Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 122 (2004) 47–79
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Optimal forecast combinations under general loss functions and forecast error distributions

Graham Elliott, Allan Timmermann*

Department of Economics, University of California-San Diego, La Jolla, CA 92093-0508, USA

Accepted 23 October 2003

Abstract

Existing results on the properties and performance of forecast combinations have been derived in the context of mean squared error loss. Under this loss function empirical studies have generally found that equally-weighted combined forecasts lead to better performance than estimates of optimal forecast combination weights which in turn outperform the best individual predictions. We show that this and other results can be overturned when asymmetries are introduced in the loss function and the forecast error distribution is skewed. We characterize the optimal combination weights for the most commonly used alternatives to mean squared error loss and demonstrate how the degree of asymmetry in the loss function and skews in the underlying forecast error distribution can significantly change the optimal combination weights. We also propose estimation methods and investigate their small sample properties in simulations and in an inflation forecasting exercise.
© 2003 Elsevier B.V. All rights reserved.

JEL classification: C130; C530

Keywords: Forecast combination; Loss function; Asymmetric loss

1. Introduction

Economic decision makers are often presented with numerous competing forecasts generated by different economic models or forecasting services with access to different information sets. A decision maker who needs a forecast could choose a single forecast from a method that has historically performed well according to a particular loss function, or some combination of the available forecasts could be considered. If a combination of forecasts proves to be superior to its individual components, a better

* Corresponding author.
E-mail address: atimmerman@ucsd.edu (A. Timmermann).

0304-4076/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2003.10.019

model should be available by pooling the information sets used to form the various forecasts. In practice all the basic information sets are often not available, so this is not a feasible approach. Combining forecasts becomes an attractive strategy in this situation and combinations have indeed proven empirically to lead to substantial improvements over the forecasts produced by the best single model. [1]

The majority of theoretical results on combination weights and empirical findings on forecast combination performance assume a symmetric, quadratic loss function. However, in economics and finance forecasting performance is increasingly evaluated under more general loss functions that account for asymmetries; see, e.g., Christoffersen and Diebold (1997), Diebold (2001), Granger and Newbold (1986), Granger and Pesaran (2000), Varian (1974), West et al. (1993) and Zellner (1986). For example, financial forecasting performance is often measured through the Sharpe ratio which divides mean returns over the standard deviation of returns on a portfolio constructed on the basis of forecasting information. [2] Similarly, it is common to have a value at risk (VaR) objective to account for the disproportionately high costs associated with large losses. This leads to a loss function that puts particular emphasis on draws from the left tail of the forecast error distribution. Central Banks may also have asymmetric preferences which will affect their optimal policies, cf. Peel and Nobay (1998).

This paper examines forecast combinations under general loss functions and forecast error distributions. We establish in general situations the factors that are important in determining the optimal weights in population; numerical examples are then used to demonstrate the extent to which the degree of asymmetry in the loss function interacts with the skew and kurtosis of the underlying forecast errors from the individual models in determining optimal combination weights. It is shown that it is the combination of asymmetry in both loss and data that is required for weights to differ from the mean square error (MSE) optimal weights. Under elliptical symmetry, our paper shows that the linear forecast combination weights (other than the constant term) are identical for (almost) all loss functions. This invariance result follows since for any set of forecast combination weights, the constant can be adjusted to choose the bias that optimally trades off the bias and variance of the forecast error. The weights on the forecasts are left free to minimize the variance and can thus be found as the solution to a standard quadratic optimization problem.

This result has useful practical implications. Since the forecast combination weights are the same as those under MSE loss, the estimation problem reduces to a simple two-stage procedure: least squares estimation of the optimal combination weights followed by estimation of the constant term which controls the optimal bias. This simplifies the estimation problem greatly since typical estimation procedures require

[1] See, e.g., the extensive survey by Clemen (1989). More recently Chan et al. (1999), Marcellino (2002), Newbold and Harvey (2001) and Stock and Watson (1999) have found evidence that combined forecasts often outperform forecasts from the best individual model. Reasons offered for the empirical success of forecast combinations include model misspecification, changes in the underlying parameters, different models' use of heterogenous information and shrinkage estimation effects. See Hendry and Clements (2002) for a discussion of these points.

[2] Hence even if the mapping from forecasts to portfolio weights is linear, there will be a highly nonlinear mapping from the forecasts to the performance metric.

search methods which can be difficult if the dimensions of the parameter vector is large.

An often quoted 'folk theorem' in the forecast combination literature is that using average as opposed to optimal weights often works better in practice.[3] The reason is that it can be difficult to precisely estimate the optimal forecast combination weights. Equal-weighted forecast combinations may be biased but they also reduce the forecast error variance by not relying on estimated combination weights that depend on second moments of forecast errors. This finding depends, however, on the trade-off between forecast error bias and variance (as well as higher order moments in the case of asymmetric forecast errors) which in turn depends on the shape of the loss function. Consequently this 'folk theorem' can be overturned under asymmetric loss functions and asymmetrically distributed forecast errors.

We finally address estimation of optimal combination weights in a much more general context than the existing literature. For general differentiable loss functions such as linex we propose to estimate the forecast combination weights using M estimation. For the lin–lin loss function estimation can be cast in the context of a quantile regression problem. For asymmetric quadratic loss, we propose an iterated weighted least squares estimation approach. The performance of these and other estimators is examined in Monte Carlo simulations.

The plan of the paper is as follows. Section 2 reviews results for the standard case with mean squared error loss. Section 3 provides theoretical results for the general case with arbitrary loss function and forecast error distribution. Section 4 considers three commonly used asymmetric loss functions and Section 5 investigates forecast combinations in the context of Gaussian mixture distributions. Section 6 discusses estimation of the optimal combination weights, while Section 7 presents Monte Carlo simulation results. Section 8 provides an empirical application to inflation forecasting and Section 9 concludes.

2. Forecast combination under mean squared error loss

Suppose a decision maker is interested in forecasting some variable, y_t, on the basis of an m-vector of forecasts of this variable $\hat{\mathbf{y}}_t$. Each element of $\hat{\mathbf{y}}_t$ is determined ex ante and is adapted to an expanding sequence of information sets, Ω_t, which constitutes a standard filtration. Hence $\hat{\mathbf{y}}_t$ is adapted to Ω_{t-1}, whereas y_t is not.[4] Ω_{t-1} comprises y_{t-1} in addition to other variables used to predict y_t. We will assume that y_t and $\hat{\mathbf{y}}_t$ have joint distribution $\mathscr{P}((y_t \ \hat{\mathbf{y}}_t')')$ with finite first and second moments

$$\mathrm{E}\begin{pmatrix} y_t \\ \hat{\mathbf{y}}_t \end{pmatrix} = \begin{pmatrix} \mu_y \\ \boldsymbol{\mu} \end{pmatrix} \tag{1}$$

[3] See, e.g., Granger (1989), Stock and Watson (1999) and Fildes and Ord (2001).

[4] Thus we are only ruling out the uninteresting case where y_t is perfectly predictable.

and

$$\mathrm{Var}\begin{pmatrix} y_t \\ \hat{\mathbf{y}}_t \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{21}' \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \tag{2}$$

To keep the notation simple, we suppress the dependence of the joint distribution and moments on Ω_{t-1}, simply using $\mathscr{P}(.)$, E[.] and Var (.) in place of $\mathscr{P}(.|\Omega_{t-1}), \mathrm{E}[.|\Omega_{t-1}]$ and $\mathrm{Var}(.|\Omega_{t-1})$.

Motivated by the seminal paper of Bates and Granger (1969), the forecast combination literature has studied the class of linear forecast combinations, $\omega^c + \boldsymbol{\omega}'\hat{\mathbf{y}}_t$, where $\boldsymbol{\omega}$ is an m-vector of combination weights and ω^c is a scalar constant. This gives rise to a forecast error, e_t, from the combination,

$$e_t = y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t. \tag{3}$$

Under the assumed moment structure e_t has first and second moments

$$\begin{aligned} \mu_e &= \mu_y - \omega^c - \boldsymbol{\omega}'\mu, \\ \sigma_e^2 &= \sigma_y^2 + \boldsymbol{\omega}'\boldsymbol{\Sigma}_{22}\boldsymbol{\omega} - 2\boldsymbol{\omega}'\boldsymbol{\sigma}_{21}. \end{aligned} \tag{4}$$

Assuming the loss is quadratic and symmetric in the forecast error, $L(e_t) = e_t^2$, the objective is to minimize the following expression:

$$\mathrm{E}[(y_t - \boldsymbol{\omega}^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t)^2] = (\mu_y - \omega^c - \boldsymbol{\omega}'\mu)^2 + \sigma_y^2 + \boldsymbol{\omega}'\boldsymbol{\Sigma}_{22}\boldsymbol{\omega} - 2\omega'\boldsymbol{\sigma}_{21}. \tag{5}$$

The optimal population values of the constant and the vector of combination weights, ω_0^c and $\boldsymbol{\omega}_0$, are thus [5]

$$\begin{aligned} \omega_0^c &= \mu_y - \boldsymbol{\omega}_0'\boldsymbol{\mu}, \\ \boldsymbol{\omega}_0 &= \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}. \end{aligned} \tag{6}$$

The sample analog of these weights is of course the usual least squares estimator for the outcomes regressed on a constant and the vector of forecasts (Granger and Ramanathan, 1984). This solution takes an interesting form: the vector of combination weights depends only on the variance–covariance matrix of the outcome and the predictions. Given the optimal weights, $\boldsymbol{\omega}_0$, the optimal constant depends on the biases in the forecasts but not directly on the variance–covariance matrix, although it will of course depend on this indirectly through $\boldsymbol{\omega}_0$. The inclusion of a constant ensures that the combined forecast is unbiased since ω_0^c can be chosen such that $\mu_e = 0$, which is optimal under MSE loss. For any variance–covariance matrix, and hence $\boldsymbol{\omega}$, ω_0^c adjusts to ensure that the combined forecast has the optimal bias.

Bayesian solutions to the forecast combination problem under quadratic loss have also been proposed. Palm and Zellner (1992) develop a general Bayesian framework for combinations where individual forecasts can be biased and the covariance matrix of the forecast errors may be unknown.

[5] We shall generally be assuming that the covariance matrix $\boldsymbol{\Sigma}_{22}$ is invertible, but see further below for a discussion of this point.

3. Forecast combination under general loss functions and forecast error distributions

While the MSE loss function may be a useful approximation in some cases, recent interest has focussed on extending existing results to allow for more general asymmetric loss functions. Tay and Wallis (2000) provide several references to this literature. Christoffersen and Diebold (1997) develop the theory for optimal bias adjustment in the case of commonly used asymmetric loss functions when the moments of the underlying error distribution vary over time, such as under volatility clustering, but maintain the assumption that (conditionally) the forecast errors follow a Gaussian distribution.

None of these results have been extended to the forecast combination literature. To fill out this gap, this section develops results for optimal forecast combinations under general loss functions and forecast error distributions. We then explore the significance of the factors determining the optimal combination weights for the most commonly used asymmetric loss functions such as lin–lin, linex and asymmetric quadratic loss.

3.1. The loss function

We will be concerned with loss functions that only depend on the forecast error and thus take the form $L(e_t)$. We assume $L(.)$ has the following properties (Granger, 1999):

1. $L(0) = 0$;
2. $\text{Min}_e\ L(e) = 0$, so $L(e) \geqslant 0$;
3. $L(e)$ is monotonic nondecreasing as e moves away from zero:

$$L(e_1) \geqslant L(e_2) \quad \text{if } e_1 > e_2 \geqslant 0 \text{ or if } e_1 < e_2 \leqslant 0.$$

In addition to these properties, the loss function may also be symmetric ($L(-e) = L(e)$), homogenous ($L(ae)=h(a)L(e)$ for some positive function $h(a)$) and differentiable up to some order.

Under quadratic loss forecasting and decision problems can be separated and only the conditional mean of the predicted random variables matter due to certainty equivalence. This does not hold for more general loss functions. 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, cf. 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).

The decision maker's problem is to find the optimal forecast combination weights and a constant that minimize expected loss:

$$\min_{\omega^c,\boldsymbol{\omega}} \int L(e_t)\, \mathrm{d}F(e_t), \tag{7}$$

where $F(e_t)$ is the cumulated density of e_t.

Optimal forecast combination weights generally depend on the shapes of both the loss function and the forecast error distribution. This can be seen from the following result:

Proposition 1. *Suppose that* (1) *the expected loss is finite*, (2) *the loss function is analytic except for possibly at a finite number of points* (*occurring with probability zero*) *at which it is continuous*; *and* (3) *all moments of the forecast error distribution up to the highest non-zero derivative of the loss function with respect to the forecast combination weights exist.*

Then the optimal combination weights solve the expression

$$\min_{\omega^c,\boldsymbol{\omega}} \left\{ L(\mu_e) + \frac{1}{2} L''_{\mu_e} \mathrm{E}[(e_t - \mu_e)^2] + \sum_{m=3}^{\infty} L^m_{\mu_e} \sum_{i=0}^{m} \frac{1}{i!(m-i)!} \mathrm{E}[e_t^{m-i} \mu_e^i] \right\}.$$

where $L^k_{\mu_e} = \partial^k L(e_t)/\partial^k \boldsymbol{\omega}|_{e_t=\mu_e}$.

The result follows from noting that, at the points where $L(.)$ is analytic and thus can be represented by a power series, cf. Rudin (1964, p. 158), it lends itself to a Taylor-series expansion around μ_e:

$$L(e_t) = L(\mu_e) + L'_{\mu_e}(e_t - \mu_e) + \frac{1}{2} L''_{\mu_e}(e_t - \mu_e)^2 + \sum_{k=3}^{\infty} \left(\frac{1}{k!} \right) L^k_{\mu_e}(e_t - \mu_e)^k.$$

Taking expectations, the finite number of points where $L(.)$ is not analytic can be ignored since they are assumed to occur with probability zero. Thus:

$$\begin{aligned}
\mathrm{E}[L(e_t)] &= L(\mu_e) + \frac{1}{2} L''_{\mu_e} \mathrm{E}[(e_t - \mu_e)^2] + \sum_{k=3}^{\infty} \left(\frac{1}{k!} \right) L^k_{\mu_e} \mathrm{E}[(e_t - \mu_e)^k] \\
&= L(\mu_e) + \frac{1}{2} L''_{\mu_e} \mathrm{E}[(e_t - \mu_e)^2] + \sum_{k=3}^{\infty} \left(\frac{1}{k!} \right) L^k_{\mu_e} \sum_{i=0}^{k} \binom{k}{i} \mathrm{E}[e_t^{k-i} \mu_e^i] \\
&= L(\mu_e) + \frac{1}{2} L''_{\mu_e} \mathrm{E}[(e_t - \mu_e)^2] + \sum_{k=3}^{\infty} L^k_{\mu_e} \sum_{i=0}^{k} \frac{1}{i!(k-i)!} \mathrm{E}[e_t^{k-i} \mu_e^i].
\end{aligned}$$

The second assumption does not rule out loss functions such as lin–lin which are non-differentiable at a single point. The third assumption requires that all moments exist for which the corresponding derivative of the loss function with respect to the forecast error is non-zero. This is a strong requirement and rules out many potential combinations of loss functions and forecast error distributions. [6]

[6] For example, combining a t-distribution with three degrees of freedom with a loss function whose fourth derivative is non-zero will result in a non-existing expected loss. Another example comes from the linex loss function, whose higher order derivatives are all non-zero. All moments of the forecast error distribution must therefore exist to ensure that the expected loss is well defined.

When the loss function is not linear or quadratic, higher order moments of the forecast error distribution such as the skew generally matter. It is not surprising that asymmetry in the loss function interacts with asymmetry in the underlying distributions of the outcomes and forecasts that are being combined. Suppose that the decision maker dislikes positive forecast errors more than negative ones. Now consider moving mass of the distribution to the right, so the underlying distribution is positively skewed. The loss function prefers lower probability of positive forecast errors when choosing the combination weights. The extent of the asymmetry in the loss function drives the size of this effect. Hence the combination weights are clearly a function of the parameters of the loss function. There is also a large effect on the optimal constant. When positive forecast errors are to be avoided, the constant is chosen to ensure that the forecast error distribution is centered to the left of zero.

3.2. Elliptically symmetric data

In special cases only the constant (ω^c) depends on the shape of the loss function, while the forecast combination weights are unaltered from the case with MSE loss:

Proposition 2. *Suppose that the expected loss can be written as* $\mathrm{E}[L(e_t)] = g(\mu_e, \sigma_e^2)$ *and let* μ_e^* *be the optimal value for* μ_e. *Then if the partial derivative* $\partial g(\mu_e^*, \sigma_e^2)/\partial \sigma_e^2$ *is nonzero*,

(i) $\boldsymbol{\omega}_0 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$;
(ii) ω_0^c *is the solution to* $\partial g(\mu_e^*, \sigma_e^2)/\partial \mu_e = 0$.

The result follows since ω^c only appears in μ_e, so the optimal value for this parameter solves

$$\frac{\partial g(\mu_e, \sigma_e^2)}{\partial \omega^c} = \frac{\partial g(\mu_e, \sigma_e^2)}{\partial \mu_e}\frac{\partial \mu_e}{\partial \omega^c} = -\frac{\partial g(\mu_e, \sigma_e^2)}{\partial \mu_e} = 0.$$

This is the result in part (ii) of the proposition. This choice for ω^c sets μ_e at its optimal point (μ_e^*). The optimal value for ω therefore solves

$$\frac{\partial g(\mu_e^*, \sigma_e^2)}{\partial \boldsymbol{\omega}} = \frac{\partial g(\mu_e^*, \sigma_e^2)}{\partial \sigma_e^2}\frac{\partial \sigma_e^2}{\partial \boldsymbol{\omega}} = \mathbf{0}.$$

Furthermore, $\partial \sigma_e^2/\partial \boldsymbol{\omega} = 2(\boldsymbol{\Sigma}_{22}\boldsymbol{\omega} - \boldsymbol{\sigma}_{21})$ so the solution is as stated so long as $\partial g(\mu_e^*, \sigma_e^2)/\partial \sigma_e^2 \neq 0$. The second order condition is satisfied as long as the loss function satisfies conditions (1)–(3) above.

Proposition 2 shows that the optimal forecast combination weights are identical to the MSE weights for a wide set of loss functions provided that the marginal distribution of the forecast errors depends only on the first two moments of the forecast

errors.[7] The general class of distributions for which this is true is for $\mathcal{P}((y_t \hat{\mathbf{y}}_t')')$ elliptically symmetric.[8] The multivariate normal is a special case of this family, as is the multivariate t-distribution. We shall for simplicity refer to distributions satisfying the assumptions of Proposition 2 as elliptically symmetric.

Proposition 2 has two important implications. First, estimation of forecast combination weights under asymmetric loss simplifies to a two-stage procedure (see Section 6.2 below for details). The second implication is that departures from elliptical symmetry are needed to drive the forecast combination weights away from their solution under MSE loss. If, under elliptical symmetry, a forecast adds information (in the sense that it has a non-zero combination weight) under MSE loss, it will also add information under arbitrarily asymmetric loss functions. The converse also holds—if the forecast is not useful under MSE loss then it will not be useful for any other loss function:

Corollary 1. *Suppose that the conditions of Proposition* 2 *hold. Then a forecast gets assigned a non-zero weight in the combined forecast under a general loss function if and only if its weight under MSE loss is non-zero.*

This result stands in stark contrast to the results we obtain under general forecast error distributions. For sufficient deviation from elliptical symmetry (particularly when there is skew in the forecast errors) the weights on the forecasts will depend on the loss function. However, under elliptically symmetric forecast errors, decision makers agree on the value of prediction signals irrespective of their loss function.[9] They could well disagree about the need to include an intercept as this depends on the shape of the loss function.

When elliptical symmetry is abandoned, a particular forecast or subset of forecasts may contain information under MSE loss but not under asymmetric loss. The converse may also hold: a forecast could be assigned a non-zero weight under asymmetric loss but not under MSE loss. We state this result in Corollary 2 which is the natural converse of Corollary 1 which was established under elliptical symmetry:

Corollary 2. *Suppose that the conditions of Proposition* 2 *do not hold. Then*

(i) *a forecast may contain information under MSE loss but not under alternative* (*asymmetric*) *loss functions*;

(ii) *a forecast may contain information under asymmetric loss but not under MSE loss.*

[7] This implication is closely related to the result in finance that the distribution of a portfolio is determined by the mean and variance in these situations, see Chamberlain (1993). Under symmetry restrictions on both the loss function and the forecast error distribution Granger (1969) obtains a related result.

[8] $\mathbf{X}$ is elliptically symmetric, $ES(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density takes the form $k|\boldsymbol{\Sigma}|^{-1/2} g((\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu}))$, where k is an arbitrary constant and $g(.)$ is any function that represents a density.

[9] Notice also the relationship between Corollary 1 and the result in Diebold et al. (1998) that decision makers with different loss functions cannot generally agree in their ranking of misspecified predictive density models. Provided that the forecast errors are elliptically symmetric, our result says that decision makers can agree about what is a useful forecast even if the forecasts are individually misspecified.

The above results allow for arbitrary correlations between the forecast errors, as long as the covariance matrix Σ_{22} is invertible. This condition is unimportant since linearly dependent forecasts would always be dropped from the combination. Serial correlation, on the other hand, is more difficult to handle in our setup. One possibility is to include extra regressors in the forecast combination and use these to absorb any serial correlation that is still present in the combined forecast. An iterative procedure would (a) estimate the combined forecast under the assumed loss function; (b) test for serial correlation in the errors from the combined forecast; (c) add the lagged combined or individual forecasts as extra 'forecasts' if there is evidence of serial correlation in the (combined) forecast error; repeat step (a) and continue the iteration until there is no serial correlation. An alternative strategy, proposed by Diebold (1988), is to use an estimator that allows the forecast errors from the combination model to follow a finite-order ARMA process.

3.3. Forecast encompassing tests

An alternative to the strategy of forecast combination is to simply use a single forecast. This situation arises when the forecast produced by one model, B, does not add anything to the forecast from another model, A. In this situation, model A is said to forecast encompass model B. Under MSE loss usually this is tested by least squares estimation of the equation

$$y_{t+1} = \alpha + \beta_A \hat{y}^A_{t+1} + \beta_B \hat{y}^B_{t+1} + \varepsilon_{t+1}. \tag{8}$$

If the joint hypothesis that $\beta_A = 1$, $\beta_B = 0$ cannot be rejected, model A forecast encompasses model B.

An implication of the previous two propositions and corollaries is the following

Proposition 3. *Suppose that the assumptions of Proposition* 2 *hold. Then if model A forecast encompasses model B under MSE loss, model A also encompasses model B for any other loss function.*

Conversely, if these assumptions do not hold, it is possible that model A forecast encompasses model B under MSE loss but not under an alternative loss function or that model A does not forecast encompass model B under MSE loss, but does so under the alternative loss function.

When forecast errors are not generated by an elliptically symmetric distribution, it follows that it is not possible to set up a universal encompassing test since the outcome will depend on the specific loss function. Encompassing tests will be specific to the loss function.

Min and Zellner (1993) determine the factors that may lead a single forecast to dominate the strategy of combining multiple forecasts. When the individual forecasts are either unbiased or have a bias of the same magnitude, some degree of forecast combination is often optimal. In contrast, when one forecast is more heavily biased than the others and a constant is not included in the forecast combination, it becomes more likely that individual forecasts will dominate forecast combinations.

4. Results for specific loss functions

To demonstrate the general results from the previous section, this section considers three commonly entertained asymmetric loss functions, namely linex, lin–lin and asymmetric quadratic loss.

4.1. Linex loss

The linex loss function was introduced by Varian (1974) in the context of losses associated with real estate assessments. Zellner (1986) gave a comprehensive treatment of Bayesian predictions for this loss function. Linex loss takes the form

$$L(e_t) = \exp(ae_t) - ae_t - 1. \tag{9}$$

The parameter a controls the extent of asymmetry. If $a > 0$, there are large losses from positive forecast errors and the losses are higher the larger the value of a. If $a < 0$, large losses result from negative forecast errors and the losses are larger, the smaller is a. The expected loss is

$$\mathrm{E}[L(e_t)] = M_e(a) - a\mu_e - 1, \tag{10}$$

where $M_e(a)$ is the moment generating function of the forecast errors and thus depends on their marginal distribution. Eq. (10) shows that the expected linex loss only exists when the forecast error distribution has an infinite number of moments.[10] Differentiating the expected loss with respect to ω^c and $\boldsymbol{\omega}$ yields the first order conditions

$$\begin{aligned} \frac{\partial \mathrm{E}[L(e_t)]}{\partial \omega^c} &= \frac{\partial M_e(a)}{\partial \omega^c} + a = 0, \\ \frac{\partial \mathrm{E}[L(e_t)]}{\partial \boldsymbol{\omega}} &= \frac{\partial M_e(a)}{\partial \boldsymbol{\omega}} + a\boldsymbol{\mu} = \mathbf{0}. \end{aligned} \tag{11}$$

These equations fully characterize the solution to the optimal weights as a function of the parameters of the underlying forecast error distribution and of the loss function (a).

In common with most other asymmetric loss functions, (11) is generally a highly nonlinear function in the forecast combination weights and simple closed form solutions are typically not available. An exception arises when the forecast errors are normally distributed. In this case $M_e(a) = \exp\{a\mu_e + (a^2/2)\sigma_e^2\}$, and the optimal constant is given by

$$\omega_0^c = \mu_y - \boldsymbol{\omega}_0'\boldsymbol{\mu} + \frac{a}{2}\sigma_e^2.$$

[10] This follows from Proposition 1 since all derivatives of the linex loss function are nonzero. The primitive condition in Proposition 2 requires that

$$\frac{\partial M_e(a)}{\partial \sigma_e} = a\sigma_e \exp\left\{a\mu_e + \frac{a^2}{2}\sigma_e^2\right\} \neq 0.$$

This is always satisfied, except in the trivial case with perfect forecasts ($\sigma_e = 0$).

Under normality of the forecast errors, the optimal population bias from the combined forecasts is $\mu_e = -(a/2)\sigma_e^2$. This is the optimal bias derived in Christoffersen and Diebold (1997) for the univariate linex forecasting problem. The expression has an intuitive interpretation: when $a > 0$, large losses follow from positive forecast errors and it is optimal to choose a positive constant so that the forecast errors have a negative mean. The forecast error distribution is hence biased towards negative losses to avoid the high penalties associated with positive losses. The reverse is true when the decision maker tries to avoid negative losses (i.e. $a < 0$). In both cases, the larger is a the larger the penalty so the larger the bias. The larger the variance of the forecast errors the higher the chance of large forecast errors and hence the more the combination gets biased to reduce the chance of large losses.

4.2. Lin–lin loss

A popular way of capturing asymmetries in the loss function is to let losses take one form for forecast errors over and above a certain threshold, κ, and another form for forecast errors below this value:

$$L(e_t) = \begin{cases} L_1(e_t) & \text{if } e_t > \kappa, \\ L_2(e_t) & \text{if } e_t \leqslant \kappa. \end{cases}$$

$L_1(\kappa) = L_2(\kappa)$ ensures continuity. Setting $\kappa = 0$ and assuming that $L(.)$ is piecewise linear, we get the lin–lin loss function:

$$L(e_t) = \begin{cases} (1-\theta)|e_t| & \text{if } e_t > 0, \\ \theta|e_t| & \text{if } e_t \leqslant 0, \end{cases} \tag{12}$$

where $0 < \theta < 1$. Using an indicator function, the lin–lin loss function can be rewritten as

$$L(e_t) = \{-\theta + 1_{e_t > 0}\}e_t. \tag{13}$$

Minimizing this expression over the constant and combination weights is equivalent to solving the problem

$$\min_{\omega^c, \boldsymbol{\omega}} \{-\theta + 1_{e_t > 0}\}e_t.$$

The expected loss to be minimized is thus

$$\mathrm{E}[L(e_t)] = \int_0^\infty e_t \, \mathrm{d}F(e_t) - \theta\mu_e,$$

where $F(e_t)$ is the cumulative density function of the forecast error. Expressing e_t as $e_t = \mu_e + \sigma_e z_t$, so z_t is the centered and standardized forecast error with density $f_z(.)$ and cumulative density $F_z(z)$, the expected loss is

$$\mathrm{E}[L(e_t)] = \mu_e \left(1 - \theta - F_z\left(-\frac{\mu_e}{\sigma_e}\right)\right) + \sigma_e \int_{-\mu_e/\sigma_e}^\infty z_t \, \mathrm{d}F_z(z_t). \tag{14}$$

Derivatives of $L(.)$ with respect to ω^c and $\boldsymbol{\omega}$ are difficult to interpret as in general the marginal distribution $F_z(.)$ will itself depend on $(\omega^c, \boldsymbol{\omega})$. An exception to this occurs

when e_t is Gaussian and, using Leibnitz' rule of integration, the first order condition for the constant simplifies to

$$1-\theta-\Phi\left(-\frac{\mu_e}{\sigma_e}\right)=0,$$

where $\Phi(.)$ is the cumulative density of a standard normal variate.

Recalling that $\mu_e=\mu_y-\omega^c-\boldsymbol{\omega}_0'\boldsymbol{\mu}$ and using the symmetry of $\Phi(.)$ we obtain a closed-form solution for the scalar intercept term: [11]

$$\omega_0^c=\mu_y-\boldsymbol{\omega}_0'\boldsymbol{\mu}-\sigma_e\Phi^{-1}(\theta). \tag{15}$$

The direction of the bias can be evaluated using $\mu_e=\sigma_e\Phi^{-1}(\theta)$. For $\theta=0.5$, the loss function is symmetric and the forecast error is unbiased. For $\theta<1/2$, higher weight is placed on positive errors. To lower the probability of large losses associated with positive forecast errors the optimal combination biases the mean forecast error to be negative ($\Phi^{-1}(\theta)<0$) irrespective of the covariance matrix of the underlying forecasts. The reverse is true when $\theta>1/2$. As with linex loss, the size of the bias increases as the forecast error variance rises, to better avoid high loss outcomes.

4.3. Asymmetric squared loss

The asymmetric squared loss function takes the form

$$L(e_t)=\begin{cases}(1-\theta)e_t^2 & \text{if } e_t>0,\\ \theta e_t^2 & \text{if } e_t\leqslant 0,\end{cases} \tag{16}$$

where $0<\theta<1$. Minimizing this expression over $\omega^c,\boldsymbol{\omega}$ is equivalent to solving

$$\min_{\omega^c,\boldsymbol{\omega}}\{\theta-(2\theta-1)1_{e_t>0}\}e_t^2. \tag{17}$$

Again we can write $e_t=\mu_e+\sigma_e z_t$ and change variables in the expression for the expected loss:

$$\mathrm{E}[L(e_t)]=\theta(\sigma_e^2+\mu_e^2)-(2\theta-1)\left\{\mu_e^2(1-F_z(-\mu_e/\sigma_e))+\sigma_e^2\int_{-\mu_e/\sigma_e}^{\infty}z_t^2\,\mathrm{d}F_z(z_t)+2\mu_e\sigma_e\int_{-\mu_e/\sigma_e}^{\infty}z_t\,\mathrm{d}F_z(z_t)\right\}. \tag{18}$$

[11] The assumption of Proposition 2 can be verified under symmetry of the forecast errors by differentiating the expected loss with respect to σ_e, evaluated at μ_e^*:

$$\begin{aligned}\frac{\partial}{\partial\sigma_e}\mathrm{E}[L(e_t)]&=\int_{-\mu_e^*/\sigma_e}^{\infty}z_t\,\mathrm{d}F_z(z_t)\\&=\int_{-\mu_e^*/\sigma_e}^{\mu_e^*/\sigma_e}z_t\,\mathrm{d}F_z(z_t)+\int_{\mu_e^*/\sigma_e}^{\infty}z_t\,\mathrm{d}F_z(z_t)\\&=\int_{\mu_e^*/\sigma_e}^{\infty}z_t\,\mathrm{d}F_z(z_t)>0,\end{aligned}$$

by symmetry of $\mathrm{d}F_z(.)$ around zero.

Taking derivatives with respect to $\boldsymbol{\omega}$ yields the first order condition

$$\begin{aligned}\mathbf{0} = \theta(&\boldsymbol{\Sigma}_{22}\boldsymbol{\omega} - \boldsymbol{\sigma}_{21} - \mu_e\boldsymbol{\mu}) \\ &- (2\theta - 1)\left\{-\boldsymbol{\mu}\int_{-\mu_e/\sigma_e}^{\infty}(\mu_e + \sigma_e z_t)\,\mathrm{d}F_z(z_t)\right. \\ &\left. + \frac{1}{\sigma_e}(\boldsymbol{\Sigma}_{22}\boldsymbol{\omega} - \boldsymbol{\sigma}_{21})\int_{-\mu_e/\sigma_e}^{\infty}(\mu_e + \sigma_e z_t)z_t\,\mathrm{d}F_z(z_t)\right\}.\end{aligned}$$

The first order condition for the constant term takes the complicated form

$$\omega_0^c = \mu_y - \boldsymbol{\omega}_0'\boldsymbol{\mu} - \frac{(2\theta - 1)\sigma_e\int_{-\mu_e/\sigma_e}^{\infty} z_t\,\mathrm{d}F_z(z_t)}{\theta - (2\theta - 1)(1 - F_z(-\mu_e/\sigma_e))}. \tag{19}$$

This term does not simplify nicely even in the Gaussian case. Only when $\theta = 0.5$, do we get the closed-form solution $\boldsymbol{\omega}_0 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$ and $\omega_0^c = \mu_y - \boldsymbol{\omega}_0'\boldsymbol{\mu}$, which is identical to the MSE result from Section 2. For most other cases, the above formulas do not provide a closed-form solution to ω_0^c since μ_e is itself a function of this constant and one must again resort to solution methods. [12]

5. Mixture distributions

Section 3 showed that departures from elliptical symmetry are needed to get combination weights that differ from the standard MSE weights. To examine the effect of such departures, this section considers models where the joint distribution is a mixture of k normals. Such mixture distributions are capable of closely approximating a very wide family of distributions [13] and, when applied to our setup, take the form

$$\begin{aligned}&f(e_t|\omega^c, \boldsymbol{\omega}) = \sum_{i=1}^{k} p_i\phi_{\sigma_i}(e_t - \mu_i), \\ &\sum_{i=1}^{k} p_i = 1, \quad 0 < p_i \leqslant 1.\end{aligned} \tag{20}$$

where $\mu_i = \mu_{y_i} - \omega^c - \boldsymbol{\omega}'\boldsymbol{\mu}_i$, $\sigma_i^2 = \sigma_{yi}^2 + \boldsymbol{\omega}'\boldsymbol{\Sigma}_{22i}\boldsymbol{\omega} - 2\boldsymbol{\omega}'\boldsymbol{\sigma}_{21i}$ are the state-specific mean and variance and ϕ_{σ_i} is the normal density with mean zero and standard deviation σ_i. For example, when $k = 2$, with probability p_1, $\mathscr{P}((y_t\ \hat{\mathbf{y}}_t')')$ is $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ while with probability $(1 - p_1)$, $\mathscr{P}((y_t\ \hat{\mathbf{y}}_t')')$ is $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

This specification allows us to introduce skew and kurtosis while controlling the first and second moments. In addition, elliptically symmetric distributions can be obtained

[12] One exception is provided by Weiss (1996) who assumes a uniform distribution and derives closed-form solutions for the optimal weights under asymmetric quadratic loss.

[13] Marron and Wand (1992) show how such mixtures can be used to generate a wide assortment of distributional shapes.

as a special case of the mixtures. Let $\mathbf{X}_t$ be an $m \times 1$ vector which has a multivariate normal mixture distribution $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with k states. Then if $\boldsymbol{\mu}_i = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_i = \gamma_i \boldsymbol{\Sigma}$ for some scalar γ_i then $\mathbf{X}_t$ is elliptical. This follows as for an elliptical distribution we need to be able to write the joint density as

$$f(\mathbf{x}) = h(|\mathbf{V}|)^{-1/2} g((\mathbf{x}-\mathbf{b})'\mathbf{V}^{-1}(\mathbf{x}-\mathbf{b})). \tag{21}$$

The joint density for $\mathbf{X}_t$ is

$$f(\mathbf{x}) = \sum_{i=1}^{k} p_i (2\pi)^{-n/2} |\boldsymbol{\Sigma}_i| \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}$$

$$= (2\pi^n |\boldsymbol{\Sigma}|)^{-1/2} \sum_{i=1}^{k} p_i \gamma_i \exp\left\{-\frac{\gamma_i}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Thus the result is shown for $\mathbf{V}=\boldsymbol{\Sigma}$, $\mathbf{b}=\boldsymbol{\mu}$, $h(\mathbf{V})=2\pi^n|\mathbf{V}|$, $g(s)=\sum_{i=1}^{k} p_i \gamma_i \exp\{-(\gamma_i/2)s\}$. When there exists at least one pair, $i \neq j$, for which $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$, the mixture generally falls outside the elliptically symmetric family.

5.1. An example

Many time series in economics and finance are skewed or fat-tailed. To disentangle the effect of skewness and kurtosis on the optimal forecast combination weights under asymmetric loss, we separately consider a model with skew and a model with kurtosis. Both models have two states. The skew model assumes the first state has a probability of 0.6 with distribution such that $\boldsymbol{\mu}_1 = \mathbf{0}$ and

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.2 & 0.15 \\ 0.2 & 0.25 & 0.125 \\ 0.15 & 0.125 & 0.2 \end{pmatrix}.$$

In state two $\boldsymbol{\mu}_1 = (0.5, 0.5, 0.5)'$ and $\boldsymbol{\Sigma}_2 = 0.1\boldsymbol{\Sigma}_1$.

Effectively the second state is shifted to the right and more tightly distributed around its mean. This example is chosen to match certain properties often found in macroeconomic and financial time series. The volatility of the forecasts of such series is typically an order of magnitude smaller than the volatility of the realizations. Furthermore, different forecasts are typically strongly correlated. We chose our parameters to match these two effects.[14] It is often the case that regime-switching models identify a high and a low volatility regime, cf. Perez-Quiros and Timmermann (2001), and we capture this by the scaling of the covariance matrix in the two regimes. This means that the high volatility state has volatility about three times higher than the low volatility regime.

[14] At 0.7, the correlation between the two forecasts, $\hat{y}_{1t}$ and $\hat{y}_{2t}$, is quite high.

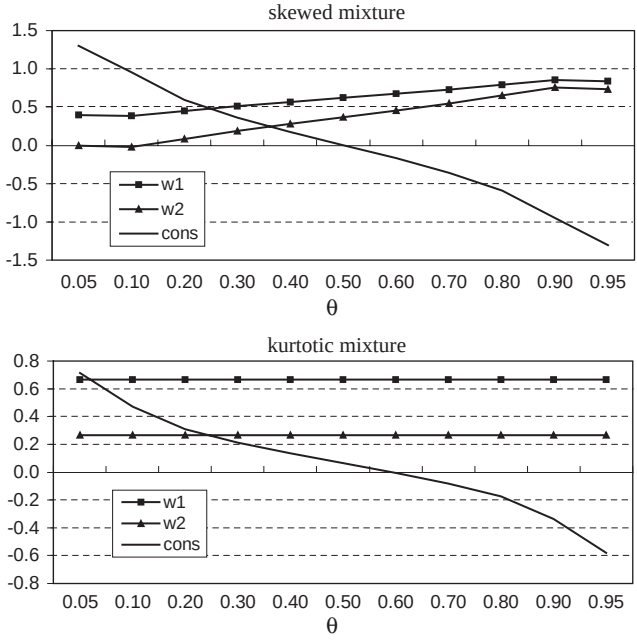


Fig. 1. Optimal combination weights and constant (lin–lin loss).

The kurtosis mixture model assumes that the probability of the first state is 0.2 with $\boldsymbol{\mu}_1 = (0.5, 0.5, 0.5)'$ and

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.2 & 0.15 \\ 0.2 & 0.25 & 0.125 \\ 0.15 & 0.125 & 0.25 \end{pmatrix}.$$

In state two $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1/15$.

The moments of each of the variables are:

	Outcome		1st forecast		2nd forecast	
	Skew	Kurtosis	Skew	Kurtosis	Skew	Kurtosis
Model 1	−0.574	3.722	−0.901	2.466	−0.942	2.213
Model 2	0.000	9.515	0.000	9.515	0.000	9.515

Model 1 has negative skew for all three variables but little or no excess kurtosis. Model 2 has strong kurtosis but no skew and is thus non-normal but nonetheless elliptically symmetric. The effect on the population weights of varying θ under lin–lin loss is shown in Fig. 1. The extent of the asymmetry clearly matters for the optimal combination weights in the skew model. In contrast—and consistent with Proposition 2—the combination weights derived under the kurtotic mixture model are independent of θ. Results are similar under asymmetric quadratic loss.

6. Estimation of forecast combination weights

In practice, forecast combination weights must be estimated from past data. This involves assumptions as to the stability over time of the joint distribution of the predicted variable and prediction signals. The empirical evidence of employing estimation methods based largely on least squares is surveyed in Clemen (1989) and further discussed by Diebold and Lopez (1996).

6.1. Moment estimators

In general we are interested in choosing the constant and the combination weights $\omega^c, \boldsymbol{\omega}$ that minimize the expected loss. The realized loss in period t, which we denote by $Q_t(\omega^c, \boldsymbol{\omega})$, can be written as

$$Q_t(\omega^c, \boldsymbol{\omega}) = L(\omega^c, \boldsymbol{\omega} | y_t, \hat{\mathbf{y}}_t, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ are the parameters of the loss function. ω^c and $\boldsymbol{\omega}$ can be obtained as an M-estimator based on the sample analog of $\mathrm{E}[Q_t]$ using n observations $\{y_t, \hat{\mathbf{y}}_t\}_{t=1}^n$:[15]

$$Q(\omega^c, \boldsymbol{\omega}) = n^{-1} \sum_{t=1}^{n} Q_t(\omega^c, \boldsymbol{\omega}).$$

An estimator that is available when we can interchange the expectation operator and the derivative of Q_t with respect to ω^c and $\boldsymbol{\omega}$,[16] $\partial Q_t / \partial((\omega^c \ \boldsymbol{\omega}')') = \mathbf{Q}_t'$, selects $\omega^c, \boldsymbol{\omega}$ to satisfy the first order condition

$$\mathrm{E}[\mathbf{Q}_t'(\omega^c, \boldsymbol{\omega})] = \mathbf{0}.$$

The corresponding sample analog is the average over the n observations. One problem with this method is that it can give rise to multiple solutions even when the M-statistic is regular.

A second method is related to instrumental variables estimation of a nonlinear regression and can therefore be handled in a GMM framework. The GMM approach estimates $\boldsymbol{\omega}$ by minimizing the quadratic form

$$V = \left(\sum_{t=1}^{n} \mathbf{Q}_t'(\omega^c, \boldsymbol{\omega}) \right)' \mathbf{A}^{-1} \left(\sum_{t=1}^{n} \mathbf{Q}_t'(\omega^c, \boldsymbol{\omega}) \right), \tag{22}$$

where $\mathbf{A}$ is some positive definite matrix. The optimal weighting matrix is standard from GMM problems. An advantage of the M-estimator is that there are a wealth of results available for consistency and asymptotic normality of the estimated weights. A similar result is true for the GMM estimator.

[15] Other examples of estimators of this form include maximum likelihood estimators, where Q is replaced by minus one times the log-likelihood.

[16] This assumption obviously imposes a restriction on the distribution of the forecast error.

6.1.1. Linex loss

These estimation methods can be demonstrated for the linex loss function. For this case, $Q_t(\omega^c, \boldsymbol{\omega})$ takes the form

$$Q_t(\omega^c, \boldsymbol{\omega}) = \exp(a(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t)) - a(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t) - 1.$$

GMM estimation is based on the derivative, $\mathbf{Q}_t'$:

$$\mathbf{Q}_t' = a(1 - \exp(a(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t))) \begin{pmatrix} 1 \\ \hat{\mathbf{y}}_t \end{pmatrix}. \tag{23}$$

The solution to this is equivalent to the nonlinear IV estimator obtained from the regression

$$1 = \exp\{a(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t)\} + u_t$$

using $(1, \hat{\mathbf{y}}_t')'$ as instruments.

6.1.2. Lin–lin loss

Recall that the lin–lin loss can be written

$$L(e_t) = \{(1-\theta)|e_t|1_{e_t>0} + \theta|e_t|1_{e_t\leqslant 0}\}.$$

This is precisely the loss function associated with the quantile regression problem, cf. Koenker and Basset (1978) and Buchinsky (1992).

Assuming independently and identically distributed forecast errors, the function to be minimized becomes

$$Q_t(\omega^c, \boldsymbol{\omega}) = \frac{1}{T}\left\{(1-\theta)\sum_{\{t:y_t>\omega^c+\boldsymbol{\omega}'\hat{\mathbf{y}}_t\}}|y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t| + \theta\sum_{\{t:y_t\leqslant\omega^c+\boldsymbol{\omega}'\hat{\mathbf{y}}_t\}}|y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t|\right\}.$$

Buchinsky (1992) shows that this can be rewritten to fit in a method of moments framework. Define the sign function $s(\lambda) = 1_{\lambda>0} - 1_{\lambda\leqslant 0}$. Then the objective function takes the form

$$Q_t(\boldsymbol{\omega}) = \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{2} - \theta + \frac{1}{2}s(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t)\right\}(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t). \tag{24}$$

Differentiating with respect to $(\omega^c, \boldsymbol{\omega})$, the first order conditions are

$$\mathbf{Q}_t' = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{2} - \theta + \frac{1}{2}s(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t)\right\} \\ \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{2} - \theta + \frac{1}{2}s(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t)\right\}\hat{\mathbf{y}}_t \end{pmatrix} = \mathbf{0}, \tag{25}$$

which again fits into the GMM Framework. Buchinsky (1992) suggests using the simplex method to solve for the parameters.

6.1.3. Asymmetric squared loss

The asymmetric squared loss function can be written as

$$L(e_t) = \{(1-\theta)e_t^2 1_{e_t>0} + \theta e_t^2 1_{e_t\leqslant 0}\}.$$

The loss function,

$$L^*(e_t) = |1-\theta - 1_{e_t\leqslant 0}|e_t^2 \tag{26}$$

is in the same homogenous class as $L(e_t)$ and the solution $(\omega^c, \boldsymbol{\omega}')'$ minimizing the two is the same. $L^*(e_t)$ is the loss function associated with the "expectile" regression problem examined in Newey and Powell (1987).

We choose the constant and forecast combination weights $(\omega^c, \boldsymbol{\omega}')'$ to minimize

$$Q_t(\omega^c, \boldsymbol{\omega}) = \frac{1}{T}\sum_{t=1}^{T} |1-\theta - 1_{y_t\leqslant \omega^c + \boldsymbol{\omega}'\hat{\mathbf{y}}_t}|(y_t - \omega^c - \boldsymbol{\omega}'\hat{\mathbf{y}}_t)^2. \tag{27}$$

This function is continuously differentiable and iterated weighted least squares can therefore be used to estimate the forecast combination weights. The estimator at iteration n becomes

$$\hat{\boldsymbol{\omega}}_n = \left(\sum_{t=1}^{T} k_{t,n-1}\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'\right)^{-1}\left(\sum_{t=1}^{T} k_{t,n-1}\tilde{\mathbf{y}}_t y_t\right), \tag{28}$$

where $\tilde{\mathbf{y}}_t = (1\ \hat{\mathbf{y}}_t')'$ and the scalar weights $k_{t,n-1}$ are constructed according to

$$k_{t,n-1} = |1-\theta - 1_{y_t<\hat{\boldsymbol{\omega}}_{n-1}'\tilde{\mathbf{y}}_t}|. \tag{29}$$

6.2. Estimation under elliptical symmetry

When the forecast errors are elliptically symmetric, a simple two stage procedure is suggested by our theoretical results in Section 3:

(i) Estimate $\hat{\boldsymbol{\omega}}$ by OLS, regressing y_t on a constant and $\hat{\mathbf{y}}_t$
(ii) Use the constructed variable $y_t - \hat{\boldsymbol{\omega}}'\hat{\mathbf{y}}_t$ to estimate $\hat{\omega}^c$ based on the relevant loss function (as above).

In this case the estimation problem is much simpler since the optimization procedure only has to search over a one dimensional parameter space regardless of the number of forecasts we are combining.

The methods require $\mathcal{P}((y_t\ \hat{\mathbf{y}}_t')')$ to be elliptically symmetric for optimality but may work well for minor deviations from this distribution. This suggests that it is important to have a test for elliptical symmetry. Zhu and Neuhaus (2000) propose nonparametric Monte Carlo tests that could be useful. Another strategy is to test for elliptical symmetry on the marginals. This is a useful procedure since, under joint elliptical symmetry, the marginals are also symmetric and departures from symmetry of the marginals imply that the joint distribution is not elliptically symmetric. Thus simple tests for skewness in the marginals will have power to detect deviations from elliptical symmetry.

Table 1
Average loss for linex loss function

a	Loss	M-est	OLS	OLSc	AVE	AVEc	For1	For1c	For2	For2c
Skewed mixture										
−1.25	0.28	1.08	1.36	1.08	1.30	1.05	1.35	1.09	1.46	1.16
−1.00	0.27	1.06	1.23	1.07	1.18	1.04	1.22	1.07	1.31	1.14
−0.75	0.26	1.05	1.14	1.05	1.10	1.03	1.13	1.06	1.21	1.13
−0.25	0.32	1.04	1.05	1.04	1.01	1.02	1.04	1.05	1.11	1.11
0.25	0.32	1.04	1.05	1.04	1.01	1.02	1.05	1.05	1.11	1.11
0.75	0.26	1.05	1.14	1.06	1.10	1.03	1.14	1.06	1.21	1.13
1.00	0.27	1.06	1.23	1.07	1.19	1.04	1.23	1.08	1.32	1.15
1.25	0.28	1.07	1.36	1.08	1.31	1.06	1.36	1.09	1.47	1.17
Kurtotic mixture										
−1.25	0.13	1.17	1.21	1.11	1.12	1.05	1.15	1.07	1.33	1.22
−1.00	0.12	1.13	1.16	1.10	1.08	1.04	1.10	1.05	1.25	1.20
−0.75	0.11	1.10	1.12	1.09	1.05	1.03	1.06	1.05	1.20	1.18
−0.25	0.16	1.08	1.08	1.08	1.02	1.02	1.03	1.04	1.15	1.16
0.25	0.16	1.08	1.08	1.08	1.02	1.02	1.03	1.04	1.15	1.16
0.75	0.11	1.11	1.12	1.09	1.05	1.03	1.06	1.05	1.20	1.18
1.00	0.12	1.13	1.16	1.10	1.08	1.04	1.10	1.05	1.25	1.20
1.25	0.13	1.17	1.21	1.11	1.12	1.05	1.15	1.07	1.33	1.22

Note: The first column reports the value of the parameter, a, controlling asymmetry. The second column gives the population expected loss scaled by the square of the asymmetry parameter. Other entries give average loss over 5,000 Monte Carlo replications based on estimates of the forecast combination weights and the constant from 100 observations. These losses are scaled by the population loss. OLS refers to the estimation of the combination weights by OLS; AVE refers to the equal-weighted forecast combination. For$i$ refers to forecast i alone. Names ending with c refer to the corresponding estimation technique for the weights followed by optimal estimation of the constant.

7. Monte Carlo results

This section reports on the performance of the various estimation procedures applied to the linex, lin–lin and asymmetric quadratic loss functions under the two mixture models in Section 5. In each case we combine two forecasts and assume that a sample of 100 observations is available to estimate the combination weights. All results are based on 5,000 Monte Carlo simulations.

7.1. Expected losses

Tables 1–3 examine the expected loss under linex, lin–lin and asymmetric loss. We report the ratio of the expected loss under the estimated weights relative to the loss under the optimal population weights:

$$\frac{\mathrm{E}[L(e|\hat{\boldsymbol{\omega}})]}{\mathrm{E}[L(e|\boldsymbol{\omega}^{\mathrm{opt}})]}.$$

Results are shown for each of the estimation methods considered in Section 6. The first column reports the value of the asymmetry parameter of the loss function. The

Table 2
Average loss for lin–lin loss function

θ	Loss	Quantile	OLS	OLSc	AVE	AVEc	For1	For1c	For2	For2c
Skewed mixture										
0.1	0.13	1.06	2.08	1.08	2.04	1.07	2.07	1.09	2.13	1.12
0.2	0.19	1.03	1.39	1.06	1.36	1.05	1.38	1.07	1.42	1.10
0.3	0.23	1.03	1.15	1.04	1.13	1.03	1.15	1.04	1.18	1.07
0.4	0.26	1.03	1.05	1.02	1.03	1.01	1.05	1.03	1.08	1.06
0.5	0.26	1.03	1.02	1.02	1.00	1.01	1.02	1.02	1.05	1.05
0.6	0.26	1.03	1.05	1.03	1.03	1.01	1.04	1.03	1.08	1.06
0.7	0.23	1.03	1.15	1.04	1.13	1.02	1.14	1.04	1.18	1.07
0.8	0.19	1.03	1.38	1.06	1.35	1.05	1.37	1.06	1.41	1.09
0.9	0.13	1.06	2.06	1.08	2.03	1.07	2.06	1.08	2.12	1.12
Kurtotic mixture										
0.1	0.08	1.08	1.85	1.05	1.79	1.02	1.80	1.03	1.91	1.09
0.2	0.11	1.05	1.35	1.04	1.30	1.01	1.31	1.02	1.39	1.08
0.3	0.13	1.04	1.15	1.04	1.11	1.01	1.12	1.02	1.19	1.08
0.4	0.14	1.03	1.07	1.04	1.03	1.01	1.04	1.02	1.10	1.08
0.5	0.15	1.03	1.04	1.04	1.00	1.01	1.01	1.02	1.07	1.08
0.6	0.14	1.03	1.07	1.04	1.03	1.01	1.04	1.02	1.10	1.08
0.7	0.13	1.04	1.15	1.04	1.11	1.01	1.12	1.02	1.19	1.08
0.8	0.11	1.05	1.35	1.04	1.30	1.01	1.31	1.02	1.39	1.08
0.9	0.08	1.08	1.85	1.05	1.79	1.02	1.80	1.03	1.91	1.09

Note: The first column reports the value of the parameter, θ, controlling asymmetry. Other columns are as in Table 1.

second column gives the expected loss using the population weights. The third column evaluates the expected loss when the forecast combination weights and the constant are based on the preferred (among estimators that consistently estimate the weights) estimation method. The fourth and fifth columns give results from OLS estimation and OLS estimation of the weights followed by estimation of the constant term (which is the procedure proposed above if we knew the data were elliptically symmetric), respectively.[17] We also report losses for the average (equal-weighted) and individual forecasts (unit or zero weight on each forecast) as well as losses for each of these where an optimally estimated constant is employed to obtain the correct bias (these are subscripted c in the tables). Separate results are reported for the skewed and kurtotic mixture models.

The first forecast is quite good on its own, the second quite a bit less so. The fact that combination methods do better than simply using the first forecast on its own highlights the value of the forecast combination approach. Consistent with theory, even when the first forecast is obviously better than the second, the latter can still be very useful in constructing a better combined forecast.

The results are similar qualitatively—and to an extent quantitatively as well—across loss functions. Under both the skewed and kurtotic mixture models the 'optimal'

[17] The expectation is taken with respect to the true distribution of the forecast errors.

Table 3
Average loss for asymmetric quadratic loss function

θ	Loss	WLS	OLS	OLSc	AVE	AVEc	For1	For1c	For2	For2c
Skewed mixture										
0.1	0.14	1.06	1.89	1.08	1.82	1.05	1.88	1.09	1.99	1.15
0.2	0.20	1.05	1.34	1.06	1.30	1.03	1.34	1.07	1.42	1.13
0.3	0.24	1.04	1.15	1.05	1.11	1.02	1.14	1.05	1.21	1.12
0.4	0.26	1.04	1.06	1.04	1.03	1.02	1.06	1.05	1.12	1.11
0.5	0.26	1.04	1.04	1.04	1.00	1.01	1.03	1.04	1.10	1.11
0.6	0.26	1.04	1.06	1.04	1.03	1.02	1.06	1.05	1.12	1.11
0.7	0.24	1.04	1.15	1.05	1.11	1.02	1.14	1.05	1.21	1.11
0.8	0.20	1.05	1.34	1.06	1.29	1.03	1.33	1.06	1.41	1.13
0.9	0.15	1.07	1.88	1.07	1.82	1.05	1.87	1.08	1.98	1.14
Kurtotic mixture										
0.1	0.07	1.16	1.65	1.09	1.55	1.04	1.58	1.06	1.76	1.18
0.2	0.09	1.11	1.29	1.08	1.22	1.03	1.24	1.04	1.38	1.17
0.3	0.10	1.09	1.16	1.08	1.09	1.03	1.11	1.04	1.24	1.16
0.4	0.10	1.08	1.09	1.08	1.03	1.02	1.05	1.04	1.17	1.16
0.5	0.10	1.08	1.08	1.08	1.01	1.02	1.03	1.04	1.15	1.16
0.6	0.10	1.08	1.09	1.08	1.03	1.02	1.05	1.04	1.17	1.16
0.7	0.10	1.09	1.16	1.08	1.09	1.03	1.11	1.04	1.24	1.16
0.8	0.09	1.11	1.29	1.08	1.22	1.03	1.24	1.04	1.38	1.17
0.9	0.07	1.16	1.65	1.09	1.55	1.04	1.58	1.06	1.76	1.18

Note: Columns as in Table 2.

estimators (M estimation for linex, quantile regression for lin–lin and WLS for asymmetric squared loss) produce expected losses that are 3–16% higher than the expected loss under population parameter values. For the skewed mixture model, use of the suboptimal OLS weights can entail an increase in the average loss of more than 100 percent relative to the loss under the population parameter values and grows as the degree of asymmetry in the loss function increases. The loss from using OLS estimates remains large in the elliptically symmetric model due to the use of an inconsistent estimator of the constant term. However, use of the OLS weights with a corrected (optimal) estimate of the constant entails a fairly small loss over the optimal estimator and sometimes even has a better small sample performance.[18] This follows from the good properties of the OLS estimator for this Monte Carlo design. As expected this is even more the case for the elliptically symmetric kurtosis model for which OLS consistently estimates the optimal weights.

[18] Note that the 'optimal bias' in the constant term adjusts for the non-optimal estimation of the combination weights. Under the 'constant adjustment' strategy the estimated forecast combination weights, and hence the relevance of the individual forecasts in the construction of the combined forecast, are the same. This means that if one uses this method (which is generally simpler in practice and also less dependent on the shape and parameters of the loss function) then the usefulness of the individual forecasts rests entirely on their usefulness in the mean square error loss forecast combination.

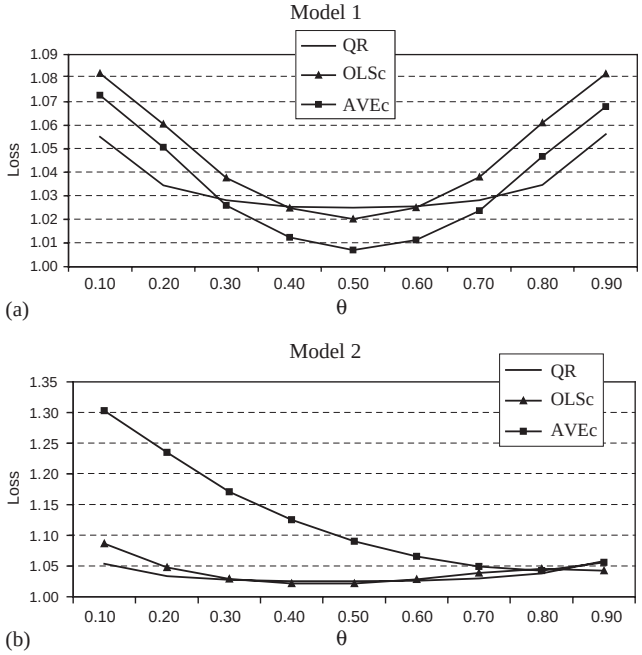


Fig. 2. Average relative loss: (a) Model 1; (b) Model 2. Note: Model 1 refers to the skewed mixture. Model 2 is as the skewed mixture but with a different variance–covariance matrix. The modified entries for state 1 are: $(1,3) = -0.05$; $(2,2) = 0.5$; and $(3,3) = 1/3$.

Similar results arise under averaging which performs poorly on its own but does much better after the bias adjustment. When the intercept is corrected (using the relevant estimator for each loss function) the losses perform very well in the Monte Carlo design examined here. This, however, strongly depends on the underlying distribution of the data. To demonstrate this point Fig. 2a graphs the average loss for the quantile regression, OLSc and AVEc against the asymmetry parameter, θ. For this case the AVEc method does very well against the other methods. The 'folk theorem' on the superiority of using even weights reflects that when the weights are 'close' to even, imposing this constraint does not result in a large bias and reduces the variance due to estimation error. However, more diverse weights overturn this result, as can be seen in Fig. 2b. In this case we have designed the Monte Carlo so that the weight on the first forecast ranges from 0.4 to 0.6 (close to one-half) but the weight on the second forecast ranges from -0.5 to 0. In this case the constant adjusted average weighting method does extremely poorly. More generally it is easy to find cases where this method does well and cases where it does poorly. Note that the OLSc method still works very well for this Monte Carlo design.

We also examine Diebold and Mariano (1995) tests that compare various estimators, reporting results for the lin–lin case only as results for the other loss functions are similar. Table 4 shows rejection frequencies for a size of 5% using the quantile

Table 4
Diebold–Mariano tests for lin–lin loss function

θ	OLS	OLSc	AVE	AVEc	For1	For1c	For2	For2c
Skewed mixture								
0.1	1.00	0.13	1.00	0.22	1.00	0.27	1.00	0.39
0.2	1.00	0.08	1.00	0.21	1.00	0.29	1.00	0.43
0.3	0.83	0.01	0.82	0.10	0.86	0.18	0.92	0.38
0.4	0.13	0.00	0.26	0.05	0.37	0.13	0.57	0.33
0.5	0.00	0.00	0.06	0.03	0.16	0.11	0.38	0.31
0.6	0.12	0.00	0.26	0.05	0.38	0.13	0.58	0.32
0.7	0.81	0.02	0.83	0.09	0.87	0.18	0.92	0.36
0.8	1.00	0.07	1.00	0.17	1.00	0.24	1.00	0.42
0.9	1.00	0.12	1.00	0.20	1.00	0.26	1.00	0.37
Kurtotic mixture								
0.1	1.00	0.01	1.00	0.05	1.00	0.07	1.00	0.20
0.2	1.00	0.01	1.00	0.05	0.99	0.08	1.00	0.29
0.3	0.81	0.00	0.83	0.05	0.84	0.08	0.93	0.34
0.4	0.19	0.00	0.29	0.06	0.32	0.09	0.64	0.39
0.5	0.01	0.00	0.09	0.06	0.12	0.09	0.47	0.40
0.6	0.20	0.01	0.28	0.05	0.32	0.08	0.62	0.38
0.7	0.81	0.01	0.83	0.05	0.84	0.07	0.92	0.35
0.8	0.99	0.01	0.99	0.05	1.00	0.07	1.00	0.28
0.9	1.00	0.01	1.00	0.06	1.00	0.08	1.00	0.20

Note: The first column reports the value of the parameter, a, controlling asymmetry. Other entries give the percentage of rejections of one-sided Diebold–Mariano tests (5% significance level) of the optimal combination vs. the corresponding entry. The tests are based on estimates of the forecast combination weights and the constant from 100 observations and 5,000 Monte Carlo replications. OLS refers to the estimation of the combination weights by OLS; AVE refers to equal-weighted forecast combination. For$i$ refers to forecast i alone. Names ending with c refer to the corresponding estimation technique for the weights followed by optimal estimation of the constant.

estimator as the base model and testing the null that the alternate methods give similar results against the one-sided alternative that the other methods are worse. A result of 0.05 indicates that other methods perform equally well, a larger frequency of rejection suggests that the base estimator does better and a number lower than 0.05 suggests that it does better for less than the expected number of cases under the null. In most cases the OLSc method produces results around zero suggesting either that we would prefer the OLSc method if we used this test or (far more likely) that the test is not sufficiently powerful to distinguish between the forecasts from the two methods. The exception is when there is extreme loss asymmetry. The AVEc method tends also to do well, especially under symmetry. But for the most asymmetric cases it is found to be worse by this metric. Part of the reason for this is that the AVEc forecasts are less variable than the other forecasts and thus contribute less to the standard error. Hence when this method is better it is more likely to be statistically significantly so. These results, as we expect from the theory in Section 3, are more pronounced for the skewed mixture than for the elliptically symmetric kurtotic model.

Table 5
Small sample results for lin–lin loss function

	Population weights			Quantile estimate bias			OLS estimate bias			OLSc	AVEc
θ	ω_1	ω_2	ω^c	ω_1	ω_2	ω^c	ω_1	ω_2	ω^c	ω^c	ω^c
Skewed mixture											
0.1	0.39	−0.02	0.95	0.00	0.03	0.00	0.23	0.39	−0.95	−0.12	−0.10
0.2	0.45	0.08	0.59	0.00	0.01	0.00	0.17	0.28	−0.59	−0.10	−0.10
0.3	0.51	0.19	0.36	0.00	0.01	0.00	0.11	0.18	−0.36	−0.10	−0.10
0.4	0.57	0.28	0.17	0.00	0.01	0.00	0.05	0.08	−0.17	−0.05	−0.06
0.5	0.62	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.67	0.46	−0.17	0.00	0.00	0.00	−0.05	−0.09	0.17	0.05	0.05
0.7	0.73	0.55	−0.36	0.00	0.00	0.00	−0.11	−0.18	0.36	0.10	0.09
0.8	0.79	0.65	−0.59	0.00	−0.01	0.00	−0.17	−0.29	0.59	0.13	0.12
0.9	0.85	0.76	−0.95	0.00	−0.04	0.00	−0.23	−0.39	0.95	0.11	0.10
Kurtotic mixture											
0.1	0.67	0.27	0.47	−0.01	0.00	0.02	−0.01	0.00	−0.39	0.01	−0.07
0.2	0.67	0.27	0.31	−0.01	0.00	0.01	−0.01	0.00	−0.23	0.02	−0.06
0.3	0.67	0.27	0.21	−0.01	0.00	0.01	−0.01	0.00	−0.13	0.01	−0.07
0.4	0.67	0.27	0.14	−0.01	0.00	0.01	−0.01	0.00	−0.06	0.01	−0.07
0.5	0.67	0.27	0.07	0.00	0.00	0.01	−0.01	0.00	0.01	0.01	−0.07
0.6	0.67	0.27	0.00	0.00	−0.01	0.01	−0.01	0.00	0.08	0.01	−0.06
0.7	0.67	0.27	−0.08	0.00	−0.01	0.01	−0.01	0.00	0.16	0.01	−0.06
0.8	0.67	0.27	−0.18	0.00	0.00	0.01	−0.01	0.00	0.25	0.01	−0.06
0.9	0.67	0.27	−0.34	0.00	−0.01	0.00	−0.01	0.00	0.42	0.01	−0.06

Note: The first column reports the value of the parameter, θ, that controls the asymmetry of the loss function. The second through fourth columns give the population values of the optimal forecast combination weights and the constant. All other entries are bias estimates based on a sample of 100 observations and 5,000 Monte Carlo replications.

7.2. Biases in the estimators

Table 5 shows both population values and the bias arising from various estimation methods under lin–lin loss. For the other loss functions results were again similar. The first column shows the value assumed for the asymmetry parameter in the loss function (θ). The next three columns show the population values of the optimal weights. Then follows the bias in the estimates from the preferred estimation method and the bias arising from ordinary least squares estimation. The final columns report the bias in the constant term when the forecast combination weights are estimated by OLS or by averaging followed by optimal estimation of the constant.

For the lin–lin loss function, the quantile regression method only introduces very minor biases in the estimated forecast combination weights and the constant term, irrespective of whether the data generating process is the skewed or kurtotic mixture model and even under very large degrees of asymmetry in the loss

function.[19] The OLS estimator, optimal when $\theta = 0.5$, is quite heavily biased even for small departures from symmetric loss. The biases in the constant are mitigated to a large extent when estimating the forecast combination weights by OLS followed by quantile estimation of the constant.[20] Nevertheless, the estimates have a bias of around 10% for the most asymmetric loss functions and skewed data.

These simulations show that, on the whole, the simple two-step procedure that estimates the combination weights by OLS followed by intercept correction tends to work very well, provided that the degree of asymmetry in the loss function is not very large. This procedure is of course very simple to use in practice and eliminates most of the loss from using the simple OLS estimator to estimate both the combination weights and the intercept.

These results also carry over to data that simultaneously display skew and kurtosis. Table 6 shows results for a mixture model that sets $p_1 = p_2 = 0.5$ has the same covariance matrices as in the kurtosis mixture model, zero means in state 1 and means of -0.25 in state 2. This model matches more closely the moments of the empirical data in the next section. We report results for the asymmetric quadratic loss function but the results are very similar for linex and lin–lin loss. OLS continues to generate substantial biases in the parameter estimates under asymmetric loss and this leads to a large increase in the expected loss. In contrast the weighted least squares estimator or the bias-adjusted least squares method produce only small biases and relatively modest values of the expected loss.

8. Empirical application to inflation forecasting

To demonstrate the theoretical ideas developed so far, we consider an application that combines predictions of changes in the consumer price index (CPI) from a simple autoregressive time-series model with survey predictions. The source of the latter is the Livingston Survey data base maintained by the Federal Reserve Bank of Philadelphia.[21] This provides a time series of predictions of the consumer price index six months ahead in time over the period 1946:I to 2001:I. All forecasts are thus one-step-ahead for a forecast horizon of 6 months. The forecasts are likely to be based on diverse information sets that comprise a much larger set of public and private information than is typically considered in econometric models.

We combine the aggregate survey predictions[22] with predictions from a simple autoregressive forecasting model. Schwarz's information criterion supported a simple first-order autoregressive (AR(1)) model which we use to predict the inflation

[19] It can be difficult to interpret the absolute size of the bias in these simulations as it depends on the scaling of the problem which is arbitrary. We use the terms "small" and "large" bias in a relative sense, i.e. relative to the biases observed in the other simulations.

[20] This reduces to a simple LAD estimation problem.

[21] This survey data has previously been used in numerous studies. For a comprehensive list of these, see the web site maintained by the Federal Reserve Bank of Philadelphia.

[22] These are computed as the arithmetic mean of the individual forecasters' predictions.

Table 6
Small sample results for asymmetric quadratic loss: mixture with skew and kurtosis matching empirical data

Small sample results

	Population weights			WLS estimate bias			OLS Estimate bias			OLSc	AVEc
θ	ω_1	ω_2	ω^c	ω_1	ω_2	ω^c	ω_1	ω_2	ω^c	ω^c	ω^c
0.2	0.69	0.48	0.36	0.00	−0.01	−0.01	−0.07	−0.12	−0.36	−0.03	−0.02
0.5	0.62	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.8	0.55	0.25	−0.36	0.00	0.00	0.01	0.07	0.12	0.36	0.02	0.02

Average loss

θ	Loss	WLS	OLS	OLSc	AVE	AVEc	For1	For1c	For2	For2c
0.2	0.18	1.06	1.31	1.06	1.26	1.03	1.29	1.06	1.37	1.12
0.5	0.22	1.05	1.05	1.05	1.00	1.01	1.03	1.04	1.10	1.11
0.8	0.17	1.06	1.31	1.06	1.26	1.03	1.30	1.06	1.37	1.12

DM tests

θ	OLS	OLSc	AVE	AVEc	For1	For1c	For2	For2c
0.2	0.90	0.00	0.77	0.03	0.82	0.09	0.88	0.24
0.5	0.01	0.01	0.03	0.02	0.09	0.07	0.29	0.25
0.8	0.90	0.00	0.78	0.03	0.82	0.10	0.88	0.26

Note: The first column of each panel reports the value of the parameter, θ, controlling asymmetry. For the first panel, the second through fourth columns give the population values of the optimal forecast combination weights and the constant. All other entries are bias estimates. For the second panel, the second column gives the population expected loss. Other entries give average loss based on estimates of the forecast combination weights and the constant. These losses are scaled by the population loss. For the third panel, the entries give the % of rejections of one-sided Diebold–Mariano tests (5% significance level) of the optimal combination vs. the corresponding entry. For all panels, the estimations and tests are based on estimates of the forecast combination weights and the constant from 100 observations and 5,000 Monte Carlo replications. OLS refers to the estimation of the combination weights by OLS; AVE refers to equal-weighted forecast combination. For*i* refers to forecast *i* alone. Names ending with *c* refer to the corresponding estimation technique for the weights followed by optimal estimation of the constant.

rate—defined as the log-difference in the *CPI*, ΔCPI — one-step, or six-months, ahead. One-step-ahead forecasts from the autoregressive model are computed using recursively updated parameter estimates.[23]

The estimated correlation between the forecast errors from the autoregressive model and the Livingston survey data is 0.63 over the full sample. While high, this estimate only reflects the linear correlation between the two sets of forecast errors. Inspection of the density of the forecast errors from the AR(1) model and the Livingston survey data suggested that the two forecast error distributions have more probability mass in the center and in the tails than a Gaussian distribution, thus indicating leptokurtosis.

[23] Two observations are used to define the first-difference of the logarithm of the CPI and its lagged value and we use the first 50 observations as an initial sample for parameter estimation.

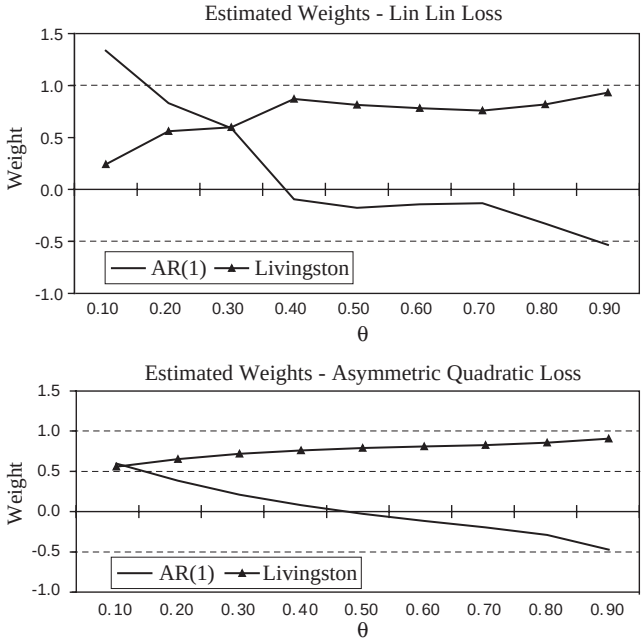


Fig. 3. Forecast combination weights under different loss functions.

The density generated by the autoregressive model also has more probability mass in the right shoulder and tail than the density associated with the Livingston data which has a thicker left shoulder.

The two sets of forecasts thus contain very different information. For small values of θ the Livingston forecasts were found to strongly outperform the AR(1) forecasts. As θ gets larger, this difference declines. When θ is 0.7 or larger the AR(1) forecasts outperform the Livingston forecasts in terms of the average loss over the sample.

To explore the importance the loss function plays in determining the optimal combination weights, we varied the asymmetry parameter, θ, between 0.1 and 0.9. This moves us from strong aversion against large positive forecast errors via a symmetric loss function towards strong aversion against large negative forecast errors. The optimal combination weights under lin–lin loss are shown in the upper window of Fig. 3. Under standard, symmetric loss ($\theta = 0.5$), the optimal weight on the AR(1) forecast is close to zero. However as the loss function becomes increasingly asymmetric (in either direction), it becomes optimal to put a non-zero weight on the time-series forecast. For small values of θ a positive weight is put on both the AR(1) forecasts and the density forecasts. For large values of θ, the optimal weight on the AR(1) forecasts becomes negative while the weight on the Livingston data increases. Strong aversion against positive forecast errors thus means assigning positive combination weights to both the

Table 7
Descriptive statistics and skew and Jarque–Bera tests for normality of forecast errors

Series	Mean	Std	Skew	Kurtosis	Skew test	JB test
Actual	0.020	0.016	1.119	4.238	19.621	25.621
AR(1)	0.017	0.009	1.237	4.769	23.954	36.217
Livingston	0.021	0.016	0.727	3.524	8.273	9.348
$\theta = 0.1$	−0.014	0.009	−1.028	4.080	7.918	10.107
$\theta = 0.5$	−0.006	0.009	−0.567	3.635	2.409	3.164
$\theta = 0.9$	0.002	0.009	0.305	3.441	0.700	1.064

Note: The first to fourth columns are estimates of the mean, standard deviation, skew (zero is no skew) and kurtosis (a value of three is no excess kurtosis), respectively. Reported in the last two columns are Skew tests and Jarque–Bera tests for normality respectively. The 95% critical value is 3.84 for the Skew tests and 5.99 for the JB tests. Both tests reject for larger values. The first three rows are for the realized series (actual) and forecasts, respectively. Values in rows marked θ refer to forecast errors estimated using weights under the asymmetric quadratic loss function.

autoregressive and Livingston predictions. Under strong aversion against negative forecast errors the combination weight on the autoregressive model is strongly negative while the weight on the Livingston prediction is strongly positive. The lower window of Fig. 3 repeats this exercise for the asymmetric quadratic loss function. The actual estimates for the weights differ somewhat but the story is basically the same as under lin–lin loss.

One way to understand the negative weight on the forecasts from the AR(1) model when θ is large is by considering scenarios under which large negative forecast errors occur. The tail correlation between the autoregressive and the Livingston forecast errors is greater than the correlations in the center of the forecast error distribution. To avoid large negative forecast errors (which occur under positive combination weights if the two forecast errors are simultaneously negative or under negative weights if the forecast errors are simultaneously positive), the optimal weights have opposite signs. Combinations with weights of opposite signs are similar to setting up a spread in a financial trading strategy, a practice commonly used when return correlations are higher the more volatile markets are. Notice, however, that while large positive forecast errors are more common for the AR(1) model than in the Livingston data, the positive weight on the Livingston forecast is almost twice as large as the negative weight on the AR(1) forecast.

Since the forecast combination weights are very sensitive to the parameters of the loss function, we expect that the distribution of the forecast errors (and indeed the individual forecasts as well as the change in inflation) are far from symmetric. In Table 7 we provide descriptive statistics for the first four moments of the various forecasts and also test for skew and kurtosis for each of the underlying variables as well as the forecast errors for three values of θ, namely $\theta = 0.1, 0.5, 0.9$. These tests merely serve the purpose of characterizing the distribution of the various forecasts. As expected we strongly reject the null of no skew for the individual forecasts and the actual time series while this null is only rejected for the forecast error distribution

Table 8
Average losses from combined and equal weighted forecasts

	WLS estimates					Losses		
θ	$w1$	$w2$	Constant	OLSc constant	AVEc constant	WLS	OLSc	AVEc
0.1	0.594	0.595	0.006	0.013	0.007	0.028	0.028	0.023
0.5	−0.001	0.824	0.004	0.004	0.002	0.053	0.053	0.051
0.9	−0.428	0.926	0.002	−0.003	−0.004	0.027	0.037	0.035
θ	ω	ω^{AR}	ω^{A}	ω^{B}	ω^{I}	WLS	Equal weights	
0.1	0.53	0.56	0.00	−0.27	0.87	0.027	0.033	
0.5	0.61	−0.03	0.30	−0.09	0.61	0.052	0.045	
0.9	0.23	−0.51	0.48	−0.10	0.60	0.026	0.057	

Note: WLS estimates are estimates of the parameters using the WLS method described for the asymmetric quadratic loss function with asymmetry parameter θ. The loss reported is average out of sample one step ahead loss. In the OLSc case OLS weights are used for the forecast combination weights but the constant is estimated using WLS. AVEc gives results for even weights on the forecasts and a constant estimated using WLS. The second panel is based on combinations of the autoregressive (AR) and disaggregated academic (A), bank (B) and industry (I) forecasts.

when $\theta = 0.1$ and a strong negative skew emerges. Notice also the shift from negative to positive skew as θ increases from 0.1 to 0.9.[24]

Finally, an implication of the Monte Carlo results in Section 7 is that simply getting the constant correct (imposing the correct bias) is a major part of the attempt to obtain combined forecasts that perform relatively well. In the numerical work we confirmed that simply using the OLS weights on the actual forecasts with the 'optimal' method for estimating the constant entailed only minor additional losses. The top panel of Table 8 examines this possibility with the inflation rate data using out-of-sample forecast errors under asymmetric quadratic loss. Except when $\theta = 0.9$, the average loss from using OLS forecast combination weights and WLS to estimate the constant is very similar to the loss based on the iterative WLS estimates. When $\theta = 0.9$ the WLS estimation method does lead to better results than both the OLS and the averaging methods followed by adjustment.

8.1. Disaggregated survey predictions

The Livingston survey data lists predictions according to the forecasters' broad affiliation. The most frequent affiliations are nonfinancial business ('industry'), academic institution ('academic') and commercial and investment banking ('banking'). Forecasters from the same area of business are perhaps more likely to use the same information sets and/or models, so we next consider how the weights on these disaggregated

[24] We also computed forecast encompassing tests. It was generally rejected that the AR(1) model forecast encompasses the Livingston predictions, while the reverse could not be rejected at standard levels of significance.

forecasts depend on the loss function. We compute averages of the optimal combination weights on the AR(1) predictions and the three survey predictions as a function of θ. All predictions are out-of-sample and the superscripts AR, A, B, I refer to the autoregressive, academic, banking and industry predictions. The bottom panel of Table 8 shows the outcome. When the loss function assigns a very small weight to large negative forecast errors ($\theta = 0.1$), the academic forecasts get a zero weight. Otherwise this weight is large and positive. The weight on the 'banking' inflation forecasts is always negative and tends to be small. Most weight is put on the forecasts produced by professionals in nonfinancial business. These consistently receive a weight above 0.5 irrespective of the shape of the loss function. The weight on the autoregressive forecast does not appear to be much affected by using disaggregate rather than aggregate survey data. It continues to be large and positive when $\theta = 0.1$ and large and negative when $\theta = 0.9$.

We finally investigated whether the often quoted finding that simple averages of forecasts outperform estimated optimal weights depends on the shape of the loss function. Our results—reported in the bottom panel of Table 8—confirm the frequent finding that equal-weights outperform estimated optimal weights under MSE loss ($\theta = 0.5$). However, they also show very clearly that this result is overturned under asymmetric loss where use of estimated optimal weights leads to far smaller average losses out-of-sample.

In an analysis of forecasts of GNP/GDP growth from the Blue Chip Economic Indicators, Laster et al. (1999) find evidence that independent forecasters are more likely to produce 'extreme' forecasts that differ the most from the consensus or median forecast. They propose an economic model where such behaviour is rational. Forecasters attract two types of clients, namely intensive users (who demand consistently precise forecasts) and occasional users that select last period's star forecaster, i.e. the forecaster whose prediction was closest to the outcome. Producers of forecasts have an incentive to attract new occasional use clients. They can only do this by being more accurate in their prediction than other forecasters. This gives an incentive to generate extreme forecasts. The incentive is stronger, the more the forecaster relies on occasional use clients. Differences in opinions across forecasters may thus be driven by forecasters with different ratios of occasional and intensive forecast users. This story is consistent with our framework where different forecasts are driven either by different information sets or by different loss functions. In fact, if the forecaster producers act as agents for forecast users, asymmetric loss could well be caused by differences across forecasters in the ratios of occasional and intensive use clients.

9. Conclusion

Several conclusions emerge from our analysis. If the forecast error distribution is elliptically symmetric then the forecast combination weights are identical for a wide range of loss functions, including the popular mean squared error loss function. This implies that if a particular forecast is informative under MSE loss, then this forecast will be informative for a very wide range of loss functions. If there is little departure

from elliptical symmetry, then the optimal forecast combination weights are near the standard weights derived under MSE loss. This has practical implications since such weights are often much easier to compute.

Under departures from elliptically symmetric distributions, the optimal combination weights under asymmetric loss can be very different than under MSE loss. For example, if negative forecast errors are associated with much lower losses than positive ones, then it will be optimal to select combination weights that give rise to a much larger upward bias than under MSE loss.

Irrespective of the distribution of the forecast errors, the constant term in the forecast combination is affected by the loss function. Our simulations suggest that even with heavily skewed forecast errors, the loss from using MSE combination weights (which are suboptimal) and optimally adjusting the constant results in fairly small increases in expected loss. Since this strategy is simple and has good small sample properties, for many applications this approach will perhaps work well.

Our empirical application showed that the optimal weights in a combination of inflation survey forecasts and forecasts from a simple autoregressive model strongly depend on the degree of asymmetry in the loss function. In the absence of loss asymmetry, the autoregressive forecast does not add much information. However, under asymmetric loss (in either direction), both sets of forecasts appear to contain information and have non-zero weights in the combined forecast. The application confirmed the frequent finding that equal-weights outperform estimated optimal weights under MSE loss. However, it also showed very clearly that this result can be overturned under asymmetric loss where use of estimated optimal weights led to much smaller average losses out-of-sample.

There are many directions for generalizing our results. While we do not study time-variations in higher order conditional moments, it is possible to extend our framework to include such effects. Elliott and Timmermann (2002) consider a framework with time-varying mixtures of Gaussian distributions that generate time-varying higher order moments. First order conditions for the optimal combination weights can be derived under a variety of loss functions. In general these depend on the full set of parameters that determine the conditional density. In this framework the optimal intercept and combination weights become time-varying. Many of the ideas in this paper can also be applied to multi-step-ahead forecasting. Some GARCH models imply that the one-step-ahead forecast errors are elliptically symmetric, while multi-step-ahead forecast errors are not. This raises the possibility of letting the estimation method depend on the forecast horizon.

Another possibility is to adopt the Bayesian methods to forecast combination that have been developed by Palm and Zellner (1992), including shrinkage methods. There are good theoretical reasons to consider such methods. Non-Bayesian estimators need not be admissible in cases where Bayesian methods may have this property. Similarly, Bayesian model averaging can be shown to provide better predictive ability than individual models along a logarithmic scoring rule. These points would suggest potentially important gains from using Bayesian methods. In their extensive analysis of a wide range of prediction models for international growth rates Min and Zellner (1993) did not find much evidence that combination methods improved MSE

performance. However they did find that the best Bayesian combination methods dominated their non-Bayesian counterparts. Similarly, recent work by Raftery et al. (1997) finds promising gains from Bayesian model averaging techniques in cases with many predictors. In future work we plan to extend these methods to cover asymmetric loss functions such as the ones considered in this paper.

Acknowledgements

We wish to thank three anonymous referees, an associate editor and the editor, Arnold Zellner, for making many constructive suggestions. Seminar participants at Bocconi and Emory University and the 2001 NBER conference on macroeconomics and forecasting, particularly Frank Diebold, Carlo Favero, Massimiliano Marcellino and Paul Ruud, also provided helpful comments. Carlos Capistran provided excellent research assistance. We are grateful for financial support from the NSF under grant SES 0111238.

References

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Operations Research Quarterly 20, 451–468.

Buchinsky, M., 1992, Methodological issues in quantile regression. Manuscript, Brown University.

Chamberlain, G., 1993. A characterization of the distributions that imply mean–variance utility functions. Journal of Economic Theory 29, 185–201.

Chan, Y.L., Stock, J., Watson, M.W., 1999. A dynamic factor model framework for forecast combination. Spanish Economic Review 1, 91–121.

Christoffersen, P., Diebold, F.X., 1997. Optimal prediction under asymmetrical loss. Econometric Theory 13, 806–817.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5, 559–581.

Diebold, F.X., 1988. Serial correlation and the combination of forecasts. Journal of Business and Economic Statistics 6, 105–111.

Diebold, F.X., 2001. Elements of Forecasting. 2nd Edition. South-Western Publishing.

Diebold, F.X., Lopez, J.A., 1996. Forecast evaluation and combination. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics 14. Elsevier, Amsterdam, pp. 241–268.

Diebold, F.X., Mariano, R., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–265.

Diebold, F.X., Gunther, T.A., Tay, A.S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39, 863–883.

Elliott, G., Timmermann, A., 2002. Optimal forecast combination weights under regime switching. Manuscript, University of California, San Diego.

Fildes, R., Ord, K., 2001. Forecasting competitions: their role in improving forecasting practice and research. In: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. Blackwell Publishers, Oxford, pp. 322–353.

Granger, C.W.J., 1969. Prediction with a generalized cost of error function. Operational Research 20, 199–207.

Granger, C.W.J., 1989. Combining forecasts—twenty years later. Journal of Forecasting 8, 167–173.

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

Granger, C.W.J., Newbold, P., 1986. Forecasting Economic Time Series, 2nd Edition. Academic Press, New York.

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

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 3, 197–204.

Hendry, D.F., Clements, M.P., 2002. Pooling of forecasts. Econometrics Journal 5, 1–26.

Koenker, R., Basset, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Laster, D., Bennett, P., Geoum, I.S., 1999. Rational bias in macroeconomic forecasts. Quarterly Journal of Economics 114, 293–318.

Marcellino, M., 2002, Forecast pooling for short time series of macroeconomic variables. Mimeo, IGIER, Bocconi University.

Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Annals of Statistics 20, 712–736.

Min, C.-k., Zellner, A., 1993. Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. Journal of Econometrics 56, 89–118.

Newbold, P., Harvey, D.I., 2001. Forecast combination and encompassing. In: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. Blackwells, Oxford.

Newey, W., Powell, J., 1987. Asymmetric least squares estimation and testing. Econometrica 55, 819–847.

Palm, F.C., Zellner, A., 1992. To combine or not to combine? Issues of combining forecasts. Journal of Forecasting 11, 687–701.

Peel, D.A., Nobay, A.R., 1998. Optimal monetary policy in a model of asymmetric central bank preferences. FMG discussion paper 0306.

Perez-Quiros, G., Timmermann, A., 2001. Business cycle asymmetries in stock returns: evidence from higher order moments and conditional densities. Journal of Econometrics 103, 259–306.

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92, 179–191.

Rudin, W., 1964. Principles of Mathematical Analysis, 2nd Edition. McGraw-Hill, New York.

Stock, J.H., Watson, M.W., 1999. Forecasting inflation. Journal of Monetary Economics 44, 293–375.

Tay, A., Wallis, K., 2000. Density forecasting: a survey. Journal of Forecasting 19, 235–254.

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

Weiss, A.A., 1996. Estimating time series models using the relevant cost function. Journal of Applied Econometrics 11, 539–560.

West, K.D., Edison, H.J., Cho, D., 1993. A utility based comparison of some models of exchange rate volatility. Journal of International Economics 35, 23–46.

Whittle, P., 1983. Prediction and regulation by linear least square methods. 2nd Edition, Revised. University of Minnesota Press, Minneapolis, MN.

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

Zhu, L.-X., Neuhaus, G., 2000. Nonparametric Monte Carlo tests for multivariate distributions. Biometrika 87, 919–928.