Evaluating and Comparing Possibly Misspecified Forecasts*

Andrew J. Patton
Duke University

First version: 27 September 2014. This version: 27 March 2015

Abstract

This paper considers the evaluation of forecasts of a given statistical functional, such as a mean, quantile, or distribution. Recent work has emphasized the importance of evaluating such forecasts using a loss function that is consistent for the functional of interest, of which there are an infinite number. If forecasters all use correctly specified models, and if the information sets of the competing forecasters are nested, then the construction, evaluation, and comparison of competing forecasts are invariant to the choice of consistent loss function. However, the presence of misspecified models, parameter estimation error, or nonnested information sets, leads to sensitivity to the choice of (consistent) loss function. Thus, rather than merely specifying the target functional, which narrows the set of relevant loss functions only to the class of loss functions consistent for that functional, this paper proposes that forecast consumers or survey designers should specify the single specific loss function that will be used to evaluate forecasts. An application to survey forecasts of US inflation illustrates the results.

Keywords: Survey forecasts, point forecasting, density forecasting, Bregman distance, proper scoring rules, consistent loss functions.

J.E.L. codes: C53, C52, E37.

AMS 2010 Classifications: 62M20, 62P20.

*I thank Tim Bollerslev, Dean Croushore, Frank Diebold, Tilmann Gneiting, Jia Li and Allan Timmermann for helpful comments. Contact address: Andrew Patton, Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham NC 27708-0097. Email: andrew.patton@duke.edu.

1 Introduction

Recent work in the theory of prediction has (re-)emphasized the importance of the choice of loss function used to evaluate the performance of a forecaster. In particular, there is a growing recognition that the loss function used must "match" the quantity that the forecaster was asked to predict, whether it is the mean, the median, the probability of a particular outcome (e.g., rain, negative economic growth), etc. For example, in the widely-cited "Survey of Professional Forecasters," conducted by the Federal Reserve Bank of Philadelphia, experts are asked to predict a variety of economic variables, with questions such as "What do you expect to be the annual average CPI inflation rate over the next 5 years?" (Section 7 of the survey). In the Thomson Reuters/University of Michigan Survey of Consumers, respondents are asked "By about what percent do you expect prices to go (up/down) on the average, during the next 12 months?" (Question A12b of the survey). The presence of the word "expect" in these questions is an indication (at least to statisticians) that the respondents are being asked for their mathematical expectation of future inflation. The oldest continuous survey of economists' expectations, the Livingston survey, on the other hand, simply asks "What is your forecast of the average annual rate of change in the CPI?," leaving the specific type of forecast unstated.

In point forecasting, a loss function is said to be "consistent" for a given statistical functional (e.g., the mean, median, etc.), if the expected loss is minimized when the given functional is used as the forecast, see Gneiting (2011a) and discussion therein. For example, a loss function is consistent for the mean if no other quantity (median, mode, etc.) leads to a lower expected loss than the mean. The class of loss functions that is consistent for the mean is known as the Bregman class of loss functions, see Savage (1971), Banerjee, *et al.* (2005) and Bregman (1967), and includes the squared-error loss function as a special case. In density or distribution forecasting the analogous idea is that of a "proper" scoring rule, see Gneiting and Raftery (2007): a scoring rule is proper if the expected loss under distribution P is minimized when using P as the distribution forecast. Evaluating forecasts of a given functional using consistent loss functions or proper scoring rules is a minimal requirement for sensible rankings of the competing forecasts.

Gneiting (2011a, p757) summarizes the implications of the above work as follows: *"If point forecasts are to be issued and evaluated, it is essential that either the scoring function be specified ex ante, or an elicitable target functional be named, such as the mean or a quantile of the predictive distribution, and scoring functions be used that are consistent for the target functional."* This paper contributes to this literature by refining this recommendation to reflect some real-word deviations from the ideal predictive environment, and suggests that only the first part of Gneiting's recommendation should stand; specifying the target functional is generally *not* sufficient to elicit a forecaster's best (according to a given, consistent, loss function) prediction. Instead, forecasters should be told the single, specific loss function that will be used to evaluate their forecasts.

Firstly, I show that when two competing forecasts are generated using correctly specified models and the information sets of one of the forecasters nests the other, then the ranking of these forecasts based on a single consistent loss function is sufficient for their ranking using *any* consistent loss function (subject of course to integrability conditions). This is established for the problem of mean forecasting, quantile forecasting (nesting the median as a special case), and distribution forecasting.

Secondly, and with more practical importance, I show that when competing forecasts are based on nonnested information sets, misspecified models, or models with estimated parameters, the ranking of the forecasts is generally sensitive to the choice of consistent loss function. This result has important implications for survey forecast design and for forecast evaluation more generally.

This result also has implications for the use of multiple loss functions to evaluate forecasts. If the loss functions used are not all consistent for the same statistical functional, then existing arguments from Engelberg, *et al.* (2007), Gneiting (2011a) and Patton (2011) apply, and it is not surprising that the rankings may differ across loss functions. If the loss functions are all consistent for the same functional, then in the absence of misspecified models and nonnested information sets, using multiple measures of accuracy adds *no* information beyond using just one measure. (Note, however, that these loss functions may have different sampling properties, and so careful choice of the loss function to use may lead to improved efficiency.) In the presence of these real-world forecasting complications, using multiple measures of forecast accuracy can lead to clouded results: a forecaster could be best under one loss function and worst under another; averaging the performance across

multiple measures could mask true out-performance under one specific loss function.

This paper also shows that if the target variable has a parametric conditional mean function, and the forecaster's model for this is correctly specified, then minimizing the expected loss under *any* Bregman loss function yields a consistent estimator of the model's parameters, and of course such a forecast will pass any forecast optimality test. However, under misspecification the choice of (Bregman) loss function used in estimation will generally lead to estimators that converge to different probability limits, and lower average loss for the forecast consumer can be attained by optimizing the misspecified model using the consumer's loss function.

The focus in this paper is on applications where the target functional (mean, quantile, etc.) is known, and the task is to find the "best" forecast of this functional. In contrast, in some economic applications, the target functional is not known or stated explicitly, and instead the decision in which the forecast will be used is specified, which "traces out" a particular economic loss function (and in turn implies a particular statistical functional as the target). See Leitch and Tanner (1991), West, *et al.* (1993) and Skouras (2007) for examples. While distinct, the recommendation from this paper is related: in the presence of potential model misspecification or nonnested information sets, forecast producers should be told the specific loss function that will be used to evaluate their predictions. When the target functional is known, the given loss function should of course be consistent for that functional, but in the presence of model misspecification or nonnested information, merely specifying the target functional is not sufficient.

I illustrate these ideas in this paper with a study of the inflation forecasting performance of respondents to the Survey of Professional Forecasters (SPF) and the Michigan Survey of Consumers. Under squared-error loss, I find that the SPF consensus forecast and the Michigan consensus forecast are very similar in accuracy, with slight preference to the SPF, but when a Bregman loss function is used that penalizes over- or under-predictions more heavily, the ranking of these forecasts switches. I also consider comparisons of individual respondents to the SPF, and find cases where the ranking of two forecasters is very sensitive to the particular choice of Bregman loss function, and cases where the ranking is robust across a range of Bregman loss functions.

This paper is related to several recent papers on related topics. Elliott, *et al.* (2014) study the

problem of forecasting binary variables with binary forecasts, and the evaluation and estimation of models based on consistent loss functions. They obtain several useful, closed-form, results for this case. Merkle and Steyvers (2013) also consider forecasting binary variables, and provide an example where the ranking of forecasts is sensitive to the choice of consistent loss function. Holzmann and Eulert (2014) show in a very general framework that forecasts based on larger information sets lead to lower expected loss, and apply their results to Value-at-Risk forecasting. This paper builds on these works, and the important work of Gneiting (2011a), to show the strong conditions under which the evaluation and comparison of a forecast of a given statistical functional is insensitive to the choice of loss function, even when that choice is constrained to the set of loss functions that are consistent for the given functional. Examples and illustrations designed to resemble those faced in economic forecasting applications highlight the relevance of the problem, and provide support for the key recommendation of this paper: when conducting surveys or forecast competitions, forecast producers should be told not only the statistical functional of interest, but rather the specific loss function that will be used to evaluate their predictions.

The remainder of the paper is structured as follows. Section 2 presents positive and negative results on forecast comparison in the absence and presence of real-world complications like nonnested information sets and misspecified models, covering mean, quantile and distribution forecasts. Section 3 considers optimal approximations and forecast adjustments in the presence of model misspecification. Section 4 considers realistic simulation designs that illustrate the main ideas of the paper, and Section 5 presents an analysis of US inflation forecasts. The appendix presents proofs, and a web appendix contains additional details.

2 Comparing forecasts using consistent loss functions

2.1 Mean forecasts and Bregman loss functions

The most well-known loss function is the quadratic or squared-error loss function:

$$L\left(y, \hat{y}\right) = \left(y - \hat{y}\right)^2 \tag{1}$$

Under quadratic loss, the optimal forecast of a variable is well-known to be the (conditional) mean:

$$\hat{Y}_t^* \equiv \arg\min_{\hat{y}\in\mathcal{Y}} E\left[L\left(Y_t,\hat{y}\right)|\mathcal{F}_t\right] \quad (2)$$

$$= E\left[Y_t|\mathcal{F}_t\right], \text{ if } L\left(y,\hat{y}\right)=\left(y-\hat{y}\right)^2 \quad (3)$$

where $\mathcal{F}_t$ is the forecaster's information set. More generally, the conditional mean is the optimal forecast under any loss function belonging to a general class of loss functions known as Bregman loss functions (see Banerjee, *et al.*, 2005 and Gneiting, 2011a). The class of Bregman loss functions is then said to be "consistent" for the (conditional) mean functional. Elements of the Bregman class of loss functions, denoted $\mathcal{L}_{Bregman}$, take the form:

$$L\left(y,\hat{y}\right)=\phi\left(y\right)-\phi\left(\hat{y}\right)-\phi'\left(\hat{y}\right)\left(y-\hat{y}\right) \quad (4)$$

where $\phi:\mathcal{Y}\rightarrow\mathbb{R}$ is any strictly convex function, and $\mathcal{Y}$ is the support of Y_t. Moreover, this class of loss functions is also *necessary* for conditional mean forecasts, in the sense that if the optimal forecast is known to be the conditional mean, then it must be that the forecast was generated by minimizing the expected loss of some Bregman loss function. Two prominent examples of Bregman loss functions are quadratic loss (equation (1)) and QLIKE loss, which is applicable for strictly positive random variables:

$$L\left(y,\hat{y}\right)=\frac{y}{\hat{y}}-\log\frac{y}{\hat{y}}-1 \quad (5)$$

The quadratic and QLIKE loss functions are particularly special, in that they are the only two Bregman loss functions that only depend on the difference (Savage, 1971) or the ratio (Patton, 2011) of the target variable and the forecast.

To illustrate the variety of shapes that Bregman loss functions can take, two parametric families of Bregman loss for variables with support on the real line are presented below. The first was proposed in Gneiting (2011a), and is a family of homogeneous loss functions, where the "shape" parameter determines the degree of homogeneity. It is generated by using $\phi\left(x;k\right)=\left|x\right|^k$ for $k>1$:

$$L\left(y,\hat{y};k\right)=\left|y\right|^k-\left|\hat{y}\right|^k-k\,\mathrm{sgn}\left(\hat{y}\right)\left|\hat{y}\right|^{k-1}\left(y-\hat{y}\right), \quad k>1 \quad (6)$$

This family nests the squared-error loss function at $k=2$. (The non-differentiability of ϕ can be ignored if Y_t is continuously distributed, and the absolute value components can be dropped

altogether if the target variable is strictly positive, see Patton, 2011). A second, non-homogeneous, family of Bregman loss can be obtained using $\phi(x;a) = 2a^{-2}\exp\{ax\}$ for $a \neq 0$:

$$L(y,\hat{y};a) = \frac{2}{a^2}\left(\exp\{ay\} - \exp\{a\hat{y}\}\right) - \frac{2}{a}\exp\{a\hat{y}\}(y-\hat{y}), \quad a \neq 0 \tag{7}$$

This family nests the squared-error loss function as $a \to 0$, and is convenient for obtaining closed-form results when the target variable is Normally distributed. This loss function has some similarities to the "Linex" loss function, see Varian (1974) and Zellner (1986), in that it involves both linear and exponential terms, however a key difference is that the above family implies that the optimal forecast is the conditional mean, and does not involve higher-order moments.

Figure 1 illustrates the variety of shapes that Bregman loss functions can take and reveals that although all of these loss functions yield the mean as the optimum forecast, their shapes can vary widely: these loss functions can be asymmetric, with either under-predictions or over-predictions being more heavily penalized, and they can be strictly convex or have concave segments. Thus restricting attention to loss functions that generate the mean as the optimum forecast does *not* require imposing symmetry or other such assumptions on the loss function. Similarly, in the literature on economic forecasting under asymmetric loss (see Granger, 1969, Christoffersen and Diebold, 1997, and Patton and Timmermann, 2007, for example), it generally thought that asymmetric loss functions necessarily lead to optimal forecasts that differ from the conditional mean (they contain an "optimal bias" term). Figure 1 reveals that asymmetric loss functions can indeed still imply the conditional mean as the optimal forecast. (In fact, Savage (1971) shows that of the infinite number of Bregman loss functions, only one is symmetric: the quadratic loss function.)

[INSERT FIGURE 1 ABOUT HERE]

2.2 Correctly specified models and nested information sets

While forecasts are of course based on conditioning information, I will consider ranking forecasts by their unconditional average loss, a quantity that is estimable, under standard regularity conditions, given a sample of data. For notational simplicity, I assume strict stationarity of the data, but certain forms of heterogeneity can be accommodated by using results for heterogeneous processes,

see White (2001) for example. I use t to denote an observation, for example a time period, however the results in this paper are applicable wherever one has repeated observations, for example election forecasting across states, sales forecasting across individual stores, etc.

Firstly, consider a case where forecasters A and B are ranked by mean squared error (MSE)

$$MSE_i \equiv E\left[\left(Y_t - \hat{Y}_t^i\right)^2\right], \ \ i \in \{A, B\} \tag{8}$$

and we then seek to determine whether

$$MSE_A \lesseqgtr MSE_B \Rightarrow E\left[L\left(Y_t, \hat{Y}_t^A\right)\right] \lesseqgtr E\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \ \ \forall L \in \mathcal{L}_{Bregman} \tag{9}$$

subject to these expectations existing. The following proposition provides conditions under which the above implication holds.

Proposition 1 *Assume that (i) The information sets of the two forecasters are nested, so* $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t$ *or* $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B \ \forall t$*, and (ii) Forecasts A and B are optimal under some Bregman loss function. Then the ranking of these forecasts by MSE is sufficient for their ranking by any Bregman loss function.*

Thus under the strong assumptions of comparing only forecasters with nested information sets, and who use only correctly specified models with no estimation error, the ranking obtained by MSE is sufficient the ranking by *any* Bregman loss function. This of course also implies that ranking forecasts by a variety of different Bregman loss functions adds no information beyond ranking by any single Bregman loss function.

Holzmann and Eulert (2014) show in a general framework that forecasts based on larger information sets lead to lower expected loss, including the mean case as well as other point forecasting applications. Their method of proof is quite different, and their interpretation differs from here.

2.3 Misspecified models or non-nested information sets

Next we consider deviations from the two assumptions underlying the above result. In part (a) of the following proposition we consider the case that both forecasters are able to construct fully accurate

estimates of the conditional mean, but their information sets are non-nested. This is a plausible scenario in practice, if we consider a variety of experts trying to predict a complicated variable, all working to gather new and useful information to construct their forecasts. It is particularly relevant if we compare two "types" of forecasters, such as professional forecasters and consumers, as in the empirical application in Section 5, whose access to different types of data differs. In part (b) we consider the case that the forecasters may be using misspecified models, which we take here to also include correctly-specified models that are subject to estimation error. Like the case of non-nested information sets, and perhaps even more so, this is a very plausible scenario in practice.

Proposition 2 *Assume that (a) the information sets of the two forecasters are non-nested, so* $\mathcal{F}_t^B \nsubseteq \mathcal{F}_t^A$ *and* $\mathcal{F}_t^A \nsubseteq \mathcal{F}_t^B$ *for some* t*, but Forecasts* A *and* B *are optimal under some Bregman loss function, or (b) at least one of the forecasts is based on a misspecified model. Then the ranking of these forecasts is, in general, sensitive to the choice of Bregman loss function.*

The web appendix contains a simple example supporting the above proposition, based on two-point and three-point random variables, and Merkle and Steyvers (2013) present an example for forecasts of binary variables. An example based on a design more closely related to economic applications is presented in Section 4 below. In all cases, the key insight is that the relative weight given to large and small, and positive and negative, forecast errors by different Bregman loss functions induces different rankings of the competing forecasts, when they are based on nonnested information sets or on misspecified models.

It should be noted that it may be possible to partially relax assumptions (i) and (ii) in Proposition 1, or to place other restrictions on the problem, and retain the robustness of the ranking of forecasts to the choice of Bregman loss function. For example, if the form of the model misspecification was known, or if the target variable has a particularly simple structure (e.g., a binary random variable). I do not pursue such special cases here.

The following corollary generalizes Propositions 1 and 2 to evaluating many forecasters.

Corollary 1 *Consider evaluating* $N \geq 2$ *forecasters.*

(a) Assume (i) there exists an ordering of the forecasters such that $\mathcal{F}_t^{(1)} \subseteq \mathcal{F}_t^{(2)} \subseteq \cdots \subseteq \mathcal{F}_t^{(N)} \; \forall \; t$*, and (ii) all forecasts are optimal under some Bregman loss function. Then the ranking of these forecasts by MSE is sufficient for their ranking by any Bregman loss function.*

(b) Assume (i) there exists a forecaster i^* *such that* $\bigcup_{i \neq i^*} \mathcal{F}_t^{(i)} \subseteq \mathcal{F}_t^{(i^*)} \; \forall \; t$*, and (ii) forecast* i^* *is optimal under some Bregman loss function. Then forecaster* i^* *will have the lowest average loss using any Bregman loss function, including MSE. The ranking of the other (non* i^**) forecasters will, in general, be sensitive to the choice of loss function.*

2.4 Quantile forecasts

This section presents results for quantile forecasts that correspond to those above for mean forecasts. The corresponding result for the necessity and sufficiency of Bregman loss for mean forecasts is presented in Saerens (2000), see also Komunjer (2005), Gneiting (2011b) and Thomson (1979): the loss function that is necessary and sufficient for quantile forecasts is called a "generalized piecewise linear" (GPL) loss function, denoted $\mathcal{L}_{GPL}^{\alpha}$:

$$L\left(y, \hat{y}; \alpha\right) = \left(\mathbf{1}\left\{y \leq \hat{y}\right\} - \alpha\right)\left(g\left(\hat{y}\right) - g\left(y\right)\right) \tag{10}$$

where g is a nondecreasing function, and $\alpha \in (0,1)$ indicates the quantile of interest. A prominent example of a GPL loss function is the "Lin-Lin" (or "tick") loss function, which is obtained when g is the identity function:

$$L\left(y, \hat{y}; \alpha\right) = \left(\mathbf{1}\left\{y \leq \hat{y}\right\} - \alpha\right)\left(\hat{y} - y\right) \tag{11}$$

and which nests absolute error (up to scale) when $\alpha = 1/2$. However, there are clearly an infinite number of loss functions that are consistent for the α quantile. The following is a homogeneous parametric GPL family of loss functions (for variables with support on the real line) related to that proposed by Gneiting (2011b):

$$L\left(y, \hat{y}; \alpha, b\right) = \left(\mathbf{1}\left\{y \leq \hat{y}\right\} - \alpha\right)\left(\operatorname{sgn}\left(\hat{y}\right)\left|\hat{y}\right|^b - \operatorname{sgn}\left(y\right)\left|y\right|^b\right)/b, \quad b > 0 \tag{12}$$

Figure 2 presents some elements of this family of loss functions for $\alpha = 0.5$ and $\alpha = 0.25$, and reveals that although the optimal forecast is always the same under all of these loss functions (with the same α), their individual shapes can vary substantially.

when the loss function belongs to the GPL family, the optimal forecast satisfies

$$\alpha = E\left[\mathbf{1}\left\{Y_t \leq \hat{Y}_t^*\right\}|\mathcal{F}_t\right] \equiv F_t\left(\hat{Y}_t^*\right) \tag{13}$$

where $Y_t|\mathcal{F}_t \sim F_t$, and if the conditional distribution function is strictly increasing, then $\hat{Y}_t^* = F_t^{-1}(\alpha|\mathcal{F}_t)$. Now we seek to determine whether the ranking of two forecasts by Lin-Lin loss is sufficient for their ranking by any GPL loss function (with the same α parameter). That is, whether

$$LinLin_A^\alpha \lesseqgtr LinLin_B^\alpha \Rightarrow E\left[L\left(Y_t, \hat{Y}_t^A\right)\right] \lesseqgtr E\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \ \forall L \in \mathcal{L}_{GPL}^\alpha \tag{14}$$

subject to these expectations existing. Under the analogous conditions to those for the conditional mean, a sufficiency result obtains.

Proposition 3 *Assume that (i) The information sets of the two forecasters are nested, so* $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t$ *or* $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B \ \forall t$*, and (ii) Forecasts A and B are optimal under some* $\mathcal{L}_{GPL}^\alpha$ *loss function. Then the ranking of these forecasts by expected Lin-Lin loss is sufficient for their ranking by any* $\mathcal{L}_{GPL}^\alpha$ *loss function.*

Next again consider deviations from the two assumptions underlying the above result, namely allowing the information sets of the two forecasters to be nonnested, or allowing for model misspecification. As in the conditional mean case, either of these complications is enough to induce sensitivity to the choice of loss function of the ranking of two quantile forecasts. A simple example supporting this proposition is presented in the web appendix, and a more realistic example is presented in Section 4 below.

Proposition 4 *Assume that (a) The information sets of the two forecasters are non-nested, so* $\mathcal{F}_t^B \nsubseteq \mathcal{F}_t^A$ *and* $\mathcal{F}_t^A \nsubseteq \mathcal{F}_t^B$ *for some* t*, but Forecasts A and B are optimal under some GPL loss function, or (b) one or both of the* α*-quantile forecasts are based on misspecified models. Then the ranking of these forecasts is, in general, sensitive to the choice of GPL loss function.*

2.5 Density forecasts

We now consider results corresponding to the mean and quantile cases above for density or distribution forecasts. In this case the central idea is the use of a proper scoring rule. A "scoring rule,"

see Gneiting and Ranjan (2011) for example, is a loss function mapping the density or distribution forecast and the realization to a measure of gain/loss. (In density forecasting this is often taken as a gain, but for comparability with the above two sections I will treat it here as a loss, so that lower values are preferred.) A "proper" scoring rule is any scoring rule such that it is minimized in expectation when the distribution forecast is equal to the true distribution. That is, L is proper if

$$E_F\left[L\left(F,Y\right)\right] \equiv \int L\left(F,y\right) dF\left(y\right) \leq E_F\left[L\left(\tilde{F},Y\right)\right] \tag{15}$$

for all distribution functions $F,\ \tilde{F} \in \mathcal{P}$, where $\mathcal{P}$ is the class of probability measures being considered. (I will use distributions rather than densities for the main results here, so that they are applicable more generally.) Gneiting and Raftery (2007) show that if L is a proper scoring rule then it must be of the form:

$$L\left(F,y\right) = \Psi\left(F\right) + \Psi^*\left(F,y\right) - \int \Psi^*\left(F,y\right) dF\left(y\right) \tag{16}$$

where Ψ is a convex, real-valued function, and Ψ^* is a subtangent of Ψ at the point $F \in \mathcal{P}$. I denote the set of proper scoring rules satisfying equation (16) as $\mathcal{L}_{\text{Proper}}$. As an example of a proper scoring rule, consider the "weighted continuous ranked probability score" from Gneiting and Ranjan (2011):

$$wCRPS\left(F,y;\omega\right) = \int_{-\infty}^{\infty} \omega\left(z\right)\left(F\left(z\right) - \mathbf{1}\left\{y \leq z\right\}\right)^2 dz \tag{17}$$

where ω is a nonnegative weight function on $\mathbb{R}$. If ω is constant then the above reduces to the (unweighted) CRPS loss function.

Now we seek to determine whether the ranking of two forecasts by two distribution forecasts by any single proper scoring rule is consistent for their ranking by any proper scoring rule.

$$E\left[L_i\left(F_t^A, Y_t\right)\right] \lesseqgtr E\left[L_i\left(F_t^B, Y_t\right)\right] \Rightarrow E\left[L_j\left(F_t^A, Y_t\right)\right] \lesseqgtr E\left[L_j\left(F_t^B, Y_t\right)\right] \ \forall L_j \in \mathcal{L}_{\text{Proper}} \tag{18}$$

Under the analogous conditions to those for the conditional mean and conditional quantile, a sufficiency result obtains.

Proposition 5 *Assume that (i) the information sets of the two forecasters are nested, so* $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A\ \forall t$ *or* $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B\ \forall t$*, and (ii)* F_t^A *and* F_t^B *are the conditional distributions of* $Y_t|\mathcal{F}_t^A$ *and* $Y_t|\mathcal{F}_t^B$

respectively. Then the ranking of these forecasts by any given proper scoring rule is sufficient for their ranking by any other proper scoring rule.

Now consider again deviations from the two assumptions underlying the above result, where we allow the information sets of the two forecasters to be nonnested, or for model misspecification. As in the conditional mean and conditional quantile cases, either of these situations is enough to induce sensitivity to the choice of loss function of the ranking of two distribution forecasts. A simple example supporting this proposition is presented in the web appendix, and a more realistic example is given in Section 4 below.

Proposition 6 *Assume that (a) the information sets of the two forecasters are non-nested, so* $\mathcal{F}_t^B \nsubseteq \mathcal{F}_t^A$ *and* $\mathcal{F}_t^A \nsubseteq \mathcal{F}_t^B$ *for some* t*, but* F_t^A *and* F_t^B *are the conditional distributions of* $Y_t|\mathcal{F}_t^A$ *and* $Y_t|\mathcal{F}_t^B$ *respectively, or (b) one or both of the distribution forecasts are based on misspecified models. Then the ranking of these forecasts is, in general, sensitive to the choice of proper scoring rule.*

2.6 Mean forecasts of symmetric random variables

We next consider a case where some additional information about the target variable is assumed to be known, and as a leading example in economic forecasting, we consider the case that the target variable is assumed to be symmetrically distributed. In the following proposition we show if this assumption is made, then the class of loss functions that leads to the forecasters revealing their conditional mean is larger than in Section 2.1. In the first part of the following proposition we establish the relevant class of loss functions in this case. The second and third parts present results on ranking forecasters when the assumptions of correctly-specified models and nested information sets hold, or fail to hold.

Proposition 7 *Assume that (i) forecaster* j *optimizes his/her forecast with respect to a symmetric continuous distribution* $F_t^{(j)}$*, for all* j*. Then, (a) any convex combination of a Bregman and a* $GPL^{1/2}$ *loss function,* $\mathcal{L}_{Breg\times GPL} \equiv \lambda \mathcal{L}_{Bregman} + (1-\lambda)\,\mathcal{L}_{GPL}^{1/2}$*,* $\lambda \in [0,1]$*, yields the mean of* $F_t^{(j)}$ *as the optimal forecast.*

(b) Assume that (ii) The information sets of the two forecasters are nested, so $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t$ *or* $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B \ \forall t$*, and (iii) Forecasts A and B are optimal under some loss function in* $\mathcal{L}_{Breg \times GPL}$*. Then the ranking of these forecasts by MSE or MAE is sufficient for their ranking by any loss function in* $\mathcal{L}_{Breg \times GPL}$*.*

(c) Assume that (ii') the information sets of the two forecasters are non-nested, so $\mathcal{F}_t^B \nsubseteq \mathcal{F}_t^A$ *and* $\mathcal{F}_t^A \nsubseteq \mathcal{F}_t^B$ *for some t, but Forecasts A and B are optimal under some loss function in* $\mathcal{L}_{Breg \times GPL}$*, or (iii') at least one of the forecasts is based on a misspecified model. Then the ranking of these forecasts is, in general, sensitive to the choice of loss function in* $\mathcal{L}_{Breg \times GPL}$*.*

Thus if forecasters are known to be using a symmetric model for the target variable, regardless of whether that assumption is correct, then the class of loss functions that is consistent for the mean is now *even larger* than the Bregman class: it is the convex combination of the Bregman and the $\text{GPL}^{1/2}$ class of loss functions. In this case it is even more important to declare which specific loss function will be used to rank the forecasts.

3 Constructing and improving forecasts under misspecification

This section considers two related problems: constructing forecasts based on possibly misspecified models, and optimally adjusting forecasts that may have been generated from a misspecified model. Below I focus on mean forecasting and Bregman loss functions, but the key insights carry over to other forecasts using ideas related to those in the previous section.

3.1 Optimal approximations

Consider the problem of calibrating a parametric forecasting model to generate the best (in some sense) prediction. First, we show that if the model is correctly specified, then minimizing the expected loss under any Bregman loss function will yield a consistent estimator of the model's parameters. However if the model is misspecified, then different Bregman estimators will yield different estimators, in the sense that they converge to different probability limits. Elliott, *et al.* (2014) provide several useful related results on this problem when both the target variable and the

forecast are binary. They show that even in their relatively tractable case, the presence of model misspecification generally leads to sensitivity of estimated parameters to the choice of (consistent) loss function.

Assumption (i) below simply states that the conditional mean has a parametric form, with true parameter θ_0. Assumption (ii) is required for identification, imposing that the conditional mean is sensitive to changes in the parameter θ.

Proposition 8 *Assume that (i)* $E[Y_t|\mathcal{F}_t] = m(V_t;\theta_0)$, *for some* $\theta_0 \in \Theta \subseteq \mathbb{R}^k$, $k < \infty$, *and (ii)* $\partial m(V_t;\theta)/\partial\theta \neq 0$ *a.s.* $\forall\theta \in \Theta$. *Define*

$$\hat{\theta}^*_\phi \equiv \arg\min_{\theta\in\Theta} \; E[L(Y_t, m(V_t;\theta);\phi)] \tag{19}$$

where L *is a Bregman loss function characterized by the convex function* ϕ. *Then* $\hat{\theta}^*_\phi = \theta_0 \; \forall \; \phi$.

Next we consider an example where the model is misspecified, and show that different Bregman estimators can yield very different approximations. Consider the following simple DGP:

$$\begin{aligned} Y_t &= X_t^2 + Z_t + \varepsilon_t \\ [X_t, Z_t, \varepsilon_t] &\sim iid\ N\left([\mu, 0, 0], diag\left\{\left[\sigma^2, \omega^2, 1\right]\right\}\right) \end{aligned} \tag{20}$$

But the forecaster uses only a linear model:

$$Y_t = \alpha + \beta X_t + \lambda Z_t + e_t \tag{21}$$

That is, the forecaster has two predictor variables, but mistakenly treats one of them linearly. For the illustration in this subsection, we will set $\omega^2 = \lambda = 0$ and so Z_t is removed from the design. Consider a forecaster using the "Exponential Bregman" loss function, defined in equation (7), with parameter p. Using results for functions of Normal random variables (see the appendix for details) we can analytically derive the optimal linear model parameters $[\alpha, \beta]$ as a function of p, subject to the condition that $p \neq \left(2\sigma^2\right)^{-1}$:

$$\begin{aligned} \hat{\alpha}^*_p &= \sigma^2 - \frac{\mu^2}{\left(1 - 2p\sigma^2\right)^2} \\ \hat{\beta}^*_p &= \frac{2\mu}{1 - 2p\sigma^2} \end{aligned} \tag{22}$$

This simple example reveals some important features of the problem of loss function-based parameter estimation in the presence of model misspecification. Firstly, the loss function shape parameter does not always affect the optimal model parameters. In this example, if $X \sim N\left(0, \sigma^2\right)$, then $\left(\hat{\alpha}_p^*, \hat{\beta}_p^*\right) = \left(\sigma^2, 0\right)$ for all values of the loss function parameter. Second, identification issues can arise even when the model appears to be *prima facie* well identified. In this example, the estimation problem is not identified at $p = \left(2\sigma^2\right)^{-1}$. Issues of identification when estimating under the "relevant" loss function have been previously documented, see Weiss (1996) and Skouras (2007).

Finally, when $\mu \neq 0$, the optimal model parameters will vary with the loss function parameter, and thus the choice of loss function to use in estimation will affect the approximation yielded by the misspecified model. Figure 3 illustrates this point, presenting the optimal linear approximations for three choices of exponential Bregman parameter, when $\mu = \sigma^2 = 1$. The optimal parameters are presented in Table 1. The approximation yielded by OLS regression is obtained when $p = 0$, and there we see the intercept is zero and the slope coefficient is two. If we consider a loss function that places greater weight on errors that occur for low values of the forecast ($p = -0.5$) the line flattens and the upper panel of Figure 3 shows that this yields a better fit for the left side of the distribution of the predictor variable. The opposite occurs if we consider a loss function that places greater weight on errors that occur for high values of the forecast ($p = 0.25$). The lower panel of Figure 3 presents a simple nonparametric estimate of the distance between the realization and the forecast as a function of the predictor variable. For OLS ($p = 0$), the distance is lowest for X approximately in the interval $(0, 2)$, which covers most of the data, since $X \sim N(1, 1)$ in this example. For $p = -0.5$ the distance is lowest for X approximately in the interval $(-1, 1.5)$, while for $p = 0.25$ the distance is lowest for X approximately in the interval $(1, 3)$.

[INSERT FIGURE 3 AND TABLE 1 ABOUT HERE]

The above results motivate declaring the specific loss function that will be used to evaluate forecasts, so that survey respondents can optimize their (potentially misspecified) models taking the relevant loss function into account. However, it is important to note that it is not always the case that optimizing the model using the relevant loss function is optimal in finite samples:

there is a trade-off between bias in the estimated parameters (computed relative to the probability limits of the parameter estimates obtained using the relevant loss function) and variance (parameter estimation error). It is possible that an efficient (low variance) but biased estimation method could out-perform a less efficient but unbiased estimation method in finite samples. See Elliott, *et al.* (2014) for discussion and examples of this for the binary prediction problem. One interpretation of the results in this section is that if all estimators converge to the same quantity then there is no bias-variance trade-off, and one need only look for the most efficient estimator. A trade-off potentially exists when the models are misspecified and the estimators converge to different limits.

3.2 Improving forecasts from misspecified models

Consider a scenario where a forecast producer generates a forecast under Bregman loss L_p, but the forecast consumer has Bregman loss L_c. Is it possible that the consumer can improve (in terms of expected loss under L_c) the forecast received from the forecast producer? In line with the forecast environment faced in practice, we will assume that the consumer cannot see the producer's predictive information, only the reported forecast and the realization of the target variable. Consider a "forecast adjustment function," $g : \mathcal{Y} \to \mathcal{Y}$, which takes in the forecast and modifies it in an optimal way for the consumer's loss function. A simple example of this is a linear function:

$$g(\hat{y}) = \gamma + \delta\hat{y} \tag{23}$$

Of course linearity is a particular case, and more general functions could be entertained. The only constraint on g is that it nests the identity function, so that if the producer's and consumer's loss functions happen to coincide, and the producer's forecast is optimal, then no adjustment is made. If the adjustment function is parametric, then the optimal adjustment parameters can be obtained by minimizing the forecast consumer's expected loss, as in equation (24) below.

If the producer's forecast is fully optimal (i.e., it comes from a correctly specified model, and also contains no estimation error) then the proposition below shows that no complications arise: the producer's forecast is also optimal under any $L_c \in \mathcal{L}_{Bregman}$.

Proposition 9 *Assume that (i)* $\hat{Y}^*_{p,t} = E[Y_t|\mathcal{F}_t]$, *(ii)* $g(\hat{y};\theta_0) = \hat{y}$ *for some* $\theta_0 \in \Theta$, *and (iii)* $\partial g(\hat{y};\theta)/\partial\theta \neq 0$ *a.s.* $\forall\theta \in \Theta \subseteq \mathbb{R}^k$, $k < \infty$. *Define*

$$\theta^*_{c|p} = \arg\min_{\theta\in\Theta} \; E\left[L_c\left(Y_t, g\left(\hat{Y}^*_{p,t};\theta\right)\right)\right] \tag{24}$$

where $L_c \in \mathcal{L}_{Bregman}$. *Then* $\theta^*_{c|p} = \theta_0 \;\forall\; L_c$.

Combining this and the previous proposition, we observe that if the forecast producer has a correctly specified model, then he can use any Bregman loss function to obtain the model parameters (Proposition 8), and the resulting forecast will be optimal for the forecast consumer regardless of the form of her Bregman loss function (Proposition 9). However, if the forecast producer's model is misspecified, then the previous section showed that the optimal model parameters will be sensitive to the choice of Bregman loss function, and below we show that the resulting forecast may be improved (in terms of the forecast consumer's expected loss) by taking into account the specific form of the consumer's loss function.

Consider the DGP and model from equations (20)–(21) above, and set $\omega^2 = 1$. Then if the producer's loss function is exponential-Bregman with parameter p, calculations analogous to the previous section reveal that α^*_p and β^*_p are unchanged from equation (22), while $\lambda^*_p = 1$ for all p. Then define the producer's forecast as:

$$\hat{Y}^*_{p,t} \equiv \alpha^*_p + \beta^*_p X + \lambda^*_p Z \tag{25}$$

Now consider a simple linear forecast adjustment function as in equation (23), optimized using the consumer's loss function, assumed to be exponential-Bregman with parameter c:

$$\left[\gamma^*_{c|p}, \delta^*_{c|p}\right] = \arg\min_{[\gamma,\delta]} \; E\left[L\left(Y, \gamma + \delta\hat{Y}^*_p; c\right)\right] \tag{26}$$

$$\hat{Y}^*_{c|p} \equiv \gamma^*_{c|p} + \delta^*_{c|p}\hat{Y}^*_p \tag{27}$$

Using derivations similar to the previous section, we can obtain the optimal adjustment parameters as a function of the consumer's loss function parameter. The general expressions are lengthy, but if we specialize to case that $\mu = \sigma^2 = \omega^2 = 1$, and assume that the producer's loss function parameter

p is zero (i.e., the producer uses quadratic loss), then we obtain:

$$\begin{aligned} \gamma^*_{c|p} &= \frac{4c\,(7c-20)}{(5-8c)^2} \\ \delta^*_{c|p} &= \frac{5}{5-8c} \end{aligned} \tag{28}$$

These parameters are plotted in Figure 4 for a range of values of c, and we observe that when $c < p = 0$, it is optimal to dampen the weight on the forecast ($\delta^*_{c|p} < 1$) and increase the intercept ($\gamma^*_{c|p} > 0$), and vice versa for $c > p = 0$. Of course, at $c = p = 0$, we observe $\left[\gamma^*_{c|p}, \delta^*_{c|p}\right] = [0, 1]$.

While the optimally-adjusted forecast ($\hat{Y}^*_{c|p}$) is (generally) an improvement on the producer's forecast ($\hat{Y}^*_p$), it is not as good as if the producer had calibrated his model using the consumer's loss function rather than his own, to obtain $\hat{Y}^*_c$. Table 3 below shows this explicitly, presenting expected losses from the producer's forecast, the optimally adjusted forecast, and the optimal consumer's forecast, when the producer's loss function has parameter $p = 0$ (i.e., quadratic loss) and the consumer's loss function has parameter $c = 0.2$.

[INSERT FIGURE 4 AND TABLE 3 ABOUT HERE]

The problem of "forecast adjustment" has close links to standard tests of forecast optimality: if the optimal forecast adjustment is not different from the identity function, then the forecast is optimal in the directions captured by the adjustment function. For example, the famous Mincer-Zarnowitz (MZ) rationality test involves regressing the target variable on a constant and the forecast

$$Y_t = \gamma + \delta \hat{Y}_t + u_t \tag{29}$$

and then testing

$$\begin{aligned} H_0 &: [\gamma, \delta] = [0, 1] \\ \text{vs.} \quad H_a &: [\gamma, \delta] \neq [0, 1] \end{aligned} \tag{30}$$

which is analogous to the linear adjustment function considered above. Estimating $[\gamma, \delta]$ by OLS minimizes the expected quadratic loss, which is a special case of estimating these parameters by minimizing expected Bregman loss, as done in equation (24) above. Estimating the MZ parameters by minimizing average Bregman loss generates a "Bregman-MZ" test of rationality.

Consider again the DGP and model in equations (20)–(21) above, and simplify by setting $\omega^2 = \lambda = 0$. If the forecast is generated using an exponential-Bregman loss function with parameter p, and is evaluated using an exponential-Bregman loss function with parameter c, then the optimal adjustment parameters are:

$$\begin{aligned} \gamma^*_{c|p} &= \frac{2\sigma^2\left(\mu^2 + \sigma^2\left(1-2p\sigma^2\right)\left(1-2c\sigma^2\right)\right)}{\left(1-2p\sigma^2\right)\left(1-2c\sigma^2\right)^2}(p-c) \\ \delta^*_{c|p} &= 1 - \frac{2\sigma^2}{1-2c\sigma^2}(p-c) \end{aligned} \tag{31}$$

These expressions reveal some interesting features of the problem. Firstly, when $p = c$, we have the model being estimated and evaluated using the same loss function, *and* we have a forecasting model that spans the (MZ) moment conditions being used to evaluate it. In this case, we will find $\left(\gamma^*_{p|p}, \delta^*_{p|p}\right) = (0,1)\ \forall p$. Aside from the use of a correctly specified model, this is the best case scenario: the consumer is looking for a good fit in particular directions, the producer's model spans those directions, and the producer's model is calibrated using the consumer's loss function.

Secondly, when $p \neq c$, we will generally find $\left(\gamma^*_{c|p}, \delta^*_{c|p}\right) \neq (0,1)$, even though the forecast producer will think that the model is well-calibrated, since $\left(\gamma^*_{p|p}, \delta^*_{p|p}\right) = (0,1)\ \forall p$. That is, the producer will observe that his model spans the MZ conditions, and when evaluating his model using those conditions he will find no evidence against it, but the consumer will evaluate it using her loss function, and will in general find it sub-optimal.

Finally, when $\mu = 0$ we obtain some cautionary results. Firstly, we find $\left(\alpha^*_p, \beta^*_p\right) = \left(\sigma^2, 0\right)\ \forall p$. The lack of sensitivity of the estimated model parameters to the loss function parameter is suggestive of correct specification, but that is not the case here: the model remains misspecified. Secondly, if we evaluate this forecast using a Bregman-MZ test (imposing $\gamma^*_{c|p} = 0$ to overcome the perfect collinearity in the MZ regression), we find that $\delta^*_{c|p} = 1\ \forall\ p, c$, which is again suggestive of a correctly specified model, but is not the case. This case is one where the form of the model misspecification is such that the estimated model parameters happen to be insensitive to the loss function used for estimation *and* the moment conditions used to evaluate the forecast are not such that the misspecification can be detected.

Note that if the forecast rationality test was consistent (e.g., Corradi and Swanson, 2002) then,

asymptotically, *all* misspecified models will be rejected, and so the test conclusion would be the same regardless of the Bregman loss used in implementation. The sensitivity of forecast rationality tests to the choice of loss function are only of concern when an inconsistent test (such as the popular Mincer-Zarnowitz test) is used for evaluation.

4 Simulation-based results for realistic scenarios

In this section I present four scenarios to provide a more realistic description of the real-world complications involved with forecast construction, and show that these complications lead to sensitivity in the ranking of competing forecasts to the choice of consistent or proper loss functions.

For the first example, consider a point forecast based on a Bregman loss function, and so the target functional is the conditional mean. Assume that the data generating process is a stationary AR(5), with a strong degree of persistence:

$$Y_t = Y_{t-1} - 0.02Y_{t-2} - 0.02Y_{t-3} - 0.01Y_{t-4} - 0.01Y_{t-5} + \varepsilon_t, \;\; \varepsilon_t \sim iid\; N(0,1) \tag{32}$$

As forecasts, consider the comparison of a misspecified (but parsimonious) model with a correctly-specified model that is subject to estimation error. The first forecast is based on a random walk assumption:

$$\hat{Y}_t^A = Y_{t-1} \tag{33}$$

and the second forecast is based on a (correctly-specified) AR(5) model with parameters estimated using OLS on a rolling window of 100 observations:

$$\hat{Y}_t^B = \hat{\phi}_{0,t} + \hat{\phi}_{1,t}Y_{t-1} + \hat{\phi}_{2,t}Y_{t-2} + \hat{\phi}_{3,t}Y_{t-3} + \hat{\phi}_{4,t}Y_{t-4} + \hat{\phi}_{5,t}Y_{t-5} \tag{34}$$

where $\hat{\phi}_{j,t}$ is the estimate of ϕ_j based on data from $t-100$ to $t-1$. I simulate this design for 10,000 observations, and report the differences in average losses for a variety of homogeneous and exponential Bregman loss functions in Figure 5. From this figure we see that the ranking of these two forecasts is indeed sensitive to the choice of Bregman loss function. Under squared-error loss (corresponding to parameters 2 and 0 respectively for the homogeneous and exponential Bregman loss functions) the average loss difference is negative, indicating that the AR(5) model has larger

average loss than the random walk model, and thus the use of a parsimonious but misspecified model is preferred to the use of a correctly specified model that is subject to estimation error. However, the ranking is reversed for homogeneous Bregman loss functions with parameter above about 3.5, and for exponential Bregman loss functions with parameter greater than about 0.5 in absolute value.

[INSERT FIGURE 5 ABOUT HERE]

For our second example consider quantile forecasts for a heteroskedastic time series process. Consider a target variable governed by an AR-GARCH model, with parameters representative of those found for daily stock returns:

$$\begin{aligned} Y_t &= \mu_t + \sigma_t \varepsilon_t, \;\; \varepsilon_t \sim N(0,1) \\ \text{where } \mu_t &= 0.03 + 0.05 Y_{t-1} \\ \sigma_t^2 &= 0.05 + 0.9\sigma_{t-1}^2 + 0.05\sigma_{t-1}^2 \varepsilon_{t-1}^2 \end{aligned} \tag{35}$$

I compare two forecasts based on non-nested information sets. The first forecast exploits knowledge of the conditional mean, but assumes a constant conditional variance, while the second is the reverse:

$$\begin{aligned} \hat{Y}_t^A &= \mu_t + \bar{\sigma}\Phi^{-1}(\alpha) \\ \hat{Y}_t^B &= \bar{\mu} + \sigma_t \Phi^{-1}(\alpha) \end{aligned} \tag{36}$$

where $\bar{\mu} = E[Y_t]$ and $\bar{\sigma}^2 = V[Y_t]$. I consider these forecasts for two quantiles, a tail quantile ($\alpha = 0.05$) and a quantile somewhere between the tail and the center of the distribution ($\alpha = 0.25$). I compare these forecasts using the homogeneous GPL loss function in equation (12), and report the results based on a simulation of 10,000 observations.

In the right panel of Figure 6, where $\alpha = 0.05$, we see that the forecaster who has access to volatility information (Forecaster B) has lower average loss, across all values of the loss function parameter, than the forecaster who has access only to mean information. This is consistent with previous empirical research on the importance of volatility on estimates of tails. However, when looking at a quantile somewhere between the tails and the center, $\alpha = 0.25$, we see that the ranking

of these forecasts switches: for loss function parameter values less than about one, the forecaster with access to mean information has lower average loss, while for loss function parameter values above one we see the opposite.

[INSERT FIGURE 6 ABOUT HERE]

In a third example, consider the problem of forecasting the distribution of the target variable. I use a GARCH(1,1) specification (Bollerslev, 1986) for the conditional variance, and a left-skewed t distribution (Hansen, 1994) for the standardized residuals:

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= 0.05 + 0.9\sigma_{t-1}^2 + 0.05\sigma_{t-1}^2\varepsilon_{t-1}^2 \\ \varepsilon_t &\sim iid\ Skew\ t\left(0, 1, 6, -0.25\right) \end{aligned} \tag{37}$$

I take the first distribution forecast to be based on the Normal distribution, with mean zero and variance estimated using the past 100 observations:

$$\hat{F}_{A,t}\left(x\right) = \Phi\left(\frac{x}{\hat{\sigma}_t}\right), \text{ where } \hat{\sigma}_t^2 = \frac{1}{100}\sum\nolimits_{j=1}^{100} Y_{t-j}^2 \tag{38}$$

This is a parsimonious specification, but it imposes an incorrect model for the predictive distribution. The second forecast is based on the empirical distribution function (EDF) of the data over the past 100 observations:

$$\hat{F}_{B,t}\left(x\right) = \frac{1}{100}\sum_{j=1}^{100} \mathbf{1}\left\{Y_{t-j} \leq x\right\} \tag{39}$$

This nonparametric specification is more flexible than the first, but will inevitably contain more estimation error. I consider the weighted CRPS scoring rule from equation (17) where the weights are based on the standard Normal CDF:

$$\omega\left(z;\lambda\right) \equiv \lambda\Phi\left(z\right) + \left(1-\lambda\right)\left(1-\Phi\left(z\right)\right), \quad \lambda \in [0,1] \tag{40}$$

When $\lambda = 0$, the weight function is based on $1 - \Phi$, and thus places more weight on the left tail than the right tail. When $\lambda = 0.5$ the weighting scheme is flat and weights both tails equally. When $\lambda = 1$ the weight function places more weight on the right tail than the left tail.

This design is simulated for 10,000 observations, and the differences in average losses across weighting schemes (λ) are shown in Figure 7. We see that the ranking of these two distribution forecasts is sensitive to the choice of (proper) scoring rule: for weights below about 0.25 (i.e., those with a focus on the left tail), we find the EDF is preferred to the Normal distribution, while for weights above 0.25, including the equal-weighted case at 0.5, the Normal distribution is preferred to the EDF. Thus, the additional estimation error in the EDF generally leads to it being beaten by the parsimonious, misspecified, Normal distribution, *unless* the scoring rule places high weight on the left tail, which is long given the left-skew in the true distribution.

[INSERT FIGURE 7 ABOUT HERE]

Finally, related to the optimal approximation and forecast adjustment problem described in Section 3.1, consider the problem of approximating the true process for the conditional variance of an asset return with some misspecified model. We take the DGP to be a GARCH(1,1):

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t, \;\; \varepsilon_t \sim \text{iid } F_\varepsilon\left(0,1;\nu\right) \\ \sigma_t^2 &= \bar{\sigma}^2\left(1-\alpha-\beta\right)+\beta\sigma_{t-1}^2+\alpha y_{t-1}^2 \end{aligned} \tag{41}$$

We use a scaled and re-centered χ^2_ν distribution for F_ε, which generates (positive) skewness in the standardized residuals. (This is done so that optimization under the "QLIKE" loss function does not correspond to maximum likelihood, which has optimality properties that are not common for loss function-based estimators.) We consider approximating this variable using an ARCH(1) model:

$$\tilde{\sigma}_t^2 = \gamma_0 + \gamma_1 y_{t-1}^2 \tag{42}$$

We consider two methods for estimating the coefficients of the approximating model: the first is Gaussian quasi-maximum likelihood, which corresponds to minimizing the expected loss of a homogeneous Bregman loss function with shape parameter $k = 0$. The second is using standard OLS, which corresponds to minimizing the expected loss of a homogeneous Bregman loss function with shape parameter $k = 2$. The parameters of the approximating model are estimated using a sample of T observations, and the loss from the resulting forecast is computed in an out-of-sample

period containing P observations. In the simulation results below we set $T \in \{100, 500, 1000\}$, $P = 100$, and $\left(\bar{\sigma}^2, \beta, \alpha, \nu\right) = (1, 0.8, 0.1, 3)$. We repeat this simulation 10,000 times to obtain average out-of-sample losses. We also present the results for the infeasible case that $T \to \infty$ (approximated by using a sample of size 1,000,000) to see the results when estimation error is removed, highlighting the bias-variance trade-off that takes place in the presence of estimation error. (The limiting parameters for this case are presented in the bottom two rows of the table. They are the same regardless of the evaluation loss function.)

Table 2 below reveals that for all sample sizes, including the limit, average out-of-sample loss is smaller when the parameters of the approximating model are estimated using the same loss function as the one used for evaluation. This is consistent with the theory in the previous section, and with the binary prediction problem considered in Elliott, *et al.* (2014), but need not always be the case, in particular for smaller sample sizes. The key conclusion is that it is indeed possible, thus highlighting the potential importance of matching the loss functions used for estimation and evaluation when estimating a misspecified model.

Next, consider linear forecast adjustments as in equation (23) in Section 3.2. For this example, I will assume that the forecast producer uses quadratic loss, and the forecast consumer has a homogeneous-Bregman loss function with parameter c, which nests quadratic loss at $c = 2$. I will focus on the population values of the forecast producer's model and the optimal adjustment parameters, approximated by setting T =1,000,000, and reported in Table 2. In Figure 8 I present the coefficients of the linear adjustment function. We see that when $c = p = 2$, no adjustment is made: the adjustment parameters $\left(\gamma^*_{c|p}, \delta^*_{c|p}\right)$ equal $(0, 1)$. But when $c < p$, we observe the weight on the forecast rising above one ($\delta^*_{c|p} > 1$) while the intercept drops below zero ($\gamma^*_{c|p} < 0$), and the opposite holding when $c > p$. If the forecast producer had been using a correctly specified model (i.e., the GARCH model in equation (41)) *or* if the producer had calibrated his misspecified model using the consumer's loss function, then the adjustment parameters would be $(0, 1)$ for all values of c.

[INSERT TABLE 2 AND FIGURE 8 ABOUT HERE]

5 Empirical illustration: Evaluating professional forecasters

In this section I illustrate the above ideas using survey forecasts of U.S. inflation. Inflation forecasts are central to many important economic decisions, perhaps most notably those of the Federal Open Markets Committee in their setting of the Federal Funds rate, but also pension funds, insurance companies, and asset markets more broadly. Inflation is also notoriously hard to predict, with many methods failing to beat a simple random walk model, see Faust and Wright (2013) for a recent comprehensive survey.

Firstly, I consider a comparison of the consensus forecast (defined as the cross-respondent median) of CPI inflation from the Survey of Professional Forecasters (available from `http://goo.gl/L4A897`), and from the Thomson Reuters/University of Michigan Survey of Consumers (available from `http://goo.gl/s8dCEz`). The SPF gathers forecasts quarterly for a range of horizons from one quarter to ten years, whereas the Michigan survey gathers forecasts monthly, but only for one- and five-year horizons. For this illustration I examine only the one-year forecast. The sample period is 1982Q3 to 2014Q1, a total of 127 observations. As the "actual" series I use the 2014Q2 vintage of the "real time" CPI data (available at `http://goo.gl/AH6gAO`). A plot of the forecasts and realized inflation series is presented in Figure 9, and summary statistics are presented in Table 4.

[INSERT FIGURE 9 AND TABLE 4 ABOUT HERE]

I also consider a comparison of individual respondents to the Survey of Professional Forecasters. These respondents are identified in the database only by a numerical identifier, and I select Forecasters 20, 506 and 528, as they all have relatively long histories of responses. For the individual comparisons I focus on the one-quarter-ahead forecast, as these have the most non-empty cells.

Given the difficulty in capturing the dynamics of inflation, it is reasonable to expect that all forecasters are subject to model misspecification. Moreover, these forecasts are quite possibly based on nonnested information sets, particularly in the comparison of professional forecasters with the Michigan survey of consumers. Thus the practical issues highlighted in Section 2 are relevant here.

Figure 10 presents the results of comparisons of these forecasts, for a range of Bregman loss

functions. In the left panel I consider homogeneous Bregman loss functions (equation 6) with parameter ranging from 1.1 to 4 (nesting squared-error loss at 2) and in the right panel I consider exponential Bregman loss functions (equation 7) with parameter ranging from -1 to 1 (nesting squared-error loss at 0). In the top panel we see that the sign of the difference in average losses varies with the parameter of the loss function: the SPF forecast had (slightly) lower average loss for values of the Bregman parameter less than 2 and 0 in the homogeneous and exponential cases respectively, while the reverse holds true for parameters above these values. (The difference in average loss is very near zero for the squared-error loss case.) This indicates that the ranking of professional vs. consumer forecasts of inflation depends on whether over-predictions are more or less costly than under-predictions, see Figure 1.

In the middle panel I compare SPF forecaster 20 to forecaster 506, and we again see strong sensitivity to the choice of loss function: for loss functions that penalize under-prediction more than over-prediction (homogeneous Bregman with parameter less than two, and exponential Bregman with parameter less than zero) forecaster 20 is preferred, while when the loss functions penalize over-prediction more than under-prediction the ranking is reversed. In the lower panel we see an example of a robust ranking: Forecaster 506 has larger average loss than Forecaster 528 for all homogeneous and exponential Bregman loss functions considered.

In Figure 11 I present the intercept and slope parameters (γ and δ) of the Mincer-Zarnowitz forecast rationality test, optimized using a variety of Bregman loss functions, for the SPF and Michigan forecasts. If the forecasts were truly optimal, we would observe $(\gamma, \delta) = (0, 1)$ for all Bregman loss functions. If the forecasts are based on misspecified models, but optimally calibrated for one of the Bregman loss functions considered in Figure 11, then we would observe $(\gamma, \delta) = (0, 1)$ for that loss function. For the Michigan forecasts in the lower panel, we see little sensitivity of the MZ parameters to the Bregman loss function, while for the SPF forecasts we observe some sensitivity, particularly for $k > 2$ and $a > 0$. However, we can clearly see that the parameter restriction is not satisfied for any loss function considered in Figure 11, and joint tests of the two parameter restrictions yield p-values far less than 0.01 in all cases, rejecting forecast optimality.

[INSERT FIGURES 10 AND 11 ABOUT HERE]

6 Conclusion

This paper shows that in the presence of realistic complexities such as nonnested forecaster information sets, misspecified predictive models, and predictive models subject to estimation error, declaring the target functional is not generally sufficient to elicit a forecaster's best (according to a given, consistent, loss function) forecast. The results presented here suggest that best practice for point forecasting is to declare the single, specific loss function that will be used to evaluate forecasts, and to make that loss function consistent for the target functional of interest to the forecast consumer. Reacting to this, forecasters may then wish to estimate their predictive models, if a model is being used, based on the loss function that will evaluate their forecast. This will ensure that the estimated parameter converges to the parameter that minimizes expected loss under that loss function (the "pseudo-true" parameter), though a trade-off may exist between the variance of an estimator and the distance between its probability limit and the pseudo-true parameter. This is related to work on estimation under the "relevant cost function," see Weiss (1996), Christoffersen and Jacobs (2004), Skouras (2007) and Elliott, *et al.* (2014) for applications in economics and finance.

Appendix

Proof of Proposition 1. We will show that under assumptions (i) and (ii), $MSE_B \geq MSE_A \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t \Rightarrow E\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \geq E\left[L\left(Y_t, \hat{Y}_t^A\right)\right] \ \forall L \in \mathcal{L}_{Bregman}$.

For the first implication: We are given that $MSE_B \geq MSE_A$, and assume that $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B \ \forall t$. This means that $\min_{\hat{y}} \ E\left[(Y_t - \hat{y})^2 \,|\mathcal{F}_t^A\right] \geq \min_{\hat{y}} \ E\left[(Y_t - \hat{y})^2 \,|\mathcal{F}_t^B\right] \ a.s. \ \forall t$, and so $E\left[\left(Y_t - \hat{Y}_t^A\right)^2 |\mathcal{F}_t^A\right] \geq E\left[\left(Y_t - \hat{Y}_t^B\right)^2 |\mathcal{F}_t^B\right] \ a.s. \ \forall t$ by assumption (ii), and $E\left[\left(Y_t - \hat{Y}_t^A\right)^2\right] \geq E\left[\left(Y_t - \hat{Y}_t^B\right)^2\right]$ by the law of iterated expectations, which is a contradiction. Thus $MSE_B \geq MSE_A \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t$.

Now consider the second implication: Let

$$Y_t = \hat{Y}_t^A + \eta_t = \hat{Y}_t^B + \eta_t + \varepsilon_t \tag{43}$$

Then

$$\begin{aligned} E\left[L\left(Y_t, \hat{Y}_t^A\right) - L\left(Y_t, \hat{Y}_t^B\right)\right] &= E\left[-\phi\left(\hat{Y}_t^A\right) - \phi'\left(\hat{Y}_t^A\right)\eta_t + \phi\left(\hat{Y}_t^B\right) + \phi'\left(\hat{Y}_t^B\right)(\eta_t + \varepsilon_t)\right] \\ &= E\left[\phi\left(\hat{Y}_t^B\right) - \phi\left(\hat{Y}_t^A\right)\right] \end{aligned} \tag{44}$$

since $E\left[\phi'\left(\hat{Y}_t^A\right)\eta_t\right] = E\left[\phi'\left(\hat{Y}_t^A\right)E\left[\eta_t|\mathcal{F}_t^A\right]\right]$ by the law of iterated expectations and $E\left[\eta_t|\mathcal{F}_t^A\right] = 0$, and similarly for $E\left[\phi'\left(\hat{Y}_t^B\right)(\eta_t + \varepsilon_t)\right]$. Next, consider the second-order mean-value expansion:

$$\phi\left(\hat{Y}_t^A\right) = \phi\left(\hat{Y}_t^B\right) - \phi'\left(\hat{Y}_t^B\right)\varepsilon_t + \phi''\left(\ddot{Y}_t^A\right)\varepsilon_t^2 \tag{45}$$

where $\ddot{Y}_t^A = \lambda_t \hat{Y}_t^A + (1 - \lambda_t)\hat{Y}_t^B$, for $\lambda_t \in [0, 1]$. Thus

$$E\left[L\left(Y_t, \hat{Y}_t^A\right) - L\left(Y_t, \hat{Y}_t^B\right)\right] = E\left[\phi'\left(\hat{Y}_t^B\right)\varepsilon_t\right] - E\left[\phi''\left(\ddot{Y}_t^A\right)\varepsilon_t^2\right] \leq 0 \tag{46}$$

since $E\left[\phi'\left(\hat{Y}_t^B\right)\varepsilon_t\right] = 0$ and ϕ is convex. ■

Proof of Corollary 1. (a) The proof follows by noting that the ranking of any pair (i, j) of forecasters satisfies the conditions of Proposition 1, and by ranking all possible pairs we obtain a complete ranking of all N forecasters.

(b) Consider ranking (i, i^*). This proof requires a slight generalization of Proposition 1, to reflect the fact that only the forecaster with the larger information set (i^*, in this case) is required

to issue an optimal forecast. Under assumptions (b)(i) and (b)(ii), we will show $MSE_i \geq MSE_{i^*} \Rightarrow \mathcal{F}_t^i \subseteq \mathcal{F}_t^{i^*}\ \forall t \Rightarrow E\left[L\left(Y_t, \hat{Y}_t^i\right)\right] \geq E\left[L\left(Y_t, \hat{Y}_t^{i^*}\right)\right]\ \forall L \in \mathcal{L}_{Bregman}$.

For the first implication: We are given that $MSE_i \geq MSE_{i^*}$, and assume that $\mathcal{F}_t^{i^*} \subseteq \mathcal{F}_t^i\ \forall t$. Under assumption (b)(ii) this means that forecaster i is using an optimal forecast but forecaster i^* may not be. We then have $E\left[\left(Y_t - \hat{Y}_t^{i^*}\right)^2 |\mathcal{F}_t^{i^*}\right] \geq \min_{\hat{y}}\ E\left[(Y_t - \hat{y})^2 |\mathcal{F}_t^{i^*}\right] \geq E\left[\left(Y_t - \hat{Y}_t^i\right)^2 |\mathcal{F}_t^i\right]$ $a.s.\ \forall t$, and $E\left[\left(Y_t - \hat{Y}_t^{i^*}\right)^2\right] \geq E\left[\left(Y_t - \hat{Y}_t^i\right)_t^2\right]$ by the law of iterated expectations (LIE), which is a contradiction. Thus $MSE_i \geq MSE_{i^*} \Rightarrow \mathcal{F}_t^i \subseteq \mathcal{F}_t^{i^*}\ \forall t$.

The second implication: Let

$$\bar{L}^j \equiv E\left[L\left(Y_t, \hat{Y}_t^j; \phi_j\right)\right],\ \ j \in \{i, i^*\}$$

where $L\left(\cdot, \cdot; \phi_j\right)$ is a Bregman loss function defined by ϕ_j, a convex function. Under assumptions (i) and (ii) we know that $\hat{Y}_t^{i^*}$ is the solution to $\min_{\hat{y}}\ E\left[L^{i^*}\left(Y_t, \hat{y}; \phi_{i^*}\right) |\mathcal{F}_t^{i^*}\right]$ where $L^{i^*} \in \mathcal{L}_{Bregman}$, ϕ_{i^*} is a convex function. Thus $\hat{Y}_t^{i^*} = E\left[Y_t | \mathcal{F}_t^{i^*}\right]$ for all possible distributions of Y_t, and from Banerjee *et al.* (2005) for example, we know that this implies that $\hat{Y}_t^{i^*}$ moreover satisfies:

$$\hat{Y}_t^{i^*} = \arg\min_{\hat{y}}\ E\left[\phi\left(Y_t\right) - \phi\left(\hat{y}\right) - \phi'\left(\hat{y}\right)\left(Y_t - \hat{y}\right) |\mathcal{F}_t^{i^*}\right]$$

for any convex function ϕ. Given that $\mathcal{F}_t^i \subseteq \mathcal{F}_t^{i^*}\ \forall t$, and acknowledging the possible suboptimality of forecast i, we then have $E\left[L\left(Y_t, \hat{Y}_t^i; \phi\right) |\mathcal{F}_t^i\right] \geq \min_{\hat{y}}\ E\left[L\left(Y_t, \hat{y}; \phi\right) |\mathcal{F}_t^i\right] \geq E\left[L\left(Y_t, \hat{Y}_t^{i^*}; \phi\right) |\mathcal{F}_t^{i^*}\right]$ $a.s.\ \forall t$ for any convex function ϕ, and by the LIE we obtain the second implication. The ranking of (i, j) for $i^* \notin \{i, j\}$ involves comparing forecasters with potentially nonnested information sets and potentially misspecified models, and so thus the results of Proposition 2 apply. ■

Holzmann and Eulert (2014) present a different proof of the two results below. We present the following for comparability with the conditional mean case presented in Proposition 1 of the paper.

Proof of Proposition 3. We will show that under assumptions (i) and (ii), $LinLin_B^\alpha \geq LinLin_A^\alpha \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A\ \forall t \Rightarrow E\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \geq E\left[L\left(Y_t, \hat{Y}_t^A\right)\right]\ \forall L \in \mathcal{L}_{GPL}^\alpha$, where $LinLin_j^\alpha \equiv E\left[LinLin\left(Y_t, \hat{Y}_t^j\right)\right]$ for $j \in \{A, B\}$ and $LinLin$ is the "Lin-Lin" loss function in equation (11).

First: we are given that $LinLin_B^\alpha \geq LinLin_A^\alpha$, and assume that $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B\ \forall t$. This means that $\min_{\hat{y}}\ E\left[LinLin\left(Y_t, \hat{y}\right) |\mathcal{F}_t^A\right] \geq \min_{\hat{y}}\ E\left[LinLin\left(Y_t, \hat{y}\right) |\mathcal{F}_t^B\right]$ $a.s.\ \forall t$, and so $E\left[LinLin\left(Y_t, \hat{Y}_t^A\right) |\mathcal{F}_t^A\right] \geq$

$E\left[LinLin\left(Y_t, \hat{Y}_t^B\right)|\mathcal{F}_t^B\right]$ $a.s.\ \forall t$ by assumption (ii), and $E\left[LinLin\left(Y_t, \hat{Y}_t^A\right)\right] \geq E\left[LinLin\left(Y_t, \hat{Y}_t^B\right)\right]$ by the LIE, which is a contradiction. Thus $LinLin_B^\alpha \geq LinLin_A^\alpha \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A\ \forall t$. Next: Let

$$\bar{L}^j \equiv E\left[L\left(Y, \hat{Y}^j; \alpha, g_j\right)\right],\ \ j \in \{A, B\}$$

where $L(\cdot, \cdot; \alpha, g_j)$ is a GPL loss function defined by g_j, a nondecreasing function. Under assumption (ii) we know that $\hat{Y}^j$ is the solution to $\min_{\hat{y}}\ E\left[L\left(Y, \hat{Y}^j; \alpha, g_j\right)|\mathcal{F}^j\right]$ where $L \in \mathcal{L}_{GPL}^\alpha$, g_j is a nondecreasing function, and and $j \in \{A, B\}$. It is straightforward to show that $\hat{Y}^j$ then satisfies $\alpha = E\left[\mathbf{1}\left\{Y \leq \hat{Y}^j\right\}|\mathcal{F}^j\right]$. This holds for all possible (conditional) distributions of Y, and from Saerens (2000) and Gneiting (2011b) we know that this implies (by the necessity of GPL loss for optimal quantile forecasts) that $\hat{Y}^j$ then moreover satisfies

$$\hat{Y}^j = \arg\min_{\hat{y}}\ E\left[\left(\mathbf{1}\left\{Y \leq \hat{y}\right\} - \alpha\right)\left(g\left(\hat{y}\right) - g\left(Y\right)\right)|\mathcal{F}^j\right]$$

for any nondecreasing function g. If $\mathcal{F}^B \subseteq \mathcal{F}^A$ then by the LIE we have $\bar{L}^B(g) \geq \bar{L}^A(g)$ for any nondecreasing function g. ■

Proof of Proposition 5. We again prove this result by showing that $E\left[L\left(F_t^A, Y_t\right)\right] \leq E\left[L\left(F_t^B, Y_t\right)\right]$ for some $L \in \mathcal{L}_{\text{Proper}} \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A\ \forall t \Rightarrow E\left[L\left(F_t^A, Y_t\right)\right] \leq E\left[L\left(F_t^B, Y_t\right)\right]\ \forall\ L \in \mathcal{L}_{\text{Proper}}$. First: we are given that $E\left[L\left(F_t^A, Y_t\right)\right] \leq E\left[L\left(F_t^B, Y_t\right)\right]$, and assume that $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B\ \forall t$. Under assumption (ii), this implies that we can take F_t^B as the data generating process for Y_t. Then $E\left[L\left(F_t^B, Y_t\right)|\mathcal{F}_t^B\right] = E_{F_t^B}\left[L\left(F_t^B, Y_t\right)|\mathcal{F}_t^B\right] \leq E_{F_t^B}\left[L\left(F_t^A, Y_t\right)|\mathcal{F}_t^B\right]\ \forall t$ by assumption (ii) and the propriety of L. By the LIE this implies $E\left[L\left(F_t^B, Y_t\right)\right] \leq E\left[L\left(F_t^A, Y_t\right)\right]$ which is a contradiction. Thus $E\left[L\left(F_t^A, Y_t\right)\right] \leq E\left[L\left(F_t^B, Y_t\right)\right]$ for some $L \in \mathcal{L}_{\text{Proper}} \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A\ \forall t$. Next, using similar logic to above, given $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A$ we have that $E\left[L\left(F_t^A, Y_t\right)\right] \leq E\left[L\left(F_t^B, Y_t\right)\right]$ for any $L \in \mathcal{L}_{\text{Proper}}$, completing the proof. ■

Proof of Proposition 7. (a) Recall the forms of these two loss functions from equations (4) and (10). Then the expected loss w.r.t. distribution F_t from a convex combination of these loss

functions is

$$\begin{aligned} E_{F_t}\left[L\left(Y_t,\hat{y}\right)\right] &= \lambda E_{F_t}\left[\phi\left(Y_t\right)\right]-\lambda\phi\left(\hat{y}\right)-\lambda\phi'\left(\hat{y}\right)\left(E_{F_t}\left[Y_t\right]-\hat{y}\right) \\ &\quad +\left(1-\lambda\right)\left(E_{F_t}\left[\mathbf{1}\left\{Y_t\leq\hat{y}\right\}\right]-1/2\right)g\left(\hat{y}\right) \\ &\quad +\left(1-\lambda\right)E_{F_t}\left[\left(\mathbf{1}\left\{Y_t\leq\hat{y}\right\}-1/2\right)g\left(Y_t\right)\right] \end{aligned}$$

And the first derivative is

$$\frac{\partial}{\partial\hat{y}}E_{F_t}\left[L\left(Y,\hat{y}\right)\right]=-\lambda\phi''\left(\hat{y}\right)\left(E_{F_t}\left[Y_t\right]-\hat{y}\right)+\left(1-\lambda\right)\left(E_{F_t}\left[\mathbf{1}\left\{Y_t\leq\hat{y}\right\}\right]-1/2\right)g'\left(\hat{y}\right)$$

using assumption (i) that F_t is continuous. Then note that $E_{F_t}\left[\mathbf{1}\left\{Y_t\leq\hat{y}\right\}\right]\equiv F_t\left(\hat{y}\right)$, and recall that F_t is symmetric$\Rightarrow E_{F_t}\left[Y_t\right]=Median_{F_t}\left[Y_t\right]\Rightarrow F_t\left(E\left[Y_t|\mathcal{F}_t\right]\right)=1/2$. Thus $\hat{Y}_t^*=E_{F_t}\left[Y_t\right]$ is a solution to the optimization problem:

$$-\lambda\phi''\left(E_{F_t}\left[Y_t\right]\right)\left(E_{F_t}\left[Y_t\right]-E_{F_t}\left[Y_t\right]\right)+\left(1-\lambda\right)\left(E_{F_t}\left[\mathbf{1}\left\{Y_t\leq E_{F_t}\left[Y_t\right]\right\}\right]-1/2\right)g'\left(E_{F_t}\left[Y_t\right]\right)=0$$

Note that this result holds whether or not the target variable is truly symmetrically distributed.

(b) We will show that under assumptions (i)–(iii) , $MSE_B\geq MSE_A\Rightarrow\mathcal{F}_t^B\subseteq\mathcal{F}_t^A\ \forall t$, $MAE_B\geq MAE_A\Rightarrow\mathcal{F}_t^B\subseteq\mathcal{F}_t^A\ \forall t$, and $\mathcal{F}_t^B\subseteq\mathcal{F}_t^A\ \forall t\Rightarrow E\left[L\left(Y_t,\hat{Y}_t^B\right)\right]\geq E\left[L\left(Y_t,\hat{Y}_t^A\right)\right]\ \forall L\in\mathcal{L}_{Breg\times GPL}$. First implication: We are given that $MSE_B\geq MSE_A$, and assume that $\mathcal{F}_t^A\subseteq\mathcal{F}_t^B\ \forall t$. This means that $\min_{\hat{y}}\ E\left[\left(Y_t-\hat{y}\right)^2|\mathcal{F}_t^A\right]\geq\min_{\hat{y}}\ E\left[\left(Y_t-\hat{y}\right)^2|\mathcal{F}_t^B\right]\ a.s.\ \forall t$. Since the quadratic loss function is in $\mathcal{L}_{Breg\times GPL}$, then by part (a) and assumption (iii) we have $E\left[\left(Y_t-\hat{Y}_t^A\right)^2|\mathcal{F}_t^A\right]\geq E\left[\left(Y_t-\hat{Y}_t^B\right)^2|\mathcal{F}_t^B\right]\ a.s.\ \forall t$, and $E\left[\left(Y_t-\hat{Y}_t^A\right)^2\right]\geq E\left[\left(Y_t-\hat{Y}_t^B\right)^2\right]$ by the law of iterated expectations, which is a contradiction. Thus $MSE_B\geq MSE_A\Rightarrow\mathcal{F}_t^B\subseteq\mathcal{F}_t^A\ \forall t$. The same reasoning applies for the implication $MAE_B\geq MAE_A\Rightarrow\mathcal{F}_t^B\subseteq\mathcal{F}_t^A\ \forall t$. Finally, consider the third implication. The expected loss difference is

$$\begin{aligned} E\left[L\left(Y_t,\hat{Y}_t^A\right)-L\left(Y_t,\hat{Y}_t^B\right)\right] &= \lambda E\left[L_{Breg}\left(Y_t,\hat{Y}_t^A\right)-L_{Breg}\left(Y_t,\hat{Y}_t^B\right)\right] \\ &\quad +\left(1-\lambda\right)E\left[L_{GPL}\left(Y_t,\hat{Y}_t^A\right)-L_{GPL}\left(Y_t,\hat{Y}_t^B\right)\right] \end{aligned}$$

where $L_{Breg}\in\mathcal{L}_{Bregman}$, $L_{GPL}\in\mathcal{L}_{GPL}^{1/2}$, and $\lambda\in[0,1]$. Noting that assumptions (i)–(iii) satisfy those of Proposition 1, we then have $\mathcal{F}_t^B\subseteq\mathcal{F}_t^A\ \forall t\Rightarrow E\left[L_{Breg}\left(Y_t,\hat{Y}_t^A\right)-L_{Breg}\left(Y_t,\hat{Y}_t^B\right)\right]\leq$

$0\ \forall L \in \mathcal{L}_{Bregman}$. Similarly, from the proof of Proposition 3 we also have $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A\ \forall t \Rightarrow E\left[L_{GPL}\left(Y_t, \hat{Y}_t^A\right) - L_{GPL}\left(Y_t, \hat{Y}_t^B\right)\right] \leq 0\ \forall L \in \mathcal{L}_{GPL}^{1/2}$. Since $\lambda \in [0,1]$ we then have $E\left[L\left(Y_t, \hat{Y}_t^A\right) - L\left(Y_t, \hat{Y}_t^B\right)\right] \leq 0$ for any $L \in \mathcal{L}_{Breg \times GPL}$.

(c) The proof of this negative result requires only an example. This can be constructed using methods similar to those for Propositions 2 and 4, and is omitted in the interest of brevity. ■

Proof of Proposition 8. The first derivative of interest is

$$\begin{aligned} \frac{\partial}{\partial \theta} E\left[L\left(Y_t, m\left(V_t;\theta\right);\phi\right)\right] &= E\left[\frac{\partial}{\partial \hat{y}} L\left(Y_t, \hat{y};\phi\right)\Big|_{\hat{y}=m(Z_t;\theta)} \cdot \frac{\partial m\left(V_t;\theta\right)}{\partial \theta}\right] \\ &= E\left[\phi''\left(m\left(V_t;\theta\right)\right)\left(Y_t - m\left(V_t;\theta\right)\right)\frac{\partial m\left(V_t;\theta\right)}{\partial \theta}\right] \end{aligned} \tag{47}$$

The first-order condition for the optimization is

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} E\left[L\left(Y_t, m\left(V_t;\theta\right);\phi\right)\right]\Big|_{\theta=\hat{\theta}_\phi^*} \\ &= E\left[\phi''\left(m\left(V_t;\hat{\theta}_\phi^*\right)\right)\left(Y_t - m\left(V_t;\hat{\theta}_\phi^*\right)\right)\frac{\partial m\left(V_t;\hat{\theta}_\phi^*\right)}{\partial \theta}\right] \\ &= E\left[\phi''\left(m\left(V_t;\hat{\theta}_\phi^*\right)\right)\left(E\left[Y_t|\mathcal{F}_t\right] - m\left(V_t;\hat{\theta}_\phi^*\right)\right)\frac{\partial m\left(V_t;\hat{\theta}_\phi^*\right)}{\partial \theta}\right], \text{ by the LIE} \end{aligned} \tag{48}$$

and note that this equality holds when $\hat{\theta}_\phi^* = \theta_0$ by assumption (i), and the solution is unique since ϕ is strictly convex and $\partial m/\partial\theta \neq 0$ a.s. by assumption (ii). ■

Proof of Proposition 9. This proof is very similar to that of Proposition 8. The first-order condition for the optimization yields:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} E\left[L\left(Y_t, g\left(\hat{Y}_{p,t}^*;\theta\right);\phi\right)\right]\Big|_{\theta=\theta_{c|p}^*} \\ &= E\left[\phi''\left(g\left(\hat{Y}_{p,t}^*;\theta_{c|p}^*\right)\right)\left(Y_t - g\left(\hat{Y}_{p,t}^*;\theta_{c|p}^*\right)\right)\frac{\partial g\left(\hat{Y}_{p,t}^*;\theta_{c|p}^*\right)}{\partial \theta}\right] \\ &= E\left[\phi''\left(g\left(\hat{Y}_{p,t}^*;\theta_{c|p}^*\right)\right)\left(E\left[Y_t|\mathcal{F}_t\right] - g\left(\hat{Y}_{p,t}^*;\theta_{c|p}^*\right)\right)\frac{\partial g\left(\hat{Y}_{p,t}^*;\theta_{c|p}^*\right)}{\partial \theta}\right], \text{ by the LIE} \end{aligned} \tag{49}$$

Note that this equality holds when $\theta^*_{c|p} = \theta_0$ by assumptions (i) and (ii), and the solution is unique since ϕ is strictly convex and $\partial g/\partial\theta \neq 0$ a.s. by assumption (iii). ■

Derivations for Example 1:

The first-order condition for the optimal parameter θ is:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} E\left[L\left(Y, m\left(X;\theta\right);\phi\right)\right] \\ &= E\left[\phi''\left(m\left(X;\theta\right)\right)\left(E\left[Y|X\right] - m\left(X;\theta\right)\right)\frac{\partial m\left(X;\theta\right)}{\partial\theta}\right] \\ &= 2E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}\left(X^2 - \alpha - \beta X\right)\left[1, X\right]'\right] \end{aligned} \tag{50}$$

So the two first-order conditions are:

$$0 = E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X^2\right] - \alpha E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}\right] - \beta E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X\right] \tag{51}$$

$$0 = E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X^3\right] - \alpha E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X\right] - \beta E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X^2\right]$$

Using properties of the Normal distribution we obtain the key moments from the above expressions:

$$E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}\right] = \exp\left\{p\left(\alpha+\beta\mu\right) + p^2\frac{\beta^2}{2}\sigma^2\right\} \tag{52}$$

$$E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X\right] = \exp\left\{p\left(\alpha+\beta\mu\right) + p^2\frac{\beta^2}{2}\sigma^2\right\}\left(\mu + p\beta\sigma^2\right)$$

$$E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X^2\right] = \exp\left\{p\left(\alpha+\beta\mu\right) + p^2\frac{\beta^2}{2}\sigma^2\right\}\left(\sigma^2 + \left(\mu + p\beta\sigma^2\right)^2\right)$$

$$E\left[\exp\left\{p\left(\alpha+\beta X\right)\right\}X^3\right] = \exp\left\{p\left(\alpha+\beta\mu\right) + p^2\frac{\beta^2}{2}\sigma^2\right\}\left(\mu + p\beta\sigma^2\right)\left(3\sigma^2 + \left(\mu + p\beta\sigma^2\right)^2\right)$$

References

[1] Banerjee, A., X. Guo and H. Wang, 2005, On the Optimality of Conditional Expectation as a Bregman Predictor, *IEEE Transactions on Information Theory*, 51(7), 2664-2669.

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

[3] Bregman, L. M., 1967, The Relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming, *USSR Computational Mathematics and Mathematical Physics*, 7, 200-217.

[4] Christoffersen, P. and F. X., Diebold, 1997, Optimal Prediction Under Asymmetric Loss, *Econometric Theory*, 13, 808-817.

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

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

[7] Elliott, G., D. Ghanem, and F. Krüger, 2014, Forecasting Conditional Probabilities of Binary Outcomes under Misspecification, working paper, Department of Economics, UC-San Diego.

[8] Engelberg, J., C. F. Manski, and J. Williams, 2009, Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters, *Journal of Business & Economic Statistics*, 27, 30-41.

[9] Faust, J. and J. H. Wright, 2013, Forecasting Inflation, in G. Elliott and A. Timmermann (eds.) *Handbook of Economic Forecasting, Volume 2*, Springer Verlag.

[10] Gneiting, T., 2011a, Making and Evaluating Point Forecasts, *Journal of the American Statistical Association*, 106(494), 746-762.

[11] Gneiting, T., 2011b, Quantiles as Optimal Point Forecasts, *International Journal of Forecasting*, 27, 197-207.

[12] Gneiting, T. and R. Ranjan, 2011, Comparing Density Forecasts using Threshold- and Quantile-Weighted Scoring Rules, *Journal of Business & Economic Statistics*, 29(3), 411-422.

[13] Gneiting, T. and A. E. Raftery, 2007, Strictly Proper Scoring Rules, Prediction and Estimation, *Journal of the American Statistical Association*, 102(477), 358-378.

[14] Granger, C. W. J., 1969, Prediction with a Generalized Cost of Error Function, *OR*, 20(2), 199-207.

[15] Hansen, B. E., 1994, Autoregressive Conditional Density Estimation, *International Economic Review*, 35(3), 705-730.

[16] Holzmann, H. and M. Eulert, 2014, The Role of the Information Set for Forecasting–with Applications to Risk Management, *Annals of Applied Statistics*, 8(1), 595-621.

[17] Komunjer, I., 2005, Quasi Maximum-Likelihood Estimation for Conditional Quantiles, *Journal of Econometrics*, 128, 137-164.

[18] Leitch, G. and J. E. Tanner, 1991, Economic Forecast Evaluation: Profits versus the Conventional Error Measures, *American Economic Review*, 81(3), 580-590.

[19] Merkle, E. C., and M. Steyvers, 2013, Choosing a Strictly Proper Scoring Rule, *Decision Analysis*, 10(4), 292-304.

[20] Patton, A. J., 2011, Volatility Forecast Comparison using Imperfect Volatility Proxies, *Journal of Econometrics*, 160(1), 246-256.

[21] Patton, A. J. and A. Timmermann, 2007, Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity, *Journal of Econometrics*, 140(2), 884-918.

[22] Savage, L. J., 1971, Elicitation of Personal Probabilities and Expectations, *Journal of the American Statistical Association*, 66(336), 783-801.

[23] Skouras, S., 2007, Decisionmetrics: A Decision-Based Approach to Econometric Modelling, *Journal of Econometrics*,137, 414-40.

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

[25] Weiss, A. A., 1996, Estimating Time Series Models using the Relevant Cost Function, *Journal of Applied Econometrics*, 11 539-560.

[26] West, K. D., H. Edison and D. Cho, 1993, A Utility-Based Comparison of Some Models of Exchange Rate Volatility, *Journal of International Economics*, 35, 23-45.

[27] White, H., 2001, *Asymptotic Theory for Econometricians*, Second Edition, San Diego, Academic Press.

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

Table 1: Optimal linear approximation example

	Exponential Bregman parameter (a)		
	-0.5	*0*	*0.25*
$\hat{\alpha}_a^*$	0.75	0.00	-3.00
$\hat{\beta}_a^*$	1.00	2.00	4.00

This table presents the optimal intercept ($\hat{\alpha}_a^*$) and slope ($\hat{\beta}_a^*$) parameters from linear approximations to a nonlinear conditional mean function, for three different values of the "exponential Bregman" loss function parameter, a.

Table 2: Average MSE and QLIKE loss for different estimation methods

Evaluation loss function	**QLIKE**		**MSE**	
Estimation loss function	*QLIKE*	*MSE*	*QLIKE*	*MSE*
100	1.505	1.509	4.535	4.326
500	1.428	1.435	4.277	4.239
1000	1.412	1.418	4.228	4.206
∞	1.399	1.401	4.179	4.169
γ_0^*	0.912	0.938	0.912	0.938
γ_1^*	0.093	0.071	0.093	0.071

Notes: This table presents the average out-of-sample loss from a volatility forecast from an ARCH(1) model (equation 42) estimated by minimizing either the QLIKE loss function or the quadratic loss function (MSE) over a rolling window of $T \in \{100, 500, 1000\}$ observations. The bottom row of the top panel presents the average loss when the parameters are set to the population limits, these limiting parameters are presented in the bottom two rows of the table.

Table 3: Expected losses

		Forecast		
		$\hat{Y}_p^*$	$\hat{Y}_{c\|p}^*$	$\hat{Y}_c^*$
Loss	$p = 0$	2.97	4.14	4.88
function	$c = 0.2$	8.66	7.38	6.83

Notes: This table presents the expected loss using two loss functions (exponential Bregman with parameters $p = 0$ and $c = 0.2$), for three forecasts: a forecast optimized using the $p = 0$ loss function ($\hat{Y}_p^*$), a forecast optimized using the $c = 0.2$ loss function ($\hat{Y}_c^*$), and a forecast that adjusts the first forecast to minimize the $c = 0.2$ loss function ($\hat{Y}_{c|p}^*$).

Table 4: Summary Statistics

Panel A: Consensus forecasts

	Actual	*SPF*	*Michigan*
Mean	2.633	3.246	3.196
Standard deviation	1.202	1.282	0.729
Minimum	-1.930	1.565	0.900
Maximum	5.662	8.058	6.700

Panel B: Individual forecasts

	Actual	*Forecaster 20*	*Forecaster 506*	*Forecaster 528*
Mean	4.213	3.295	1.339	1.542
Standard deviation	3.458	1.369	1.243	1.243
Minimum	-13.668	0.800	-2.600	-2.000
Maximum	16.645	11.400	3.829	3.600

Notes: This table presents summary statistics on realized inflation and inflation forecasts. Panel A considers one-year consensus forecasts, in percent, of annual US CPI inflation from the Survey of Professional Forecasters and from the Thomson Reuters/University of Michigan Survey of Consumers. Panel B considers one-quarter forecasts from individual respondents to the Survey of Professional Forecasters, of US CPI inflation, annualized, in percent.

.

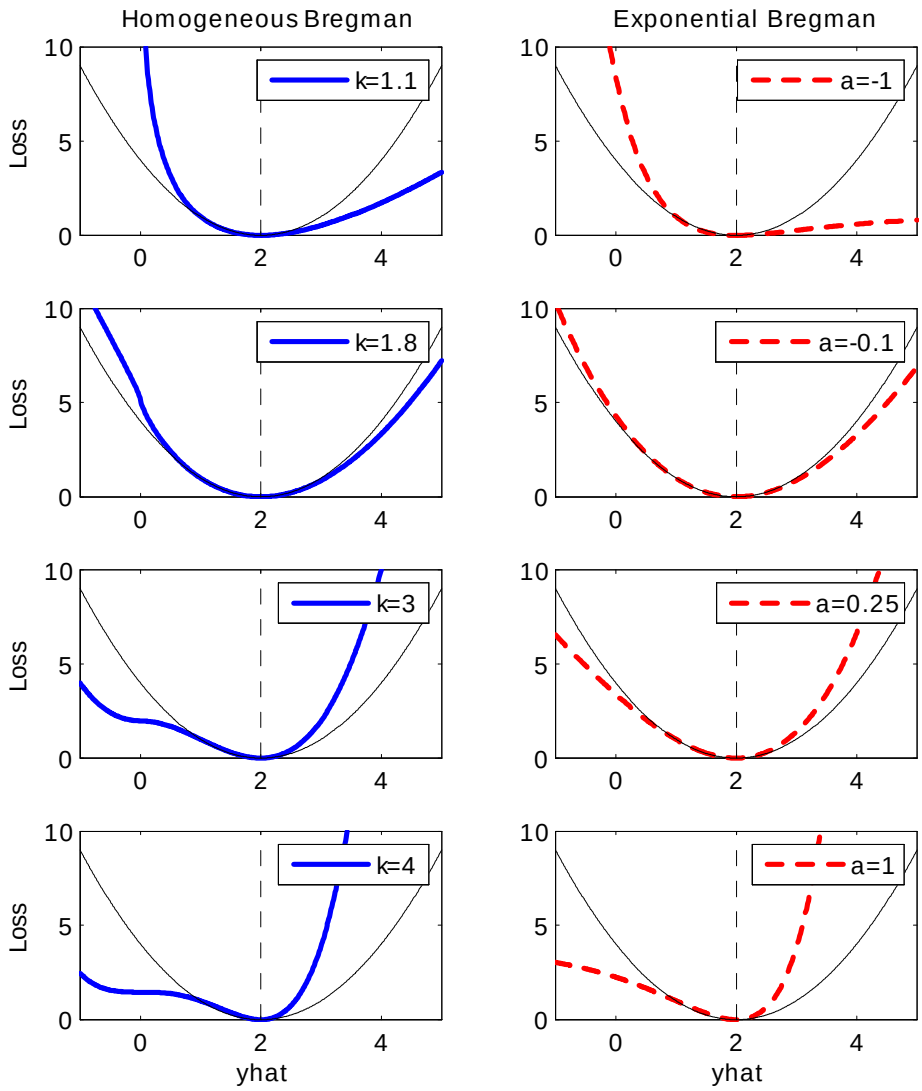

Figure 1: *Various Bregman loss functions. The left column presents four elements of the "homogeneous Bregman" family, and the right column presents four elements of the "exponential Bregman" family. The squared error loss function is also presented in each panel. In all cases the value for* $\hat{y}$ *ranges from -1 to 5, and the value of* y *is set at 2.*

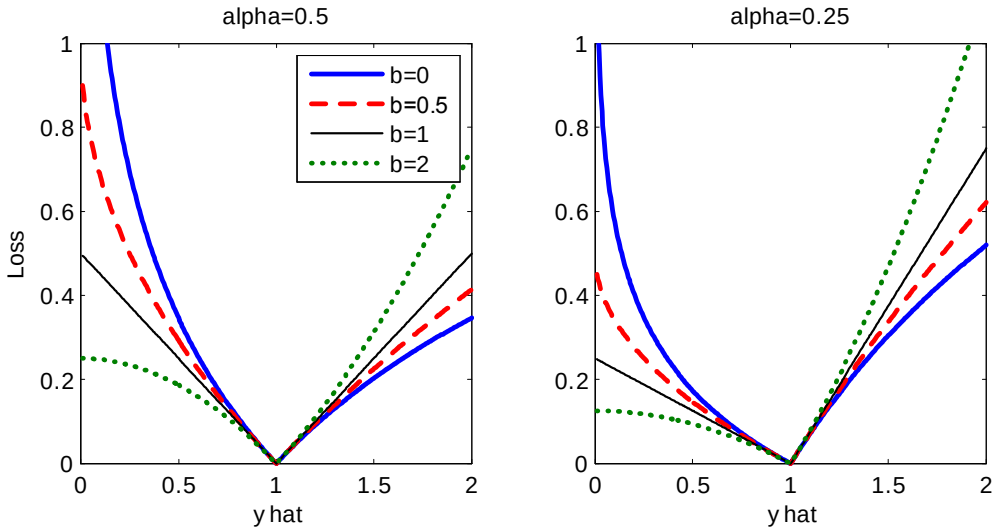

Figure 2: *Various homogenous GPL loss functions, with* $\alpha = 0.5$ *(left panel) and* $\alpha = 0.25$ *(right panel). The "Lin-Lin" (or "tick") loss function is obtained when* $b = 1$*. In both cases the value for* $\hat{y}$ *ranges from 0 to 2, and the value of* y *is set at 1.*

.

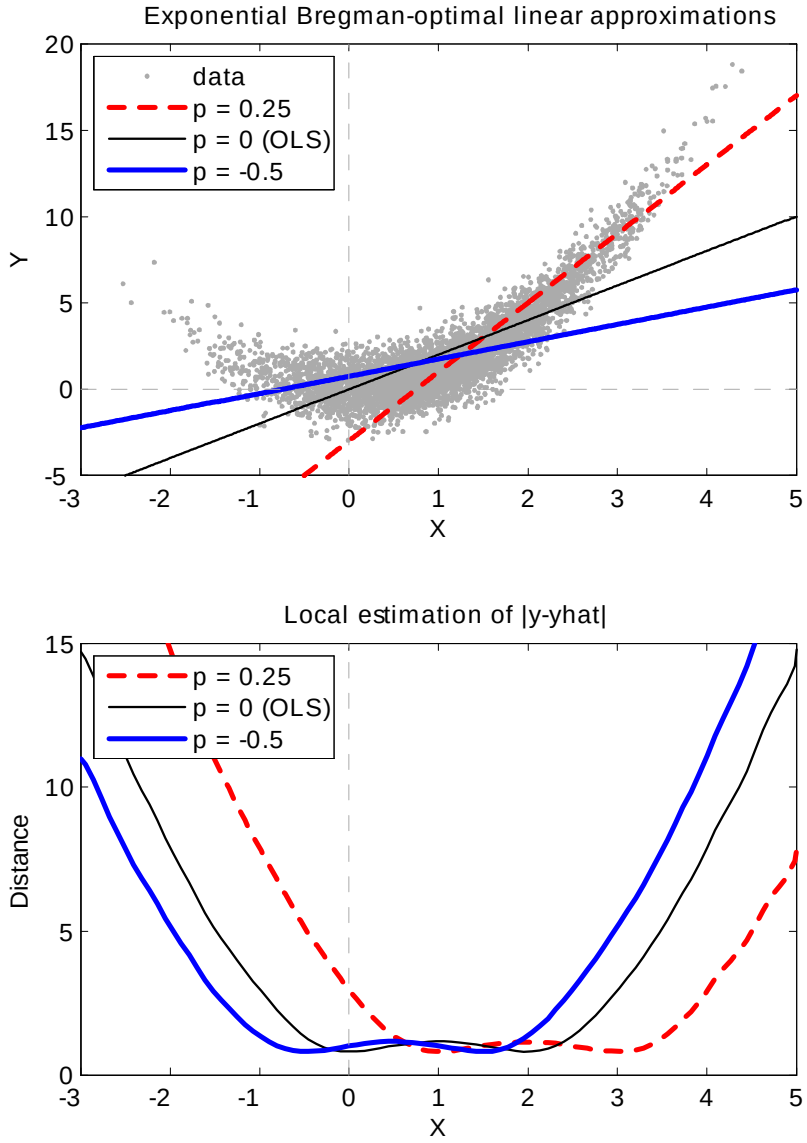

Figure 3: *The top panel presents the optimal linear approximations to a nonlinear DGP based on the exponential Bregman loss function for three choices of "shape" parameter; the choice p=0 corresponds to quadratic loss, and the fit is the same as that obtained by OLS. The lower panel presents a simple nonparametric estimate of the distance between the realization and the (linear) forecast as a function of the predictor variable (X), for the three loss function parameters.*

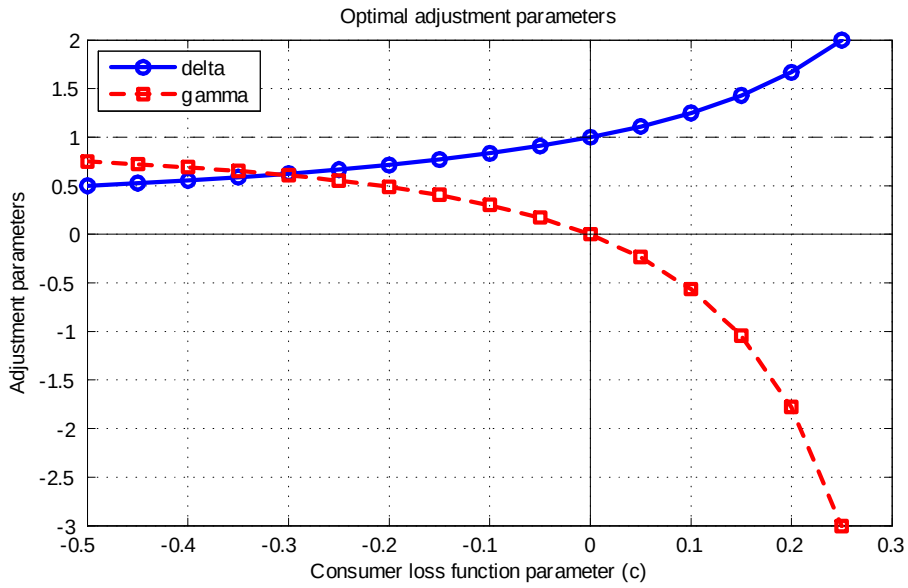

Figure 4: *Parameters of the optimal linear adjustment function, when the producer's loss function parameter is zero, for a range of values of the consumer's loss function parameter.*

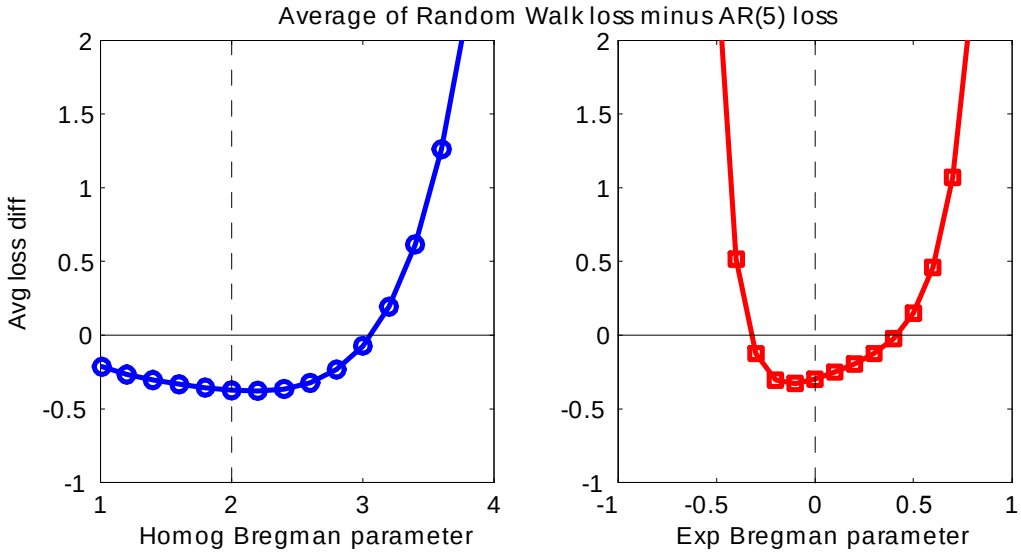

Figure 5: *Average loss from a random walk forecast minus that from an estimated AR(5) forecast, for various homogeneous (left panel) and exponential (right panel) Bregman loss functions.*

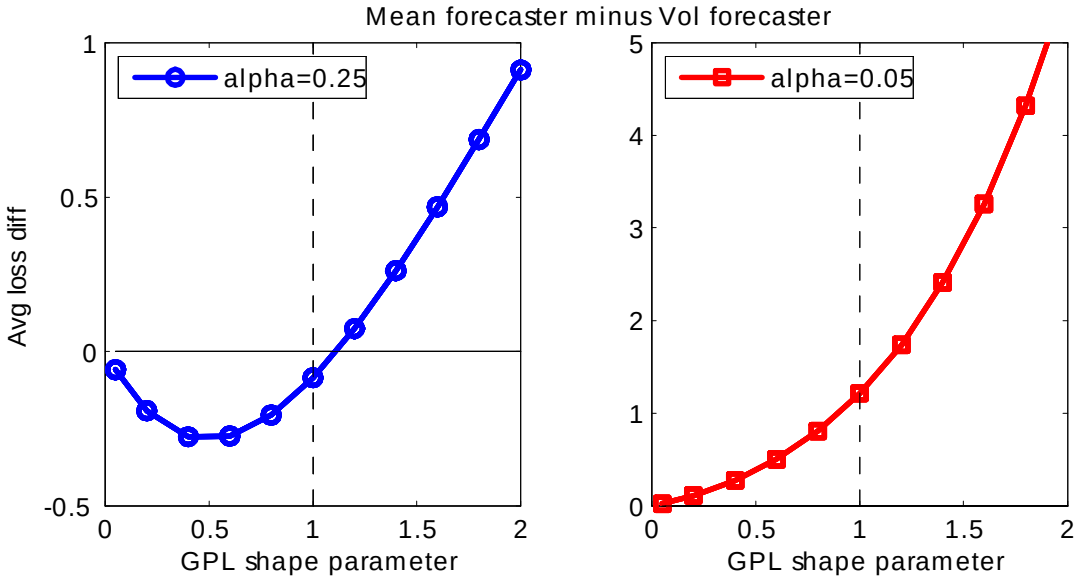

Figure 6: *Average loss from a AR-constant volatility forecsast minus that from a constant mean-GARCH forecast for various GPL loss functions. (Lin-Lin loss is marked with a vertical line at 1.) The left panel is for the 0.25 quantile, and the right panel is for the 0.05 quantile.*

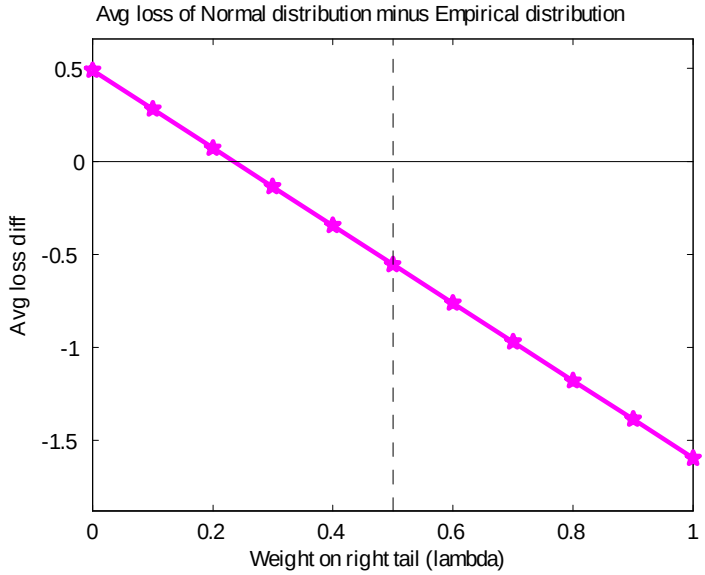

Figure 7: *Average wCRSP loss from a Normal distribution forecast minus that from an empirical distribution forecast based on 100 observations. The x-axis plots different weights on the left/right tail, with equal weight at 0.5, indicated with a vertical line.*

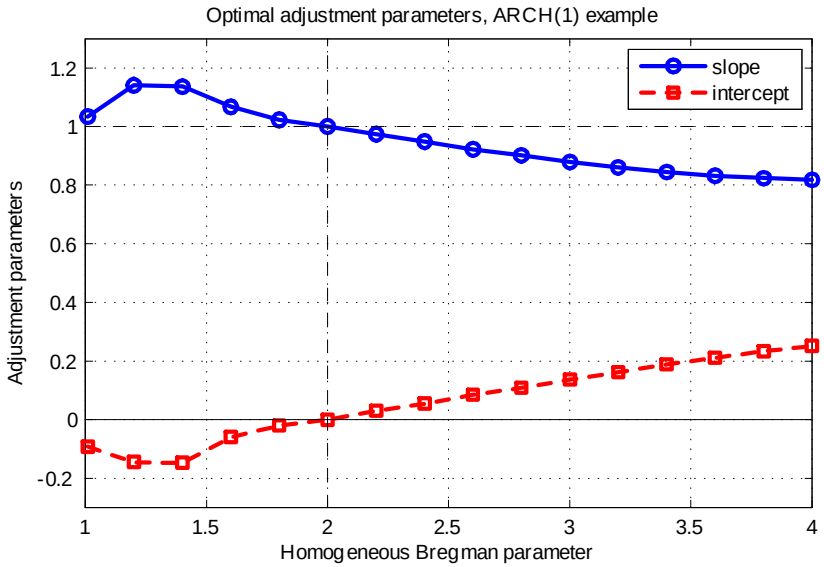

Figure 8: *Parameters of the optimal linear adjustment function for the ARCH(1) example, for a range of homogeneous Bregman loss functions.*

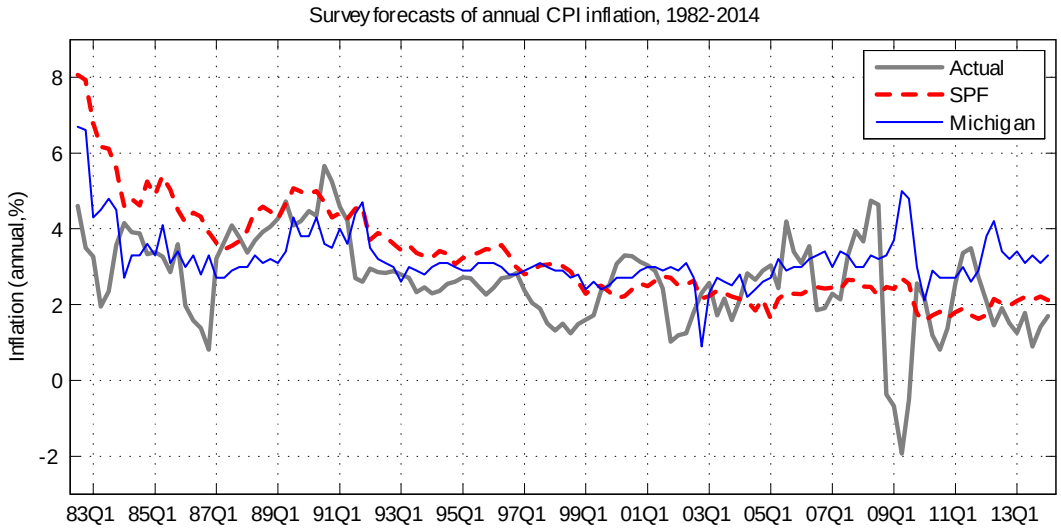

Figure 9: *Time series of actual and predicted annual US CPI inflation, updated quarterly, over the period 1982Q3–2014Q1. The inflation forecasts come from the Survey of Professional Forecasters and the Michigan survey of consumers.*

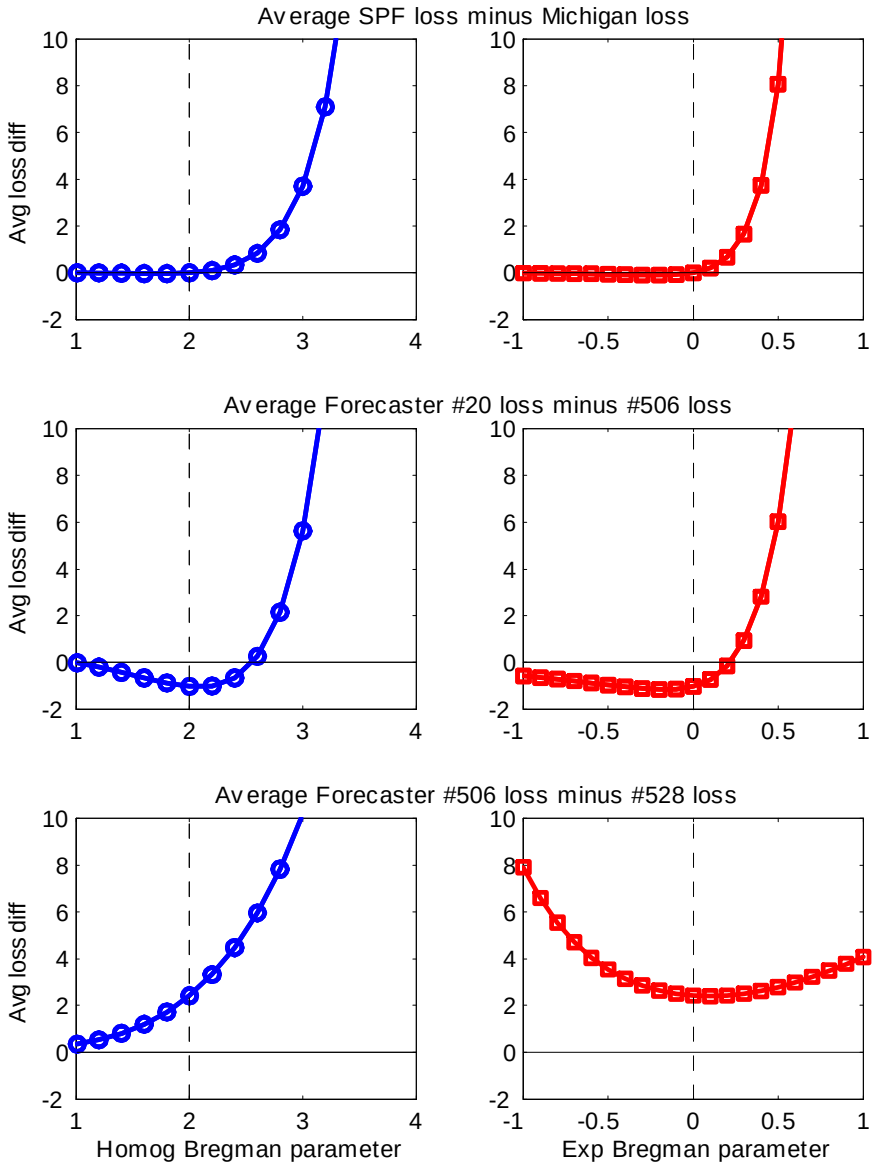

Figure 10: *Differences in average losses between two forecasts, for a range of loss function parameters. The "homogeneous Bregman" loss function is in the left column, and the "exponential Bregman" loss function is in the right column. The squared-error loss function is nested at 2 and 0 for these loss functions, and is indicated by a vertical line. The top row compares the consensus forecast from the Survey of Professional Forecasters and the Michigan survey of consumers; the lower two rows compare individual forecasters from the Survey of Professional Forecasters.*

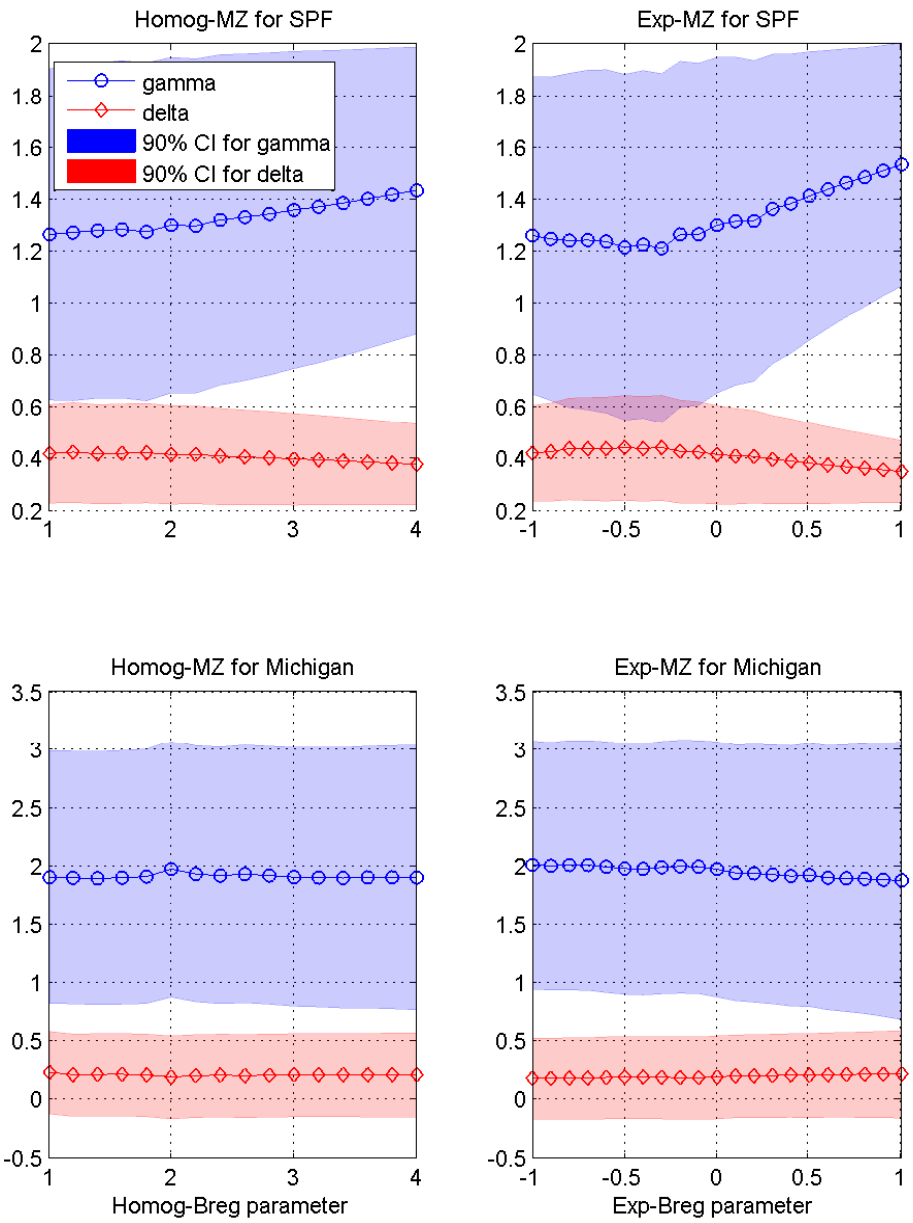

Figure 11: *Parameter estimates and 90% confidence intervals for Mincer-Zarnowitz tests of the optimality of SPF (upper row) and Michigan (lower row) forecasts, for a range of Homogeneous (left column) and Exponential (right column) Bregman loss functions.*