}\quad \left(1-\frac{a}{2}\psi_t\right)|\mathcal{F}_{t-1} &\sim \log N\left(a\left(\mu_t-\hat{Y}_t\right),a^2h_t\right) \\ \text{and}\quad \left(1-\frac{a}{2}\psi_t^*\right)|\mathcal{F}_{t-1} &\sim \log N\left(-\frac{a^2}{2}h_t,a^2h_t\right). \end{aligned} \tag{17}$$

For the numerical example, we chose values of the predicted variance, h_t, to correspond to the mean and the 0.01, 0.25, 0.75, 0.9 and 0.99 percentiles of the unconditional distribution of h_t when

the GARCH parameters are set to $(\omega, \alpha, \beta) = (0.02, 0.05, 0.93)$, which are empirically reasonable. A plot of the objective and the MSE-loss densities is given in Figure 2.

[INSERT FIGURE 2 ABOUT HERE]

In all cases we see that the MSE-loss density is shifted to the right of the objective density, in order to remove the (optimal) negative bias that is present under the objective probability distribution due to the high cost associated with positive forecast errors. The way this probability mass is shifted depends on the level of predicted volatility, and Figure 2 reveals a variety of shapes for the MSE-loss density. When volatility is low, ($h_t = 0.54$ or 0.73) the MSE-loss density remains approximately bell-shaped, and is a simple shift of location (with a minor increase in spread) so that the mean of this density is zero. When volatility is average to moderately-high, ($h_t = 1.00$ or 1.11) the MSE-loss density becomes a more rounded bell shape and remains unimodal. When volatility is high, the MSE-loss density becomes bimodal: it is approximately 'flat-topped' for the $h_t = 1.43$ case (though actually bimodal) and clearly bimodal for the $h_t = 2.45$ case. The bimodality arises from the interaction of the three components that affect the shape of the MSE-loss density: the derivative of the loss function, the shape of the objective density, and the inverse of the forecast error.

We also see that the MSE-loss density is symmetric in this example. This is not a general result: a symmetric objective density (such as in this example) combined with an asymmetric loss function will generally lead to an asymmetric MSE-loss density. It is the particular combination of the normal objective density with the linex loss function that leads to the symmetric MSE-loss function observed here. A symmetric but non-normal conditional density for ν_t, such as a mixture of normals, can be shown to lead to an asymmetric MSE-loss density.

4.1 Application to U.S. Inflation

In this section we apply the methods of this paper to inflation forecasting, which was the application in Rob Engle's original ARCH paper, Engle (1982). We use monthly CPI inflation for the U.S., $\Delta \log(CPI_t)$ over the period January 1982 to December 2006. This happens to be the period starting with the publication of the original ARCH paper, and also coincides with the period after the change in the Federal Reserve's monetary policy during the 'monetarist experiment' from 1979-82. This is widely believed to have led to a break in the inflation dynamics and volatility of

many macroeconomic time-series. We use a simple AR(4) model for the conditional mean, and a GARCH(1,1) model for the conditional variance.[9] Assuming normality for the standardized residuals from this model, we can then obtain both the MSE-optimal forecast (simply the conditional mean) and the Linex-optimal forecast, where we set the linex shape parameter to equal three, as in the previous section.[10] The data and forecasts are presented in Figure 3. In the upper panel we plot both the realized inflation (in percent per month) and the estimated conditional mean, which is labelled in the 'MSE forecast' in the lower panel. The lower panel reveals that the linex forecast is always greater than the MSE forecast, by an amount that grows in periods with high variance (as shown in the middle panel), with the average difference being 0.087%, or 1.04% per year. With average realized inflation at 3.06% per year in this sample period, the linex forecast (optimal) bias is substantial.

[INSERT FIGURE 3 ABOUT HERE]

To emphasize the importance of the loss function in considering forecast optimality, we illustrate two simple tests of optimality for each of the two forecasts.[11] The first looks for bias in the forecast, while the second looks for bias *and* first-order autocorrelation in the forecast errors. The results for the MSE and Linex forecasts are presented below, with Newey-West (1987) t-statistics presented in parentheses below the parameter estimates. The 'p-value' below reports the p-value associated with the test of the null of forecast optimality, either zero bias or zero bias and zero autocorrelation.

$$\begin{aligned}
e_t^{MSE} &= \underset{(-0.123)}{-0.002} + u_t, \quad \text{p-value} = 0.902 \\
e_t^{MSE} &= \underset{(-0.124)}{-0.002} + \underset{(0.050)}{0.003} e_{t-1}^{MSE} + u_t, \quad \text{p-value} = 0.992 \\
e_t^{Linex} &= \underset{(-6.175)}{-0.087} + u_t, \quad \text{p-value} = 0.000 \\
e_t^{Linex} &= \underset{(-6.482)}{-0.085} + \underset{(0.327)}{0.021} e_{t-1}^{Linex} + u_t, \quad \text{p-value} = 0.000
\end{aligned} \tag{18}$$

[9] The Engle (1982) LM test for ARCH in the residuals from the AR(4) model rejected the null of homoskedasticity, at the 0.05 level, for all lags up to 12.

[10] The Jarque-Bera (1987) test for the normality of the standardized residuals actually rejects the assumption of normality here. The estimated skewness of these residuals is near zero, but the kurtosis is 4.38, which is far enough from 3 for this test to reject normality. We nevertheless proceed under the assumption of normality.

[11] Formal testing of forecast optimality would use a pseudo-out-of-sample period for analysis, separate from the period used for estimation.

As expected, the MSE-optimal passes these tests. The Linex-optimal forecast fails both of these tests, primarily due to the positive bias in the linex forecasts. This is, of course, also expected, as the linex forecasts are constructed for a situation where the costs of under-predicting are much greater than those of over-predicting, see Figure 1. Thus the linex forecast is not constructed to be optimal under MSE loss, which is what the above two tests examine.

Next we consider testing for optimality under linex loss, using the generalized forecast error for that loss function and the methods discussed in Section 2. The formula for the generalized forecast for linex loss is given in equation (17) , and from that we construct ψ_t^{MSE} and ψ_t^{Linex} using the MSE forecast and the Linex forecast. We ran the same tests as above, but now using the generalized forecast error rather than the usual forecast error, and obtained the following results:

$$\begin{aligned}
\psi_t^{MSE} &= \underset{(-3.985)}{-0.210} + u_t, \quad \text{p-value} = 0.000 \\
\psi_t^{MSE} &= \underset{(-3.737)}{-0.214} - \underset{(-0.342)}{0.019}\, \psi_{t-1}^{MSE} + u_t, \quad \text{p-value} = 0.000 \\
\psi_t^{Linex} &= \underset{(-0.256)}{-0.010} + u_t, \quad \text{p-value} = 0.798 \\
\psi_t^{Linex} &= \underset{(-0.263)}{-0.010} - \underset{(-0.550)}{0.031}\, \psi_{t-1}^{Linex} + u_t, \quad \text{p-value} = 0.849
\end{aligned} \tag{19}$$

Using the test of optimality based on linex loss (with parameter equal to three), we find that the MSE forecasts are strongly rejected, while the linex forecasts are not. The contrast between this conclusion and the conclusion from the tests based on the usual forecast errors provides a clear illustration of the importance of matching the loss function used in forecast evaluation with that used in forecast construction. Failure to accurately account for the forecaster's objectives through the loss function can clearly lead to false rejections of forecast optimality.

Finally, we present the estimated objective and MSE-loss densities associated with these forecasts. We nonparametrically estimated the objective density of the standardized residuals, $\hat{\nu}_t \equiv (y_t - \hat{\mu}_t)/\sqrt{\hat{h}_t}$, where $\hat{\mu}_t$ is the conditional mean and $\sqrt{\hat{h}_t}$ is the conditional standard deviation, using a Gaussian kernel with bandwidth set to $0.9 \times \sqrt{\hat{V}[\hat{\nu}_t]} \times T^{-1/5}$, where $T = 300$ is the sample size. From this, we can then compute an estimate of the conditional (objective) density of the forecast errors:

$$\hat{f}\left(e|\hat{h}_t\right) = \hat{f}_\nu\left(\frac{e + \hat{h}_t a/2}{\sqrt{\hat{h}_t}}\right)\frac{1}{\sqrt{\hat{h}_t}} \tag{20}$$

The MSE-loss density is estimated as:

$$\tilde{f}\left(e|\hat{h}_t\right) = \frac{\frac{2}{ae}\left(1-\exp\left\{ae\right\}\right)}{\hat{E}\left[\frac{2}{ae_t}\left(1-\exp\left\{ae_t\right\}\right)|h_t\right]}\hat{f}\left(e|\hat{h}_t\right) \tag{21}$$

$$\text{where } \hat{E}\left[\frac{2\left(1-\exp\left\{ae_t\right\}\right)}{ae_t}|h_t\right] \equiv \frac{1}{T}\sum_{i=1}^{T}\frac{2\left(1-\exp\left\{a\left(\sqrt{h_t}\nu_i-\frac{a}{2}h_t\right)\right\}\right)}{a\left(\sqrt{h_t}\nu_i-\frac{a}{2}h_t\right)} \tag{22}$$

and thus uses both the nonparametric estimate of the objective density, and a data-based estimate of the normalization constant.

The estimated objective and MSE-loss densities are presented in Figure 4, using the same method of choosing values for the predicted variance: we use values that correspond to the mean and the 0.01, 0.25, 0.75, 0.9 and 0.99 percentiles of the sample distribution of $\hat{h}_t$ from our model. As in the simulation example in the previous section, we see that the objective density is centered to the left of zero, and that the centering point moves further from zero as the variance increases. A small 'bump' in the right tail of the objective density estimate is amplified in the MSE-loss estimate, particularly as the volatility increases, and the MSE-loss density is approximately centered on zero. The 'bump' in the right tail of both of these densities disappears if we impose that the standardized residuals are truly normally distributed; in that case the objective density is, of course, Gaussian, and the resulting MSE-loss density is unimodal across these values of $\hat{h}_t$.

[INSERT FIGURE 4 ABOUT HERE]

5 Conclusion

This paper derives properties of an optimal forecast that hold for general classes of loss functions in the presence of conditional heteroskedasticity. Studying these properties is important, given the overwhelming evidence for conditional heteroskedasticity that has accumulated since the publication of Engle's seminal (1982) ARCH paper. We show that irrespective of the loss function and data generating process, a generalized orthogonality principle must hold provided information is efficiently embedded in the forecast. We suggest that this orthogonality principle leads to two primary implications: (1) a transformation of the forecast error, the "generalized forecast error", must be uncorrelated with elements of the information set available to the forecaster, and (2) a transformation of the density of the forecast errors, labelled the "MSE-loss" density, must exist

which gives forecasts that are optimal under non-MSE loss the same properties as those that are optimal under MSE loss.

The first approach to testing forecast optimality has its roots in the widely-used Mincer-Zarnowitz (1969) regression, while the second approach is based on a transformation from the usual probability measure to an "MSE-loss probability measure". This transformation has its roots in asset pricing and "risk neutral" probabilities but to our knowledge has not previously been considered in the context of forecasting. Implementing the first approach empirically is relatively straightforward, although it may require estimation of the parameters of the loss function if these are unknown (Elliott et al. (2005)); implementing the second approach will require thinking about forecast (sub-)optimality in a different way, which may yield new insights into forecaster behavior.

Appendix

Proof of Proposition 1. 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 $(Y_t, Y_{t-1}, ...) \in \mathcal{F}_t$ 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 $\{X_t\}$ 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 $\tilde{Z}_t$. 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. ■

Proof of Corollary 1. This proof follows directly from the proof of Proposition 1 above, when one observes the relation between the forecast error and the generalized forecast error, $\psi^*_{t+h,t}$, for the mean squared loss case: $e^*_{t+h,t} = -\frac{1}{2\theta_h}\psi^*_{t+h,t}$, and noting that the MSE loss function satisfies assumptions L1, L3 and L4 which implies a unique interior optimum. ■

To prove Proposition 2 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 $\tilde{L}$, $0 < E_t\left[\frac{\partial L(Y_{t+h},\hat{y})/\partial\hat{y}}{\partial\tilde{L}(Y_{t+h},\hat{y})/\partial\hat{y}}\right] < \infty$ for all $\hat{y} \in \mathcal{Y}$ almost surely.

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, L5 and L6' hold for L and $\tilde{L}$. Then the "$\tilde{L}$-loss probability measure", $\tilde{F}_{e_{t+h,t}}$, defined below is a proper probability distribution function for all $\hat{y} \in \mathcal{Y}$.

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

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. $\tilde{E}\left[\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right]$, the expectation of $\tilde{L}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)$ under $\tilde{F}_e(\cdot;\hat{y})$, is a non-decreasing function of the forecast horizon when evaluated at $\hat{y}=\hat{Y}^*_{t+h,t}$.

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 L5 we have $\Lambda(e,\hat{y}) > 0$ for all e where $\Lambda(e,\hat{y})$ exists. 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,

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

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.

Part 3: Since $\tilde{E}\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\right] = 0$, from part 2, 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 $j \geq h$. Again, by part 2,

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

$\tilde{E}\left[\tilde{\psi}\left(Y_{t+h},\hat{Y}^*_{t+h,t}\right)\cdot\tilde{\psi}\left(Y_{t+h+j},\hat{Y}^*_{t+h+j,t+j}\right)\right]=0$ 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 $\{X_t\}$ 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 $\tilde{Z}_t$. ■

Proof of Proposition 2. 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. ■

References

[1] Andersen, T.G., T. Bollerslev, P.F. Christoffersen, and F.X. Diebold, 2006, Volatility and Correlation Forecasting, in G. Elliott, C.W.J. Granger, and A. Timmermann (eds.), Handbook of Economic Forecasting. Amsterdam: North-Holland.

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

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

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

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

[6] Bollerslev, T., R.F. Engle, and D.B. Nelson, 1994, ARCH Models, in the Handbook of Econometrics, R.F. Engle and D. McFadden eds., North Holland Press, Amsterdam.

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

[8] Chernov, M. and P. Mueller, 2007, The Term Structure of Inflation Forecasts, working paper, London Business School.

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

[10] Christoffersen, P.F and K. Jacobs, 2004, The Importance of the Loss Function in Option Valuation. Journal of Financial Economics, 72, 291-318.

[11] Christoffersen, P.F. and F.X. Diebold, 1996, Further Results on Forecasting and Model Selection Under Asymmetric Loss, Journal of Applied Econometrics, 11, 561-72.

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

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

[14] Elliott, G., I. Komunjer, and A. Timmermann, 2005, Estimation and Testing of Forecast Rationality under Flexible Loss. Review of Economic Studies, 72, 1107-1125.

[15] Elliott, G., I. Komunjer, and A. Timmermann, 2008, Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss? Journal of European Economic Association, 6, 122-157.

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

[17] Engle, R.F, 2004, Risk and Volatility: Econometric Models and Financial Practice, American Economic Review, 94, 405-420.

[18] Engle, R.F., T. Ito and W.L. Lin, 1990, Meteor Showers or Heat Waves? Heteroskedastic Intra-daily Volatility in the Foreign Exchange Market, Econometrica, 58, 525-542.

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

[20] Engle, R.F, and M. Rothschild, 1992, Statistical Models for Financial Volatility, Journal of Econometrics 52, 1-311.

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

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

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

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

[25] Granger, C.W.J., and M.J. Machina, 2006, Forecasting and Decision Theory, in G. Elliott, C.W.J. Granger, and A. Timmermann (eds.), Handbook of Economic Forecasting. Amsterdam: North-Holland.

[26] Hansen, P.R., and A. Lunde, 2006, Consistent Ranking of Volatility Models, Journal of Econometrics, 131, 97-121.

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

[28] Jarque, C.M. and A.K. Bera, 1987, A Test for Normality of Observations and Regression Residuals, International Statistical Review, 55, 163-172.

[29] 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.

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

[31] Patton, A.J., 2006, Volatility Forecast Comparison using Imperfect Volatility Proxies, Quantitative Finance Research Centre, University of Technology Sydney, Research Paper 175.

[32] Patton, A.J., and A. Timmermann, 2007a, Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity, Journal of Econometrics, 140, 884-918.

[33] Patton, A.J. and A. Timmermann, 2007b, Testing Forecast Optimality under Unknown Loss, Journal of American Statistical Association, 102, 1172-1184.

[34] 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.

[35] 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.

[36] West, K.D., 1996, Asymptotic Inference About Predictive Ability, Econometrica, 64, 1067-1084.

[37] West, K.D., 2006, Forecast Evaluation, in G. Elliott, C.W.J. Granger, and A. Timmermann (eds.), Handbook of Economic Forecasting. Amsterdam: North-Holland.

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

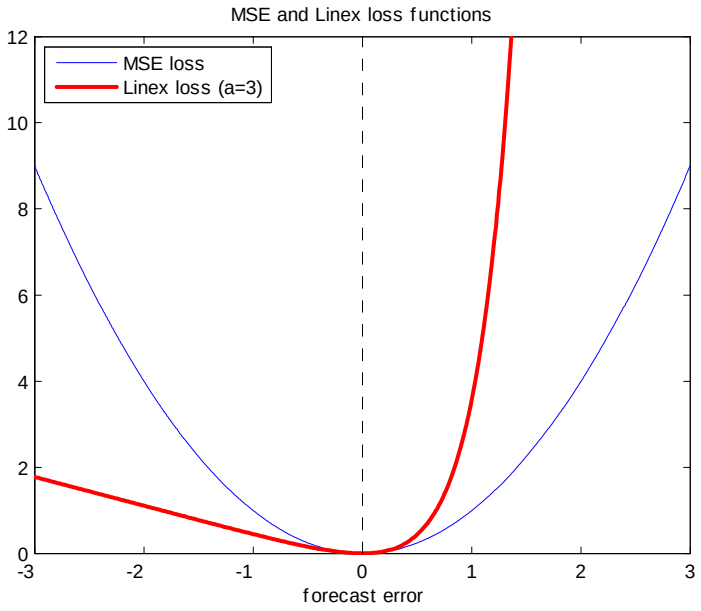

Figure 1: *MSE and Linex loss functions for a range of forecast errors.*

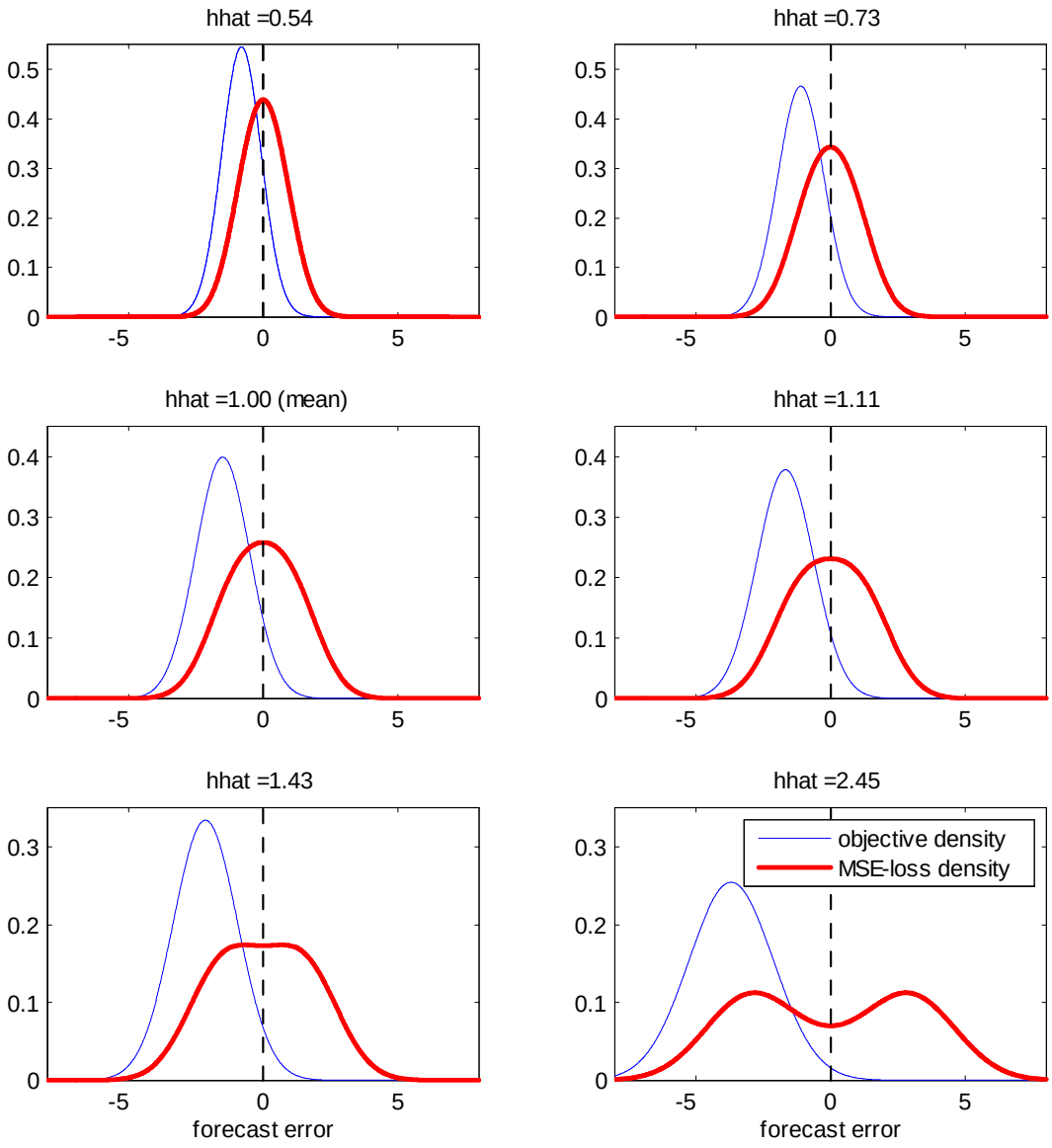

Figure 2: *Objective and "MSE-loss" error densities for a GARCH process under Linex loss, for various values of the predicted conditional variance.*

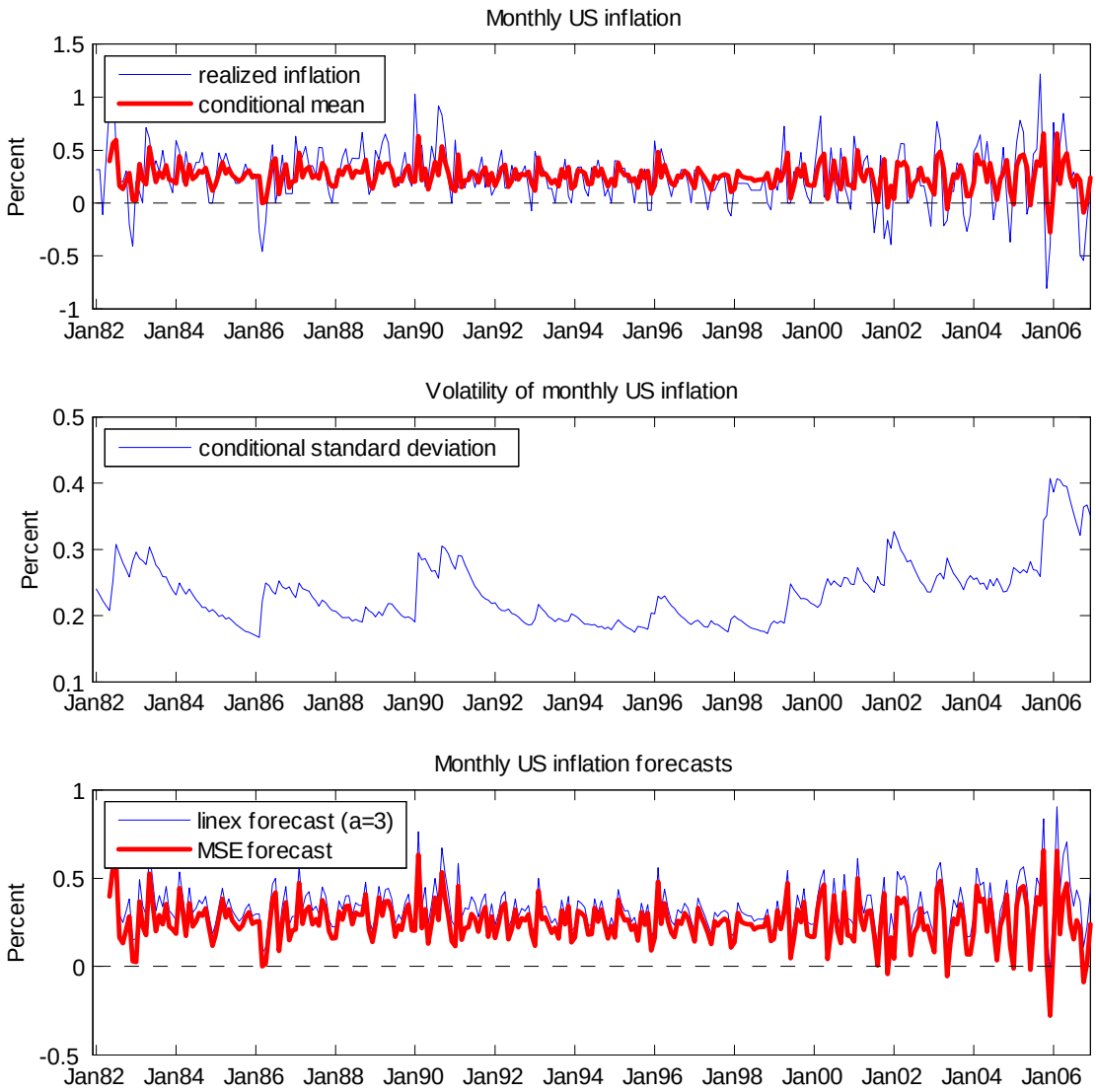

Figure 3: *Monthly CPI inflation in the US over the period January 1982 to December 2006, along with the estimated conditional mean, conditional standard deviation, and the linex-optimal forecast.*

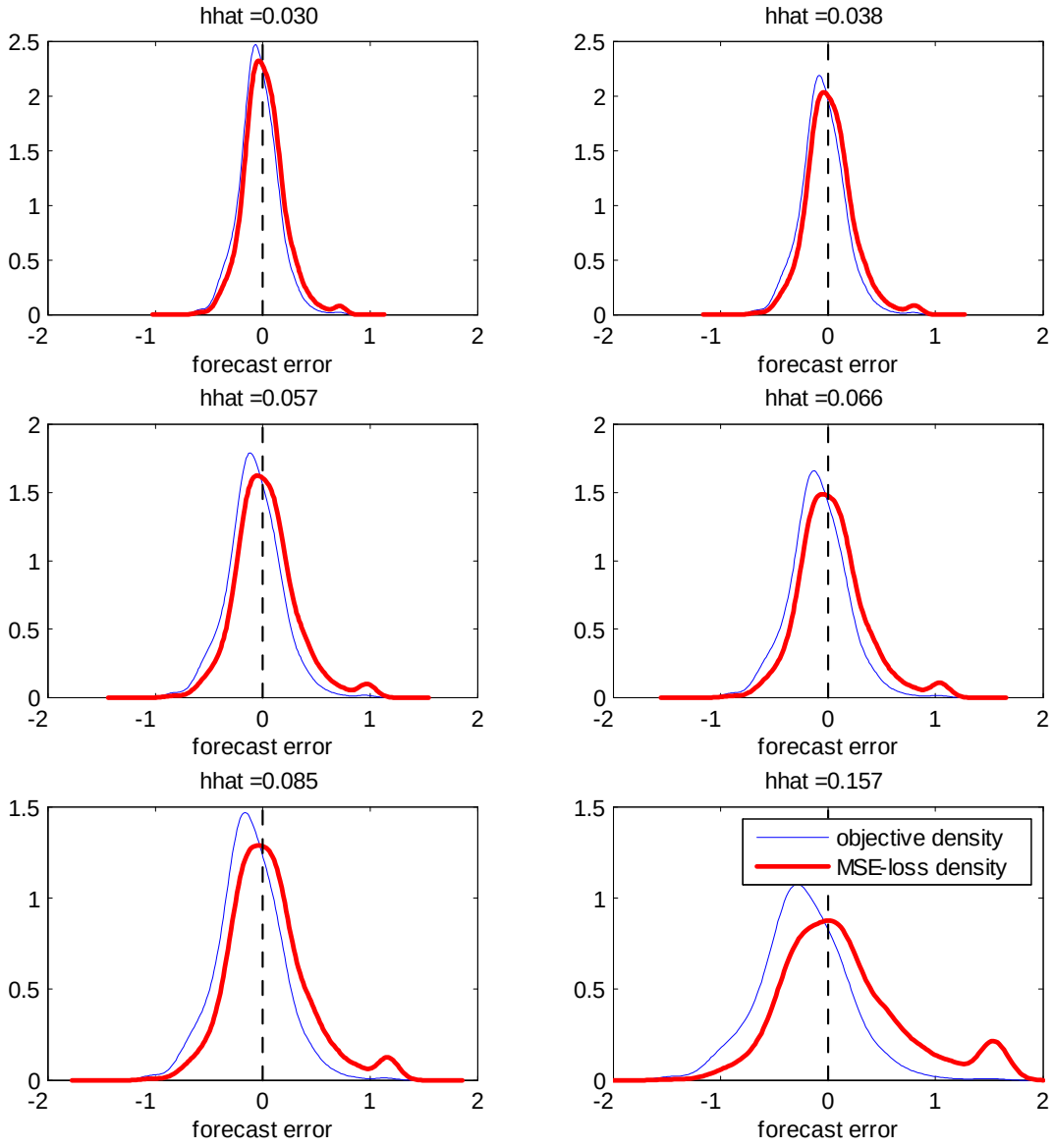

Figure 4: *Estimated objective and "MSE-loss" error densities for US inflation, for various values of the predicted conditional variance.*