Comparing Possibly Misspecified Forecasts*

Andrew J. Patton
Duke University

First version: 27 September 2014. This version: 22 December 2016

Abstract

Recent work has emphasized the importance of evaluating forecasts of a statistical functional (such as a mean, quantile, or distribution) 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 free from estimation error, and if the information sets of the competing forecasters are nested, then the ranking induced by a single consistent loss function is sufficent for the ranking by any consistent loss function. This paper shows, via analytical results and realistic simulation-based analyses, that the presence of misspecified models, parameter estimation error, or nonnested information sets, leads generally 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, 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, economic forecasting, point forecasting, model misspecification, Bregman distance, proper scoring rules, consistent loss functions.

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

AMS 2010 Classifications: 62M20, 62P20.

*For helpful comments I thank Tim Bollerslev, Dean Croushore, Frank Diebold, Tilmann Gneiting, Jia Li, Robert Lieli, Minchul Shin, Allan Timmermann and seminar participants at Boston College, Columbia, Duke, Penn, Princeton, St. Louis Federal Reserve, 8th French Economics conference, NBER Summer Institute, Nordic Econometric Society meetings, and the World Congress of the Econometric Society. 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

Misspecified models pervade the observational sciences and social sciences. In such fields, researchers must contend with limited data, which inhibits both their ability to refine their models, thereby introducing the risk of model misspecification, and their ability to estimate these models with precision, introducing estimation error (parametric or nonparametric). This paper considers the implications of these empirical realities for the comparison of forecasts, in light of recent work in statistical decision theory on the importance of the use of consistent scoring rules or loss functions in forecast evaluation, see Gneiting (2011a). This paper shows that in analyses where forecasts are possibly based on models that are misspecified, subject to estimation error, or that use nonnested information sets (e.g., expert forecasters using different proprietary data sets), the choice of scoring rule or loss function is even more critical than previously noted.

Recent work in the theory of prediction has 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," in a specific sense clarified below, the quantity that the forecaster was asked to predict, for example the mean, the median, or the probability of a particular outcome (e.g., rain, a recession), 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 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, see Savage (1971), Banerjee *et al.* (2005) and Bregman (1967), and includes the squared-error loss function as a special case. The class of loss functions that is consistent for the α-quantile is known as the generalized piecewise linear (GPL) class, see Gneiting (2011b), which nests the familiar piece-wise linear function from quantile regression, see Koenker *et al.* (2017) for example. 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 the literature by refining this recommendation to reflect real-world deviations from the ideal predictive environment, and suggests that only the first part of the above 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 models that are correctly specified, free from estimation error, and when the information sets of one of the forecasters nests the other, 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 via analytical and realistic numerical examples that when *any* of these three conditions is violated, i.e. when the competing forecasts are based on nonnested information sets, or 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.

I illustrate the 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 generally similar in accuracy, with the SPF performing better under quadratic loss, 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.

The sensitivity or insensitivity of rankings to the choice of loss function also has implications for the use of multiple loss functions to compare a given collection of 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, estimation error or nonnested information sets, the results in this paper show that 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 judicious 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. In recent work, Ehm *et al.* (2016) obtain mixture represetations for the classes of loss functions consistent for quantiles and expectiles which can be used to determine whether one forecast outperforms another across all consistent loss functions.

This paper is related to several recent papers on related topics. Elliott *et al.* (2016) 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. Lieli and Stinchcombe (2013, 2016) study the identifiability of a forecaster's loss function given a sequence of observed forecasts, and find in particular for discrete random variables that whether the forecast is constrained to have the same support as the target variable or not has crucial implications for identification. 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 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. A primary goal of this paper is to show that in many realistic prediction environments, sensitivity to the choice of consistent loss function is the norm, not the exception.

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 realistic simulation designs that illustrate the main ideas of the paper, and Section 4 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:

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

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

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

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

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:

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

where $\phi : \mathcal{Y} \to \mathbb{R}$ 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 (Patton, 2011), which is applicable for strictly positive random variables:

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

The quadratic and QLIKE loss functions are unique (up to location and scale constants) 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 $k > 1$:

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

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, which we exploit in the next subsection. 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 Forecast comparison in ideal and less-than-ideal forecasting environments

As usual in the forecast comparison literature, I will consider ranking forecasts by their unconditional average loss, a quantity that is estimable, under standard regularity conditions, given a

sample of data. (Forecasts themselves, on the other hand, are of course generally based on conditioning information.) 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)

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

and we then seek to determine whether

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

subject to these expectations existing. The following proposition provides conditions under which the above implication holds. Denote the information sets of forecasters A and B as $\mathcal{F}_t^A$ and $\mathcal{F}_t^B$.

Assumption 1 *The information sets of the forecasters are nested, so $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A \; \forall t$ or $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B \; \forall t$.*

Assumption 2 *If the forecasts are based on models, then the models are free from estimation error.*

Assumption 3 *If the forecasts are based on models, then the models are correctly specified for the statistical functional of interest.*

The above assumptions are presented somewhat generally, as we will refer to them not only in this section on mean forecasting, but also for the analyses of quantile and distribution forecasting below. Assumption 3 implies, in this section, that:

$$\exists \; \theta_{0,i} \in \Theta \text{ s.t. } \mathbb{E}\left[Y|\mathcal{F}_t^i\right] = m_i\left(Z_t^i; \theta_{0,i}\right) \;\; a.s. \text{ for some } Z_t^i \in \mathcal{F}_t^i, \text{ for } i \in \{A, B\} \tag{10}$$

where m_i is forecaster i's prediction model, which has a finite-dimensional parameter θ. The "true" parameter $\theta_{0,i}$ is allowed to vary across i as the conditional mean of Y_t given $\mathcal{F}_t^i$ will generally vary with the information set, $\mathcal{F}_t^i$. Related, Assumption 2 implies in this section that

$$\hat{Y}_t^i = m_i\left(Z_t^i; \theta_i^*\right) \;\; a.s. \text{ for all } t = 1, 2, .. \tag{11}$$

where θ_i^* is some fixed parameter.

Proposition 1 *(a) Under Assmptions 1, 2 and 3, the ranking of two forecasts by MSE is sufficient for their ranking by any Bregman loss function.*

(b) If any of Assumptions 1, 2, or 3 fail to hold, then the ranking of two forecasts may be sensitive to the choice of Bregman loss function.

The proof of part (a) is given in Appendix A. Interest in this paper focuses on part (b), and we provide analytical examples for this part below.

Under the strong assumptions of comparing only forecasters with nested information sets, and who use only correctly specified models with no estimation error, part (a) shows that 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. Related to this result, Holzmann and Eulert (2014) show in a general framework that forecasts based on larger information sets lead generally to lower expected loss.

To verify part (b) of the above proposition, we consider deviations from the three assumptions used in part (a). Consider the following example: assume that the target variable follows a persistent, but strictly stationary AR(5) process:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + ... + \phi_5 Y_{t-5} + \varepsilon_t, \quad \varepsilon_t \sim iid\ N(0,1) \tag{12}$$

where $\phi_0 = 1$ and $[\phi_1, ...\phi_5] = [0.8, 0.3, -0.5, 0.2, 0.1]$. These parameter values are stylized, but are broadly compatible with estimates for standard macroeconomic time series like US interest rates, see Faust and Wright (2013). We then consider a set of forecasting models. The first three contain no estimation error, and have parameters that are correct given their information sets:

$$\text{AR(1)} \quad \hat{Y}_t = \beta_0 + \beta_1 Y_{t-1} \tag{13}$$

$$\text{AR(2)} \quad \hat{Y}_t = \delta_0 + \delta_1 Y_{t-1} + \delta_2 Y_{t-2} \tag{14}$$

$$\text{AR(5)} \quad \hat{Y}_t = \phi_0 + \phi_1 Y_{t-1} + ... + \phi_5 Y_{t-5} \tag{15}$$

The first two models use too few lags, while the third model nests the data generating process, and will produce the optimal forecast. The parameters of the AR(1) and AR(2) models are obtained

by minimizing the (population) expecation of the Bregman loss, and the specific values are given in Appendix B, along with details on the derivation of these parameter values. As each of these models are correctly specified given their (limited) information sets, part (a) of Proposition 3 below applies, and the optimal parameters are not affected by the specific Bregman loss function used in estimation.

In the upper-left panel of Figure 2 I plot the ratio of the expected loss for a given forecast to that for the optimal forecast, as a function of the Exponential-Bregman parameter a. We see in that panel that the rankings are as expected: the AR(1) model has higher average loss than then AR(2), which in turn has higher average loss than the AR(5). These rankings hold for all values of a, consistent with part (a) of Proposition 1. More generally, the ranking method of Ehm *et al.* (2016) could be applied, and would show that these rankings hold for any Bregman loss function, not only those in the Exponential Bregman family.

Now consider comparing two misspecified models. The first is a simple random walk forecast, and the second is a "two-period average" forecast:

$$\text{Random Walk} \quad \hat{Y}_t = Y_{t-1} \tag{16}$$

$$\text{Two-period Average} \quad \hat{Y}_t = \frac{1}{2}\left(Y_{t-1} + Y_{t-2}\right) \tag{17}$$

Neither of these forecasts has any estimation error and their information sets are nested, but both models are misspecified. The lower-left panel of Figure 2 compares the average losses for these two forecasts, and we observe that the Random Walk provides the better approximation when the Exponential Bregman loss function parameter is near zero, but the Two-period Average forecast is preferred when the parameter is further from zero.

Now we consider the impact of parameter estimation error. Consider the feasible versions of the AR(2) and AR(5) forecasts, with the parameters are estimated by OLS using a rolling window of 36 observations, corresponding to three years of monthly data:

$$\widehat{\text{AR}}(2) \quad \hat{Y}_t = \hat{\delta}_{0,t} + \hat{\delta}_{1,t} Y_{t-1} + \hat{\delta}_{2,t} Y_{t-2} \tag{18}$$

$$\widehat{\text{AR}}(5) \quad \hat{Y}_t = \hat{\phi}_{0,t} + \hat{\phi}_{1,t} Y_{t-1} + \ \ldots \ + \hat{\phi}_{5,t} Y_{t-5} \tag{19}$$

We compare $\widehat{\text{AR}}(2)$ and $\widehat{\text{AR}}(5)$ to see whether any trade-off exists between goodness of fit and estimation error: $\widehat{\text{AR}}(5)$ is correctly specified, but requires the estimation of three more parameters; $\widehat{\text{AR}}(2)$ excludes three useful lags, but is less affected by estimation error. Analytical results for the finite-sample estimation error in misspecified AR(p) models are not available, and so we use 10,000 simulated values to obtain the average losses for these two models. The results are presented in the upper-right panel of Figure 2. We see that the expected loss of $\widehat{\text{AR}}(5)$ is below that of $\widehat{\text{AR}}(2)$ for values of the Exponential-Bregman parameter near zero, while the ranking reverses when the parameter is greater than approximately 0.4 in absolute value. Thus, there is indeed a trade-off between goodness-of-fit and estimation error, and the ranking switches as the loss function parameter changes. This reversal of ranking is not possible in the "ideal environment" case.

Finally, we seek to show that relaxing only Assumption 1 (nested information sets) can lead to a sensitivity in the ranking of two forecasts. For reasons explained below, consider a different data generating process, where the target variable is affected by two Bernoulli shocks, X_t and W_t:

$$\begin{aligned} Y_t &= X_t\mu_L + (1 - X_t)\,\mu_H + W_t\mu_C + (1 - W_t)\,\mu_M + Z_t \\ \text{where } X_t &\sim \textit{iid Bernoulli}\,(p)\,, \quad W_t \sim \textit{iid Bernoulli}\,(q)\,, \quad Z_t \sim \textit{iid } N\,(0,1) \end{aligned} \tag{20}$$

Forecaster X has access to a "local variation" signal X_t that is regular $(p = 0.5)$ but not very strong $(\mu_L = -1,\ \mu_H = 1)$, while Forecaster W has access to a "crisis" signal W_t that is irregular $(q = 0.05)$ but large when it arrives $(\mu_C = -5,\ \mu_M = 0)$. If both forecasters optimally use their (non-overlapping) information sets, then their forecasts are:

$$\begin{aligned} \hat{Y}_t^X &= q\mu_C + (1 - q)\,\mu_M + \mu_H + (\mu_L - \mu_H)\,X_t \\ \hat{Y}_t^W &= p\mu_L + (1 - p)\,\mu_H + \mu_M + (\mu_C - \mu_M)\,W_t \end{aligned} \tag{21}$$

The lower-right panel of Figure 2 shows that the "crisis" forecaster is preferred for Exponential-Bregman parameter values less than zero, while the "local variation" forecaster is preferred for larger parameter values.

We have thus demonstrated that relaxing *any one* of the three "ideal environment" assumptions in part (a) of Proposition 1 can lead to sensitivity of forecast rankings to the choice of Bregman loss

function. Moreover, we have demonstrated this using scenarios that are broadly similar to those faced in economic and financial forecasting applications. In Section 3 we consider more realistic scenarios using numerical simulations.

It should be noted that it may be possible to partially relax Assumptions 1–3 in Proposition 1, or to place other restrictions on the problem, and retain (some, possibly partial) robustness of the ranking of forecasts to the choice of Bregman loss function. One example is when the competing forecasts are correct given their (possibly limited) information sets, free from estimation error, and the target variable and the forecasts are Normally distributed. In this case the following proposition shows we can omit the assumption of nested information sets and retain robustness of rankings for any Exponential-Bregman loss function. (This explains the need for an alternative DGP in demonstrating sensitivity to non-nested information sets.)

Proposition 2 *If (i)* $Y_t \sim N\left(\mu, \sigma^2\right)$, *(ii)* $\hat{Y}_t^i \sim N\left(\mu, \omega_i^2\right)$ *for* $i \in \{A, B\}$, *and (iii)* $\mathbb{E}\left[Y_t | \hat{Y}_t^i\right] = \hat{Y}_t^i$ *for* $i \in \{A, B\}$, *then*

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

Other special cases of robustness may be arise if, for example, the form of the model misspecification was known, or if the target variable has a particularly simple structure (e.g., a binary random variable, see Elliott *et al.* (2016) for example). I do not pursue further special cases here.

2.3 Optimal approximations from a possibly misspecified model

We next present a result for the estimation of forecasting models: consider the problem of calibrating a parametric forecasting model to generate the best prediction. If the model is correctly specified, then part (a) of the proposition below shows that minimizing the expected loss under any Bregman loss function will yield a consistent estimator of the model's parameters. We contrast this robust outcome with the sensitivity to the choice of loss function that arises under model misspecification in part (b). Elliott *et al.* (2016) 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.

Proposition 3 *Denote the model for* $\mathbb{E}\left[Y_t|\mathcal{F}_t\right]$ *as* $m\left(Z_t;\theta\right)$ *where* $Z_t \in \mathcal{F}_t$ *and* $\theta \in \Theta \subseteq \mathbb{R}^p,\ p < \infty$. *Define*

$$\theta_\phi^* \equiv \arg\min_{\theta\in\Theta}\ \mathbb{E}\left[L\left(Y_t, m\left(Z_t;\theta\right);\phi\right)\right] \tag{22}$$

where L *is a Bregman loss function characterized by the convex function* ϕ. *Assume (i)* $\partial m\left(Z_t;\theta\right)/\partial\theta \neq 0$ *a.s.* $\forall\theta \in \Theta$ *for both (a) and (b) below.*

(a) Assume (ii) $\exists\ \theta_0 \in \Theta$ *s.t.* $\mathbb{E}\left[Y_t|\mathcal{F}_t\right] = m\left(Z_t;\theta_0\right)$ *a.s., then* $\theta_\phi^* = \theta_0\ \forall\ \phi$.

(b) Assume (ii') ∄ $\theta_0 \in \Theta$ *s.t.* $\mathbb{E}\left[Y_t|\mathcal{F}_t\right] = m\left(Z_t;\theta_0\right)$ *a.s., then* θ_ϕ^* *may vary with* ϕ.

Assumption (i) in the above proposition is required for identification, imposing that the model is sensitive to changes in the parameter θ. Assumption (ii) is a standard definition of a correctly specified parametric model, while Assumption (ii') is a standard definition of a misspecified parametric model.

The proof of part (a) is presented in the Appendix. This result is related to the theory for pseudo/quasi maximum likelihood estimation, see Gourieroux, *et al.* (1984) and White (1994), for example.

To verify part (b) consider the following illustrative example, where the DGP is:

$$\begin{aligned} Y_t &= X_t^2 + \varepsilon_t,\ \ \varepsilon_t \perp\!\!\!\perp X_s\ \forall t,s \\ X_t &\sim iid\ N\left(\mu,\sigma^2\right),\quad \varepsilon_t \sim iid\ N\left(0,1\right) \end{aligned} \tag{23}$$

but the forecaster mistakenly assumes the predictor variable enters the model linearly:

$$Y_t = \alpha + \beta X_t + e_t \tag{24}$$

To obtain analytical results to illustrate the main ideas, consider a forecaster using the "Exponential Bregman" loss function, defined in equation (7), with parameter a. Using results for functions of Normal random variables (see Appendix B for details) we can analytically derive the optimal linear model parameters $[\alpha,\beta]$ as a function of a, subject to the condition that $a \neq \left(2\sigma^2\right)^{-1}$:

$$\hat{\alpha}_a^* = \sigma^2 - \frac{\mu^2}{\left(1-2a\sigma^2\right)^2}\quad ,\quad \hat{\beta}_a^* = \frac{2\mu}{1-2a\sigma^2} \tag{25}$$

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 $\left(\hat{\alpha}_a^*, \hat{\beta}_a^*\right) = \left(\sigma^2, 0\right)$ 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 $a = \left(2\sigma^2\right)^{-1}$. 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 loss function used 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 approximation yielded by OLS regression is obtained when $a = 0$, and by equation (25) we find in that case 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 ($a = -0.5$) the line flattens (the optimal parameters are 0.75 and 1) and 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 ($a = 0.25$, yielding optimal parameters -3 and 4).

[INSERT FIGURE 3 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. It is important to note, however, 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. This is related to work on estimation under the "relevant cost function," see Weiss (1996), Christoffersen and Jacobs

(2004), Skouras (2007), Hansen and Dumitrescu (2016) and Elliott *et al.* (2016) for applications in economics and finance. One interpretation of the results in this section is that if all estimators converge to the same quantity then there is no possible 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.

2.4 Comparing 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 class of loss functions that is necessary and sufficient for quantile forecasts is called the "generalized piecewise linear" (GPL) class, denoted $\mathcal{L}_{GPL}^{\alpha}$:

$$L\left(y,\hat{y};\alpha\right)=\left(\mathbf{1}\left\{y\leq\hat{y}\right\}-\alpha\right)\left(g\left(\hat{y}\right)-g\left(y\right)\right) \tag{26}$$

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:

$$L\left(y,\hat{y};\alpha\right)=\left(\mathbf{1}\left\{y\leq\hat{y}\right\}-\alpha\right)\left(\hat{y}-y\right) \tag{27}$$

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):

$$L\left(y,\hat{y};\alpha,b\right)=\ \left(\mathbf{1}\left\{y\leq\hat{y}\right\}-\alpha\right)\left(\operatorname{sgn}\left(\hat{y}\right)\left|\hat{y}\right|^{b}-\operatorname{sgn}\left(y\right)\left|y\right|^{b}\right)/b,\quad b>0 \tag{28}$$

Figure 4 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

$$\alpha=\mathbb{E}\left[\mathbf{1}\left\{Y_t\leq\hat{Y}_t^*\right\}|\mathcal{F}_t\right]\equiv F_t\left(\hat{Y}_t^*\right) \tag{29}$$

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). That is, whether

$$LinLin_A^\alpha \lesseqgtr LinLin_B^\alpha \Rightarrow \mathbb{E}\left[L\left(Y_t, \hat{Y}_t^A\right)\right] \lesseqgtr E\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \ \forall L \in \mathcal{L}_{GPL}^\alpha \tag{30}$$

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

Proposition 4 *(a) Under Assmptions 1, 2 and 3, the ranking of these two forecasts by expected Lin-Lin loss is sufficient for their ranking by any $\mathcal{L}_{GPL}^\alpha$ loss function.*

(b) If any of Assumptions 1, 2, or 3 fail to hold, then the ranking of these two forecasts may be sensitive to the choice of $\mathcal{L}_{GPL}^\alpha$ loss function.

As in the conditional mean case, a violation of any of Assumptions 1, 2, or 3 is sufficient to induce sensitivity to the choice of consistent loss function. A proof of part (a) is given in Appendix A, and analytical examples establishing part (b) are presented in the supplemental appendix. An example based on a realistic simulation design is given in Section 3 below.

2.5 Comparing 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

$$\mathbb{E}_F\left[L\left(F, Y\right)\right] \equiv \int L\left(F, y\right) dF\left(y\right) \leq \mathbb{E}_F\left[L\left(\tilde{F}, Y\right)\right] \tag{31}$$

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:

$$L(F,y) = \Psi(F) + \Psi^*(F,y) - \int \Psi^*(F,y)\, dF(y) \tag{32}$$

where Ψ is a convex, real-valued function, and Ψ^* is a subtangent of Ψ at $F \in \mathcal{P}$. I denote the set of proper scoring rules satisfying equation (32) 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):

$$wCRPS(F,y;\omega) = \int_{-\infty}^{\infty} \omega(z)\left(F(z) - \mathbf{1}\{y \leq z\}\right)^2 dz \tag{33}$$

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.

$$\mathbb{E}\left[L_i\left(F_t^A, Y_t\right)\right] \lesseqgtr \mathbb{E}\left[L_i\left(F_t^B, Y_t\right)\right] \Rightarrow \mathbb{E}\left[L_j\left(F_t^A, Y_t\right)\right] \lesseqgtr \mathbb{E}\left[L_j\left(F_t^B, Y_t\right)\right] \;\; \forall L_j \in \mathcal{L}_{\text{Proper}} \tag{34}$$

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

Proposition 5 *(a) Under Assmptions 1, 2 and 3, the ranking of these two forecasts by any given proper scoring rule is sufficient for their ranking by any other proper scoring rule.*

(b) If any of Assumptions 1, 2, or 3 fail to hold, then the ranking of these two forecasts may be sensitive to the choice of proper scoring rule.

As in the conditional mean and quantile cases, a violation of any of Assumptions 1, 2 or 3 is enough to induce sensitivity to the choice of proper scoring rule. A proof of part (a) is given in Appendix A, and analytical examples establishing part (b) are presented in the supplemental appendix. An example based on a realistic simulation design is given in Section 3 below.

3 Simulation-based results for realistic scenarios

In this section I consider three realistic forecasting scenarios, all calibrated to standard economic applications, and show that in the presence of model misspecification, estimation error or nonnested

information sets can 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, similar to US inflation or long-term bond yields:

$$Y_t = Y_{t-1} - 0.02Y_{t-2} - 0.02Y_{t-3} - 0.01Y_{t-4} - 0.01Y_{t-5} + \varepsilon_t, \;\; \varepsilon_t \sim iid\ N(0,1) \tag{35}$$

As forecasts, consider the comparison of a parsimonious but misspecified model with a correctly-specified model that is subject to estimation error. The first forecast is based on a random walk assumption, and the second forecast is based on a correctly-specified AR(5) model with estimated parameters:

$$\hat{Y}_t^A = Y_{t-1} \tag{36}$$

$$\hat{Y}_t^B = \hat{\phi}_{0,t} + \hat{\phi}_{1,t}Y_{t-1} + \hat{\phi}_{2,t}Y_{t-2} + \hat{\phi}_{3,t}Y_{t-3} + \hat{\phi}_{4,t}Y_{t-4} + \hat{\phi}_{5,t}Y_{t-5} \tag{37}$$

where $\hat{\phi}_{j,t}$ is the OLS estimate of ϕ_j based on data from $t-100$ to $t-1$, for $j = 0, 1, ..., 5$. 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 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]

Next, consider quantile forecasts for a heteroskedastic time series process, designed to mimic daily stock returns. Such data often have some weak first-order autocorrelation, and time-varying

volatility that is well-modeled using a GARCH (Bollerslev, 1986) process:

$$\begin{aligned} Y_t &= \mu_t + \sigma_t \varepsilon_t, \;\; \varepsilon_t \sim iid\ N(0,1) \\ \text{where } \mu_t &= 0.03 + 0.05 Y_{t-1} \\ \sigma_t^2 &= 0.05 + 0.9\sigma_{t-1}^2 + 0.05\sigma_{t-1}^2 \varepsilon_{t-1}^2 \end{aligned} \tag{38}$$

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:

$$\begin{aligned} \hat{Y}_t^A &= \mu_t + \bar{\sigma}\Phi^{-1}(\alpha) \\ \hat{Y}_t^B &= \bar{\mu} + \sigma_t \Phi^{-1}(\alpha) \end{aligned} \tag{39}$$

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 an intermediate quantile between the tail and the center of the distribution ($\alpha = 0.25$). I compare these forecasts using the family of homogeneous GPL loss functions in equation (28), 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 an intermediate quantile, $\alpha = 0.25$, 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]

As a final 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, with parameters broadly designed to match daily US stock returns:

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t, \quad \varepsilon_t \sim iid\ Skew\ t(0,1,6,-0.25) \\ \sigma_t^2 &= 0.05 + 0.9\sigma_{t-1}^2 + 0.05\sigma_{t-1}^2 \varepsilon_{t-1}^2 \end{aligned} \tag{40}$$

I take the first distribution forecast to be based on the Normal distribution, with mean zero and variance estimated using the past 100 observations. 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, which is clearly more flexible than the first, but will inevitably contain more estimation error.

$$\hat{F}_{A,t}(x) = \Phi\left(\frac{x}{\hat{\sigma}_t}\right), \text{ where } \hat{\sigma}_t^2 = \frac{1}{100}\sum\nolimits_{j=1}^{100} Y_{t-j}^2 \tag{41}$$

$$\hat{F}_{B,t}(x) = \frac{1}{100}\sum_{j=1}^{100} \mathbf{1}\left\{Y_{t-j} \leq x\right\} \tag{42}$$

I consider the weighted CRPS scoring rule from equation (33) where the weights are based on the standard Normal CDF:

$$\omega(z;\lambda) \equiv \lambda\Phi(z) + (1-\lambda)(1-\Phi(z)), \quad \lambda \in [0,1] \tag{43}$$

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]

These three illustrations all use data generating processes that a realistic representations of prediction problems in economics and finance, and in all cases we observe sensitivity of forecast rankings to the choice of consistent scoring rule. Thus the sensitivity of rankings highlighted in this paper is not only of theoretical interest; it may arise in many realistic forecasting applications.

4 Empirical illustration: Evaluating forecasts of US inflation

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/Yt4y81`). 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, and I compute the implied one-year forecast for the SPF using the one-quarter forecasts for horizons 1 to 4. The sample period is 1982Q3 to 2016Q2, a total of 136 observations. As the "actual" series I use the 2016Q4 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 8, and summary statistics are presented in Table 1.

[INSERT FIGURE 8 AND TABLE 1 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 510, as they all have relatively long histories of responses. Like the consensus forecasts, I also consider the one-year forecasts from the individual respondents.

Given the difficulty in capturing the dynamics of inflation, it is reasonable to expect that all forecasters are subject to model misspecification. Further, only relatively few observations are available for forecasters to estimate their model, making estimation error relevant feature of the problem. 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 all potentially relevant here.

Figure 9 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 lower average loss for values of the Bregman parameter less than 3.5 and 0.5 in the homogeneous and exponential cases respectively, while the reverse holds true for parameters above these values. (The difference in average loss is slightly below 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 sensitivity to the choice of loss function: for loss functions that penalize under-prediction more than over-prediction (homogeneous Bregman with parameter less than 2.25, 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 510 for all homogeneous and exponential Bregman loss functions considered; in no case does the ranking reverse. This is a comparison where the ranking method of Ehm *et al.* (2016) may reveal that there is a clear ordering of these two forecasters for all Bregman loss functions.

[INSERT FIGURE 9 ABOUT HERE]

5 Conclusion

Using analytical results, realistic simulation designs, and an application to US inflation forecasting, this paper shows that the ranking of competing forecasts can be sensitive to the choice of consistent

loss function or scoring rule. In the absence of model misspecification, parameter estimation error and nonnested forecaster information sets, this sensitivity is shown to vanish, but in almost all practical applications at least one of these complications may be a concern. In the presence of these complications, a conclusion of this paper is that declaring the target functional is not generally sufficient to elicit a forecaster's best (according to a given, consistent, loss function) forecast; rather 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.

Appendix A: Proofs

Proof of Proposition 1(a). We will show that under Assumptions (1)–(3), $MSE_B \geq MSE_A \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t \Rightarrow \mathbb{E}\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \geq E\left[L\left(Y_t, \hat{Y}_t^A\right)\right] \ \forall L \in \mathcal{L}_{Bregman}$.

For the first implication: Assume that $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B \ \forall t$. This means that $\min_{\hat{y}} \ \mathbb{E}\left[(Y_t - \hat{y})^2 \,|\mathcal{F}_t^A\right] \geq \min_{\hat{y}} \ \mathbb{E}\left[(Y_t - \hat{y})^2 \,|\mathcal{F}_t^B\right] \ a.s. \ \forall t$, and so $\mathbb{E}\left[\left(Y_t - \hat{Y}_t^A\right)^2 |\mathcal{F}_t^A\right] \geq \mathbb{E}\left[\left(Y_t - \hat{Y}_t^B\right)^2 |\mathcal{F}_t^B\right] \ a.s. \ \forall t$ by Assumptions (2)-(3), and $\mathbb{E}\left[\left(Y_t - \hat{Y}_t^A\right)^2\right] \geq \mathbb{E}\left[\left(Y_t - \hat{Y}_t^B\right)^2\right]$ by the law of iterated expectations (LIE), which contradicts $MSE_B \geq MSE_A$. Thus we have, under Assumption (1), $MSE_B \geq MSE_A \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t$.

Now consider the second implication: Let

$$Y_t = \hat{Y}_t^A + \eta_t = \hat{Y}_t^B + \eta_t + \varepsilon_t \tag{44}$$

Then

$$\begin{aligned} \mathbb{E}\left[L\left(Y_t, \hat{Y}_t^A\right) - L\left(Y_t, \hat{Y}_t^B\right)\right] &= \mathbb{E}\left[-\phi\left(\hat{Y}_t^A\right) - \phi'\left(\hat{Y}_t^A\right)\eta_t + \phi\left(\hat{Y}_t^B\right) + \phi'\left(\hat{Y}_t^B\right)(\eta_t + \varepsilon_t)\right] \\ &= \mathbb{E}\left[\phi\left(\hat{Y}_t^B\right) - \phi\left(\hat{Y}_t^A\right)\right] \end{aligned} \tag{45}$$

since $\mathbb{E}\left[\phi'\left(\hat{Y}_t^A\right)\eta_t\right] = \mathbb{E}\left[\phi'\left(\hat{Y}_t^A\right)\mathbb{E}\left[\eta_t|\mathcal{F}_t^A\right]\right]$ by the LIE and $\mathbb{E}\left[\eta_t|\mathcal{F}_t^A\right] = 0$, by Assumptions (2)-(3). Similarly for $\mathbb{E}\left[\phi'\left(\hat{Y}_t^B\right)(\eta_t + \varepsilon_t)\right]$. Next, consider the second-order mean-value expansion:

$$\phi\left(\hat{Y}_t^A\right) = \phi\left(\hat{Y}_t^B\right) - \phi'\left(\hat{Y}_t^B\right)\varepsilon_t + \phi''\left(\ddot{Y}_t^A\right)\varepsilon_t^2 \tag{46}$$

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

$$\mathbb{E}\left[L\left(Y_t, \hat{Y}_t^A\right) - L\left(Y_t, \hat{Y}_t^B\right)\right] = \mathbb{E}\left[\phi'\left(\hat{Y}_t^B\right)\varepsilon_t\right] - \mathbb{E}\left[\phi''\left(\ddot{Y}_t^A\right)\varepsilon_t^2\right] \leq 0 \tag{47}$$

since $\mathbb{E}\left[\phi'\left(\hat{Y}_t^B\right)\varepsilon_t\right] = 0$ and ϕ is convex. And so $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A \, \forall t \Rightarrow \mathbb{E}\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \geq E\left[L\left(Y_t, \hat{Y}_t^A\right)\right] \; \forall L \in \mathcal{L}_{Bregman}$. ■

Proof of Proposition 2. First we note that

$$\begin{aligned}
\mathbb{E}\left[L\left(Y_t, \hat{Y}_t; a\right)\right] &= \frac{2}{a^2}\left(\mathbb{E}\left[\exp\{aY_t\}\right] - \mathbb{E}\left[\exp\left\{a\hat{Y}_t\right\}\right]\right) \\
&= \frac{2}{a^2}\left(\exp\left\{\frac{a}{2}\left(a\sigma^2 + 2\mu\right)\right\} - \exp\left\{\frac{a}{2}\left(a\omega_k^2 + 2\mu\right)\right\}\right) \\
&\to \sigma^2 - \omega_k^2 \quad \text{as } a \to 0.
\end{aligned}$$

where the first equality holds under mean-unbiasedness (assumption (ii)) and the second follows from normality of the target variable and the forecast (assumption (i)). The last line implies that whichever forecast is based on the richest information set, leading to the greatest (optimal) variability in the forecast (ω_k^2), will have the lowest MSE loss.

Then note that for non-MSE Exponential-Bregman loss (i.e., for $a \neq 0$), that if $\mathbb{E}\left[L\left(Y_t, \hat{Y}_t^A; a\right)\right] \geq \mathbb{E}\left[L\left(Y_t, \hat{Y}_t^B; a\right)\right]$, then $\exp\left\{\frac{a}{2}\left(a\omega_A^2 + 2\mu\right)\right\} \leq \exp\left\{\frac{a}{2}\left(a\omega_B^2 + 2\mu\right)\right\}$ and so $\omega_A^2 \leq \omega_B^2$ and thus $MSE_A \geq MSE_B$. The converse holds using the same derivations, proving the proposition. ■

Proof of Proposition 3(a). The first-order condition for the optimization is:

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

where the last equality holds by the LIE. Note that the first-order condition is satisfied 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). ■

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 4(a). We will show that under Assumptions (1)–(3), $LinLin_B^\alpha \geq LinLin_A^\alpha \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t \Rightarrow \mathbb{E}\left[L\left(Y_t, \hat{Y}_t^B\right)\right] \geq \mathbb{E}\left[L\left(Y_t, \hat{Y}_t^A\right)\right] \ \forall L \in \mathcal{L}_{GPL}^\alpha$, where $LinLin_j^\alpha \equiv \mathbb{E}\left[LinLin\left(Y_t, \hat{Y}_t^j\right)\right]$ for $j \in \{A, B\}$ and $LinLin$ is the "Lin-Lin" loss function in equation (27).

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 $\min_{\hat{y}} \mathbb{E}\left[LinLin\left(Y_t, \hat{y}\right)|\mathcal{F}_t^A\right] \geq \min_{\hat{y}} \mathbb{E}\left[LinLin\left(Y_t, \hat{y}\right)|\mathcal{F}_t^B\right]$ $a.s.\ \forall t$, and so $\mathbb{E}\left[LinLin\left(Y_t, \hat{Y}_t^A\right)|\mathcal{F}_t^A\right] \geq \mathbb{E}\left[LinLin\left(Y_t, \hat{Y}_t^B\right)|\mathcal{F}_t^B\right]$ $a.s.\ \forall t$ by Assumptions (2)-(3), and $\mathbb{E}\left[LinLin\left(Y_t, \hat{Y}_t^A\right)\right] \geq \mathbb{E}\left[LinLin\left(Y_t, \hat{Y}_t^B\right)\right]$ by the LIE, which is a contradiction. Thus, under Assumption (1), $LinLin_B^\alpha \geq LinLin_A^\alpha \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t$. Next: Let

$$\bar{L}^j \equiv \mathbb{E}\left[L_{GPL}\left(Y_t, \hat{Y}_t^j; \alpha, g\right)\right], \ \ j \in \{A, B\}$$

where $L_{GPL}\left(\cdot, \cdot; \alpha, g\right)$ is a GPL loss function defined by g, a nondecreasing function. Under Assumptions (2)-(3) we know that $\hat{Y}_t^j$ is the solution to $\min_{\hat{y}} \mathbb{E}\left[L_{GPL}\left(Y_t, \hat{Y}_t^j; \alpha, g\right)|\mathcal{F}_t^j\right]$. It is straightforward to show that $\hat{Y}_t^j$ then satisfies $\alpha = \mathbb{E}\left[\mathbf{1}\left\{Y_t \leq \hat{Y}_t^j\right\}|\mathcal{F}_t^j\right]$. This holds for all possible (conditional) distributions of Y_t, and from Saerens (2000) and Gneiting (2011b) we know that this implies (by the necessity of GPL loss for optimal quantile forecasts) that $\hat{Y}_t^j$ then moreover satisfies

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

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

Proof of Proposition 5(a). We again prove this result by showing that $\mathbb{E}\left[L\left(F_t^A, Y_t\right)\right] \leq \mathbb{E}\left[L\left(F_t^B, Y_t\right)\right]$ for some $L \in \mathcal{L}_{\text{Proper}} \Rightarrow \mathcal{F}_t^B \subseteq \mathcal{F}_t^A \ \forall t \Rightarrow \mathbb{E}\left[L\left(F_t^A, Y_t\right)\right] \leq \mathbb{E}\left[L\left(F_t^B, Y_t\right)\right] \ \forall \ L \in \mathcal{L}_{\text{Proper}}$. First: we are given that $\mathbb{E}\left[L\left(F_t^A, Y_t\right)\right] \leq \mathbb{E}\left[L\left(F_t^B, Y_t\right)\right]$, and assume that $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B \ \forall t$. Under Assumptions (2)-(3), this implies that we can take F_t^B as the data generating process for Y_t. Then $\mathbb{E}\left[L\left(F_t^B, Y_t\right)|\mathcal{F}_t^B\right] = \mathbb{E}_{F_t^B}\left[L\left(F_t^B, Y_t\right)|\mathcal{F}_t^B\right] \leq \mathbb{E}_{F_t^B}\left[L\left(F_t^A, Y_t\right)|\mathcal{F}_t^B\right] \ \forall t$ by Assumptions (2)-(3) and the propriety of L. By the LIE this implies $\mathbb{E}\left[L\left(F_t^B, Y_t\right)\right] \leq \mathbb{E}\left[L\left(F_t^A, Y_t\right)\right]$ which is a contradiction. Thus, under Assumption (1), $\mathbb{E}\left[L\left(F_t^A, Y_t\right)\right] \leq \mathbb{E}\left[L\left(F_t^B, Y_t\right)\right]$ 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 $\mathbb{E}\left[L\left(F_t^A, Y_t\right)\right] \leq \mathbb{E}\left[L\left(F_t^B, Y_t\right)\right]$ for any $L \in \mathcal{L}_{\text{Proper}}$, completing the proof. ■

Appendix B: Derivations

The results below draw on the following lemma, which summarizes some useful results on moments that arise when the data are Gaussian and the loss function is Exponential-Bregman.

Lemma 1 *If* $X \sim N\left(\mu, \sigma^2\right)$ *and* $(a, b) \in \mathbb{R}^2$*, then*

$$\begin{aligned}
&(i) & \mathbb{E}\left[\exp\left\{a + bX\right\}\right] &= \exp\left\{a + b\mu + \tfrac{1}{2}b^2\sigma^2\right\} \\
&(ii) & \mathbb{E}\left[\exp\left\{a + bX\right\} X\right] &= \exp\left\{a + b\mu + \tfrac{1}{2}b^2\sigma^2\right\}\left(\mu + b\sigma^2\right) \\
&(iii) & \mathbb{E}\left[\exp\left\{a + bX\right\} X^2\right] &= \exp\left\{a + b\mu + \tfrac{1}{2}b^2\sigma^2\right\}\left(\sigma^2 + \left(\mu + b\sigma^2\right)^2\right) \\
&(iv) & \mathbb{E}\left[\exp\left\{a + bX\right\} X^3\right] &= \exp\left\{a + b\mu + \tfrac{1}{2}b^2\sigma^2\right\}\left(\mu + b\sigma^2\right)\left(3\sigma^2 + \left(\mu + b\sigma^2\right)^2\right)
\end{aligned}$$

Some results below are simplified if we consider the following definition:

Definition 1 *A forecast* $\hat{Y}_t^i$ *is "mean unbiased" if* $\hat{Y}_t^i = \mathbb{E}\left[Y_t | \mathcal{F}_t^i\right]$ *a.s.*

By the law of iterated expectations, this implies that $\mathbb{E}\left[Y_t | \hat{Y}_t^i\right] = \hat{Y}_t^i$ *a.s.* Note that this does *not* require that $\mathcal{F}_t^i$ contains all relevant information for forecasting Y_t, only that that $\hat{Y}_t^i$ optimally uses all information available in $\mathcal{F}_t^i$.

Appendix B-1: Derivations for the AR(p) models in Section 2.2

The Gaussian AR(5) specification in equation (12) implies:

$$\begin{bmatrix} Y_t & Y_{t-1} & Y_{t-2} & Y_{t-3} & Y_{t-4} \end{bmatrix}' \sim N\left(\mu \boldsymbol{\iota}_5, \Sigma\right) \tag{48}$$

where $\boldsymbol{\iota}_5$ is a (5×1) vector of ones, μ is the mean of Y_t and Σ is the covariance matrix of the left-hand side vector. These can be obtained using standard methods from time series analysis, see

Hamilton (1994) for example).

$$\begin{aligned}
\text{Let } \boldsymbol{\phi} &\equiv [\phi_1, \phi_2, ..., \phi_5]' \\
\text{then } \mu &= \frac{\phi_0}{1 - \boldsymbol{\phi}'\boldsymbol{\iota}_5}, \quad \text{and } \; vec(\Sigma) = (I_5 - (F \otimes F))^{-1} vec(Q) \\
\text{where } F &= \left[\begin{array}{c|c} \boldsymbol{\phi}' & \\ \hline I_4 & \mathbf{0}_{4\times 1} \end{array}\right], \quad Q = \mathbf{e}_1\mathbf{e}_1', \quad \mathbf{e}_1 \equiv [1,0,0,0,0]'
\end{aligned} \tag{49}$$

Then note that the joint distribution of (Y_t, Y_{t-1}) is

$$\begin{aligned}
[Y_t, Y_{t-1}]' &\sim N(\mu\boldsymbol{\iota}_2, \Sigma_2) \\
\text{where } \Sigma_2 &\equiv \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix}, \quad \text{and } \; \gamma_j \equiv Cov[Y_t, Y_{t-j}], \quad \text{for } j = 0, 1, 2, ...
\end{aligned} \tag{50}$$

Denote $\rho_j \equiv \gamma_j/\gamma_0$. Then the conditional distribution of $Y_t|Y_{t-1}$ is:

$$Y_t|Y_{t-1} \sim N\left(\mu(1-\rho_1) + \rho_1 Y_{t-1}, \gamma_0\left(1-\rho_1^2\right)\right) \tag{51}$$

and so for the parameters used in the example we find $[\beta_0, \beta_1] = [\mu(1-\rho_1), \rho_1] = [1.52, 0.85]$.

Similar calculations for the AR(2) model yield:

$$[Y_t, Y_{t-1}, Y_{t-2}]' \sim N(\mu\boldsymbol{\iota}_3, \Sigma_3) \tag{52}$$

$$\text{where } \Sigma_3 \equiv \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix} \tag{53}$$

$$\begin{aligned}
Y_t|(Y_{t-1}, Y_{t-2}) &\sim N(\delta_0 + \delta_1 Y_{t-1} + \delta_2 Y_{t-2}, V_{AR2}) \\
\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} &= \frac{1}{\gamma_0^2 - \gamma_1^2}\begin{bmatrix} \gamma_1(\gamma_0 - \gamma_2) \\ \gamma_0\gamma_2 - \gamma_1^2 \end{bmatrix}, \quad \delta_0 = \mu(1 - \delta_1 - \delta_2) \\
V_{AR2} &= \gamma_0 - \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}\begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix}^{-1}\begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \gamma_0\frac{\rho_1^2 + \rho_2^2 - 2\rho_1^2\rho_2}{1-\rho_1^2}
\end{aligned} \tag{54}$$

And for the parameters used in the example we find $[\delta_0, \delta_1, \delta_2] = [1.38, 0.76, 0.10]$.

Now we derive the expected loss for the AR(1), AR(2) and AR(5) forecasts. First note that since all three of these forecasts are mean-unbiased, the expected loss of any Bregman loss function

simplifies to:

$$\mathbb{E}\left[L\left(Y_t,\hat{Y}_t;\phi\right)\right]=\mathbb{E}\left[\phi\left(Y_t\right)\right]-\mathbb{E}\left[\phi\left(\hat{Y}_t\right)\right]-\mathbb{E}\left[\phi'\left(\hat{Y}_t\right)\left(\mathbb{E}\left[Y_t|\hat{Y}_t\right]-\hat{Y}_t\right)\right]=\mathbb{E}\left[\phi\left(Y_t\right)\right]-\mathbb{E}\left[\phi\left(\hat{Y}_t\right)\right]. \tag{55}$$

For Exponential-Bregman loss, where $\phi\left(Y;a\right)=2a^{-2}\exp\left\{aY\right\}$, Lemma 1 implies

$$\mathbb{E}\left[\phi\left(Y_t\right)\right]=2a^{-2}\mathbb{E}\left[\exp\left\{aY_t\right\}\right]=2a^{-2}\exp\left\{\frac{a}{2}\left(2\mu+a\gamma_0\right)\right\} \tag{56}$$

To obtain $\mathbb{E}\left[\exp\left\{a\hat{Y}_t\right\}\right]$ for the AR(1), AR(2) and AR(5) forecasts, we exploit the fact that for this Gaussian autoregression, all of these forecasts are unconditionally normally distributed: $\hat{Y}_t^{\mathrm{ARk}}\sim N\left(\mu,V_{\mathrm{ARk}}\right)$, where $\left[V_{\mathrm{AR1}},V_{\mathrm{AR2}},V_{\mathrm{AR5}}\right]=\left[\rho_1^2\gamma_0\ ,\ \delta_1^2\delta_0+\delta_2^2\delta_0+2\delta_1^2\delta_2\ ,\ \gamma_0-1\right]$. Thus we find

$$\begin{aligned}\mathbb{E}\left[\phi\left(\hat{Y}_t^{\mathrm{ARk}}\right)\right] &= 2a^{-2}\mathbb{E}\left[\exp\left\{a\hat{Y}_t^{\mathrm{ARk}}\right\}\right]\\ &= 2a^{-2}\exp\left\{\frac{a}{2}\left(2\mu+aV_{\mathrm{ARk}}\right)\right\}\end{aligned}$$

from Lemma 1. The expected loss from an $AR\left(k\right)$ forecast is

$$\begin{aligned}\mathbb{E}\left[L\left(Y_t,\hat{Y}_t^{\mathrm{ARk}};\phi\right)\right] &= 2a^{-2}\left(\exp\left\{\frac{a}{2}\left(2\mu+a\gamma_0\right)\right\}-\exp\left\{\frac{a}{2}\left(2\mu+aV_{\mathrm{ARk}}\right)\right\}\right)\\ &\rightarrow \gamma_0-V_{\mathrm{ARk}}\text{ as }a\rightarrow 0\end{aligned}$$

Note that we know that $V_{\mathrm{AR1}}\leq V_{\mathrm{AR2}}\leq V_{\mathrm{AR5}}$ and so we immediately see that the ranking under MSE (i.e., Exponential-Bregman with $a\rightarrow 0$) is $\mathbb{E}\left[L\left(Y_t,\hat{Y}_t^{\mathrm{AR1}}\right)\right]\geq\mathbb{E}\left[L\left(Y_t,\hat{Y}_t^{\mathrm{AR2}}\right)\right]\geq\mathbb{E}\left[L\left(Y_t,\hat{Y}_t^{\mathrm{AR5}}\right)\right]$.

Appendix B-2: Derivations for the Bernoulli forecasters in Section 2.2

Since $\hat{Y}_t^X$ and $\hat{Y}_t^W$ are both optimal with respect to their limited information, they are both "mean unbiased" and so their expected Bregman loss simplifies to $2a^{-2}\left(E\left[\exp\left\{aY_t\right\}-\exp\left\{a\hat{Y}_t\right\}\right]\right)$.

For this DGP we easily find:

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{aY_t\right\}\right] &= \exp\left\{\frac{a^2}{2}+a\left(\mu_L+\mu_C\right)\right\}pq+\exp\left\{\frac{a^2}{2}+a\left(\mu_H+\mu_C\right)\right\}\left(1-p\right)q \\
&\quad +\exp\left\{\frac{a^2}{2}+a\left(\mu_L+\mu_M\right)\right\}p\left(1-q\right)+\exp\left\{\frac{a^2}{2}+a\left(\mu_H+\mu_M\right)\right\}\left(1-p\right)\left(1-q\right) \\
\mathbb{E}\left[\exp\left\{a\hat{Y}_t^X\right\}\right] &= \exp\left\{a\left(\mu_L+q\mu_C+\left(1-q\right)\mu_M\right)\right\}p+\exp\left\{a\left(\mu_H+q\mu_C+\left(1-q\right)\mu_M\right)\right\}\left(1-p\right) \\
\mathbb{E}\left[\exp\left\{a\hat{Y}_t^W\right\}\right] &= \exp\left\{a\left(p\mu_L+\left(1-p\right)\mu_H+\mu_C\right)\right\}q+\exp\left\{a\left(p\mu_L+\left(1-p\right)\mu_H+\mu_M\right)\right\}\left(1-q\right)
\end{aligned} \tag{57}$$

Figure 2 normalizes the expected loss from forecast X and W by the optimal forecast, using both signals:

$$\hat{Y}_t^{XW}=X_t\mu_L+\left(1-X_t\right)\mu_H+W_t\mu_C+\left(1-W_t\right)\mu_M \tag{58}$$

which leads to

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{a\hat{Y}_t^{XW}\right\}\right] &= \exp\left\{a\left(\mu_L+\mu_C\right)\right\}pq+\exp\left\{a\left(\mu_H+\mu_C\right)\right\}\left(1-p\right)q \\
&\quad +\exp\left\{a\left(\mu_L+\mu_M\right)\right\}p\left(1-q\right)+\exp\left\{a\left(\mu_H+\mu_M\right)\right\}\left(1-p\right)\left(1-q\right)
\end{aligned} \tag{59}$$

Appendix B-3: Derivations for the linear model in Section 2.3

The first-order condition for the optimal parameter $\theta\equiv[\alpha,\beta]$ is:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\mathbb{E}\left[L\left(Y,m\left(X;\theta\right);\phi\right)\right] \\
&= \mathbb{E}\left[\phi''\left(m\left(X;\theta\right)\right)\left(\mathbb{E}\left[Y|X\right]-m\left(X;\theta\right)\right)\frac{\partial m\left(X;\theta\right)}{\partial\theta}\right] \\
&= 2\mathbb{E}\left[\exp\left\{a\left(\alpha+\beta X\right)\right\}\left(X^2-\alpha-\beta X\right)\left[1,X\right]'\right]
\end{aligned} \tag{60}$$

So the two first-order conditions are:

$$0 = \mathbb{E}\left[\exp\left\{a\left(\alpha+\beta X\right)\right\}X^2\right]-\alpha\mathbb{E}\left[\exp\left\{a\left(\alpha+\beta X\right)\right\}\right]-\beta\mathbb{E}\left[\exp\left\{a\left(\alpha+\beta X\right)\right\}X\right] \tag{61}$$

$$0 = \mathbb{E}\left[\exp\left\{a\left(\alpha+\beta X\right)\right\}X^3\right]-\alpha\mathbb{E}\left[\exp\left\{a\left(\alpha+\beta X\right)\right\}X\right]-\beta\mathbb{E}\left[\exp\left\{a\left(\alpha+\beta X\right)\right\}X^2\right] \tag{62}$$

Using Lemma 1 above we have each of the four unique terms above in closed form. Substituting these in and solving for solving for $[\alpha,\beta]$ yields the expressions given in equation (25).

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] Ehm, W., T. Gneiting, A. Jordan, F. and Krüger, F., 2016, Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings, *Journal of the Royal Statistical Society, Series B*, 78, 505-562. With discussion and rejoinder.

[7] Elliott, G., D. Ghanem, and F. Krüger, 2016, Forecasting Conditional Probabilities of Binary Outcomes under Misspecification, *Review of Economics and Statistics*, 98(4), 742-755.

[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] Gourieroux, C., A. Monfort, and A. Trognon, 1984, Pseudo Maximum Likelihood Methods: Theory, *Econometrica*, 52(3), 681-700.

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

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

[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] 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.

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

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

[17] Hansen, P.R., and E.-I. Dumitrescu, 2016, Parameter Estimation with Out-of-Sample Objective, Working paper, Department of Economics, UNC-Chapel Hill.

[18] 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.

[19] Koenker, R., V. Chernozhukov, X. He, and L. Peng, 2017, *Handbook of Quantile Regression*, forthcoming, CRC Press.

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

[21] Lieli, R.P. and M.B. Stinchcombe, 2013, On the Recoverability of Forecasters' Preferences, *Econometric Theory*, 29, 517-544.

[22] Lieli, R.P. and M.B. Stinchcombe, 2016, Bregman and GPL Losses: Identified or Not? Working paper, Department of Economics, Central European University.

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

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

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

[26] Saerens, M., 2000, Building Cost Functions Minimizing to Some Summary Statistics, *IEEE Transactions on Neural Networks*, 11, 1263-1271.

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

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

[29] Thomson, W., 1979, Eliciting Production Possibilities from a Well-Informed Manager, *Journal of Economic Theory*, 20, 360-380.

[30] 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.

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

[32] White, H. 1994, *Estimation, Inference and Specification Analysis*, Econometric Society Monographs No. 22, Cambridge University Press.

[33] White, H., 2001, *Asymptotic Theory for Econometricians*, 2nd Ed., San Diego, Academic Press.

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

Table 1: Summary Statistics

Panel A: Consensus forecasts

	Actual	*SPF*	*Michigan*
Mean	2.760	3.162	3.163
Standard deviation	1.373	1.279	0.689
Minimum	-1.378	1.565	1.700
Maximum	6.255	8.058	6.900

Panel B: Individual forecasts

	Forecaster 20	*Forecaster 506*	*Forecaster 510*
Mean	3.432	1.640	2.261
Standard deviation	1.453	0.508	0.434
Minimum	0.853	-0.047	1.256
Maximum	12.142	2.597	3.136

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 over the period 1982Q3 to 2016Q2. Panel B considers corresponding forecasts from three individual respondents to the Survey of Professional Forecasters, across all observations available for each of the respondents.

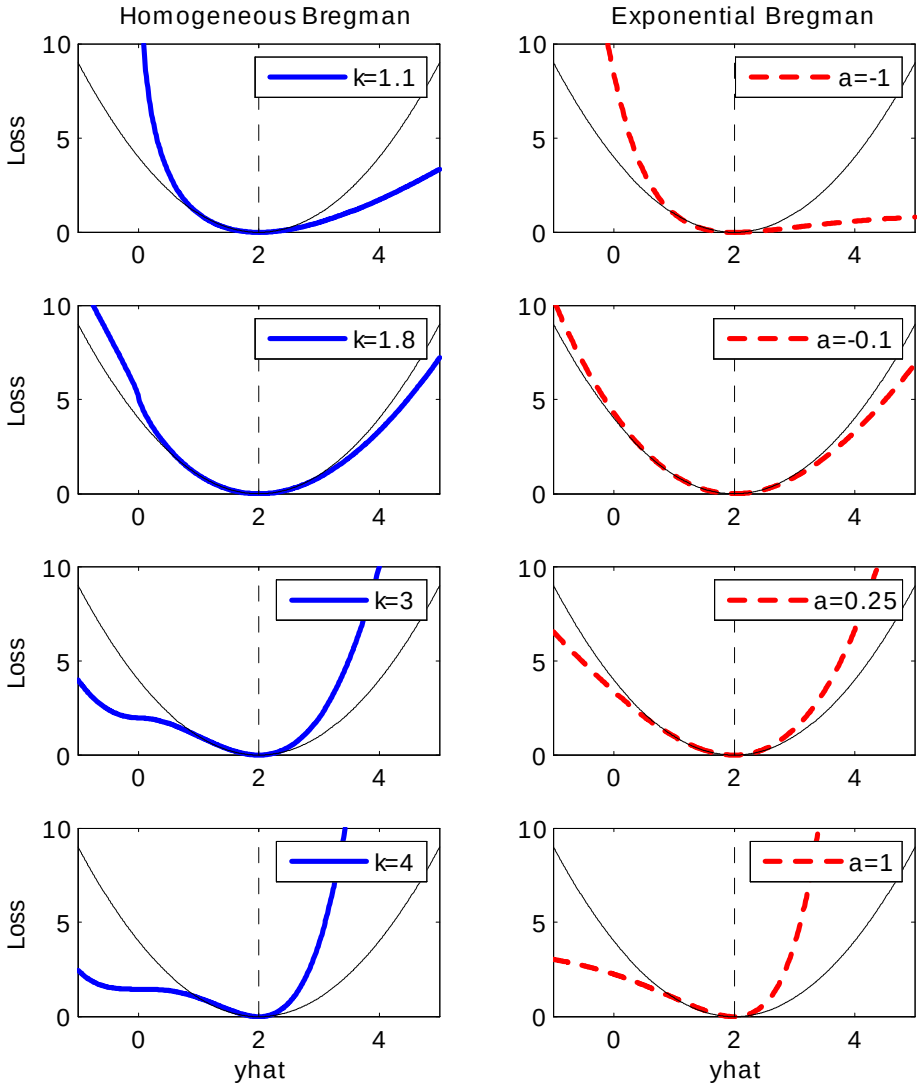

Figure 1: *Various Bregman loss functions. The left column presents four elements of the "homogeneous Bregman" family, and the right column presents four elements of the "exponential Bregman" family. The squared error loss function is also presented in each panel. In all cases the value for* $\hat{y}$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

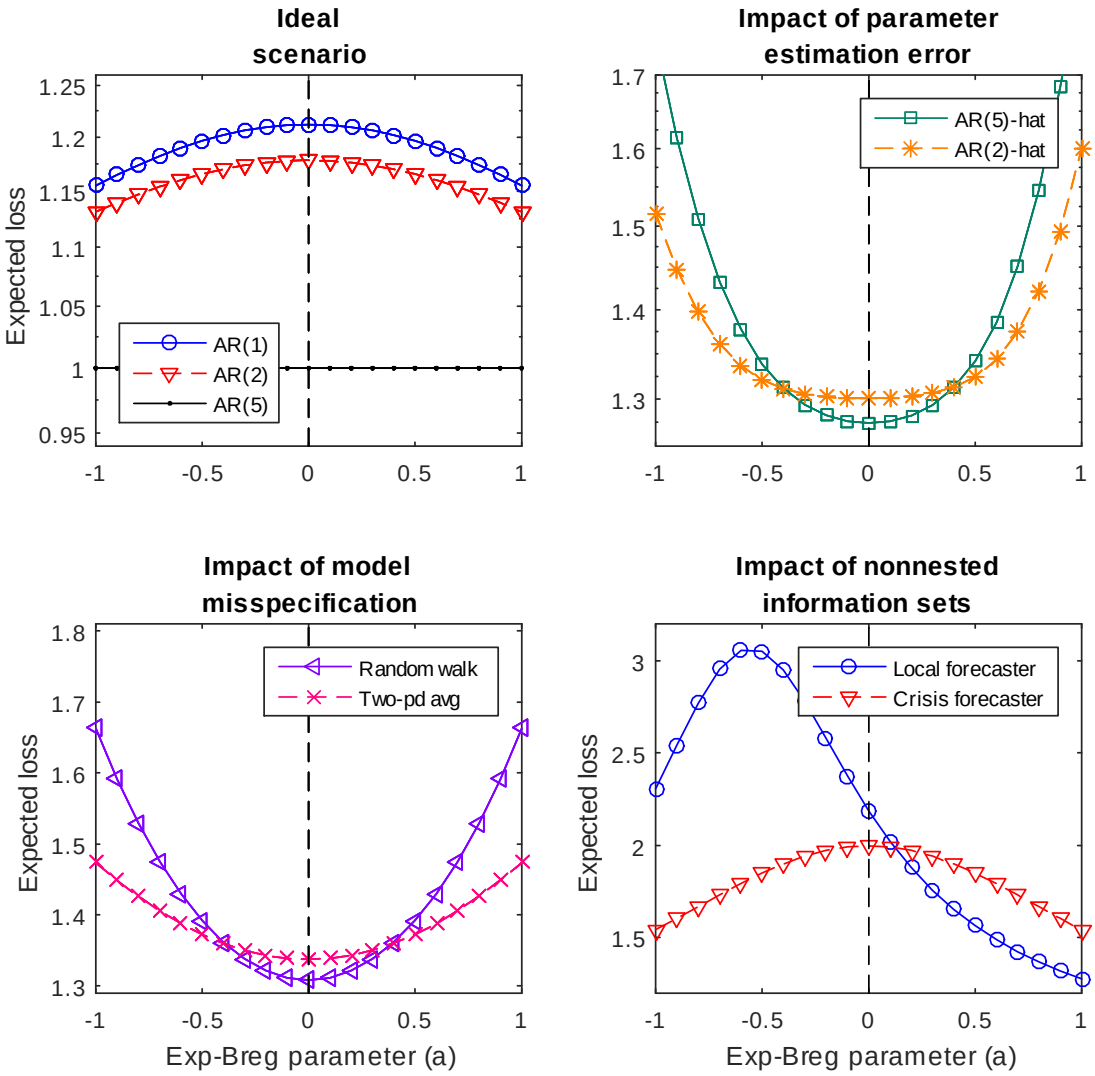

Figure 2: *This figure presents the ratio of expected Exponential Bregman loss for a given forecast to that for the optimal forecast, as a function of the Exponential Bregman parameter a.*

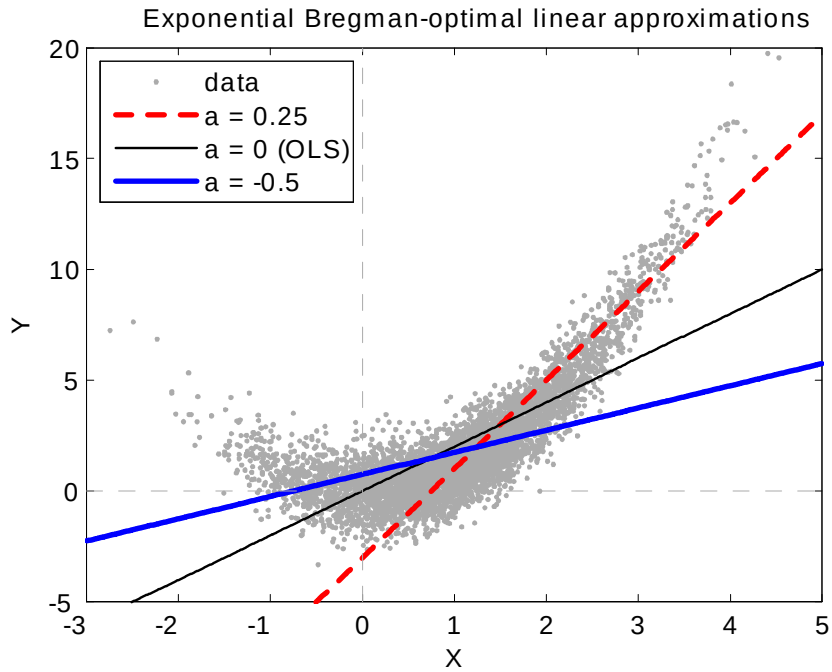

Figure 3: *This figure presents the optimal linear approximations to a nonlinear DGP based on the exponential Bregman loss function for three choices of "shape" parameter; the choice a=0 corresponds to quadratic loss, and the fit is the same as that obtained by OLS.*

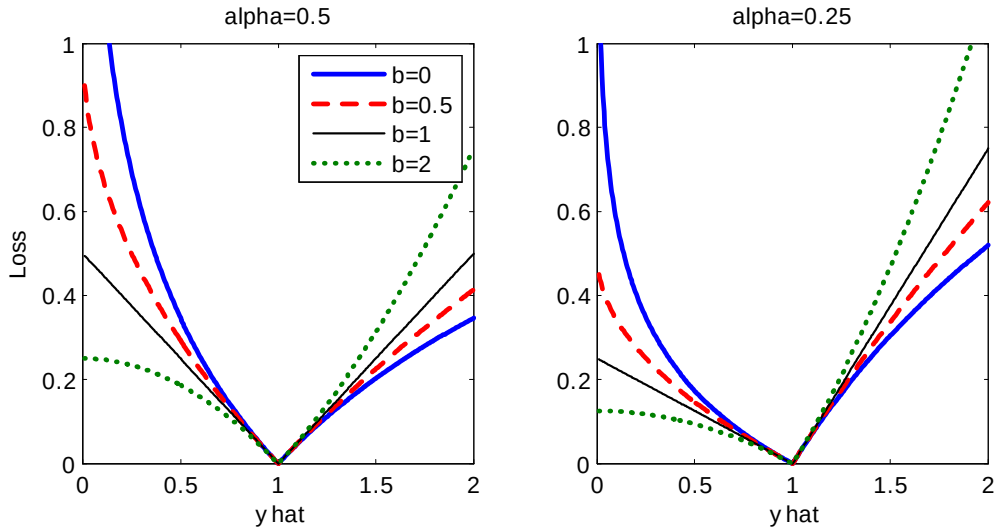

Figure 4: *Various homogenous GPL loss functions, with* $\alpha = 0.5$ *(left panel) and* $\alpha = 0.25$ *(right panel). The "Lin-Lin" (or "tick") loss function is obtained when* $b = 1$*. In both cases the value for* $\hat{y}$ *ranges from 0 to 2, and the value of* y *is set at 1.*

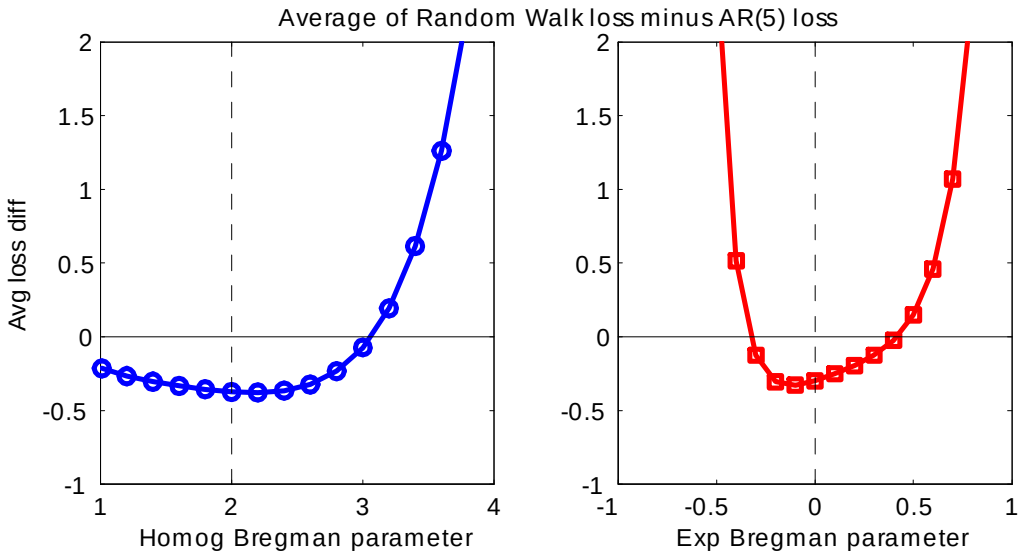

Figure 5: *Average loss from a random walk forecast minus that from an estimated AR(5) forecast, for various homogeneous (left panel) and exponential (right panel) Bregman loss functions.*

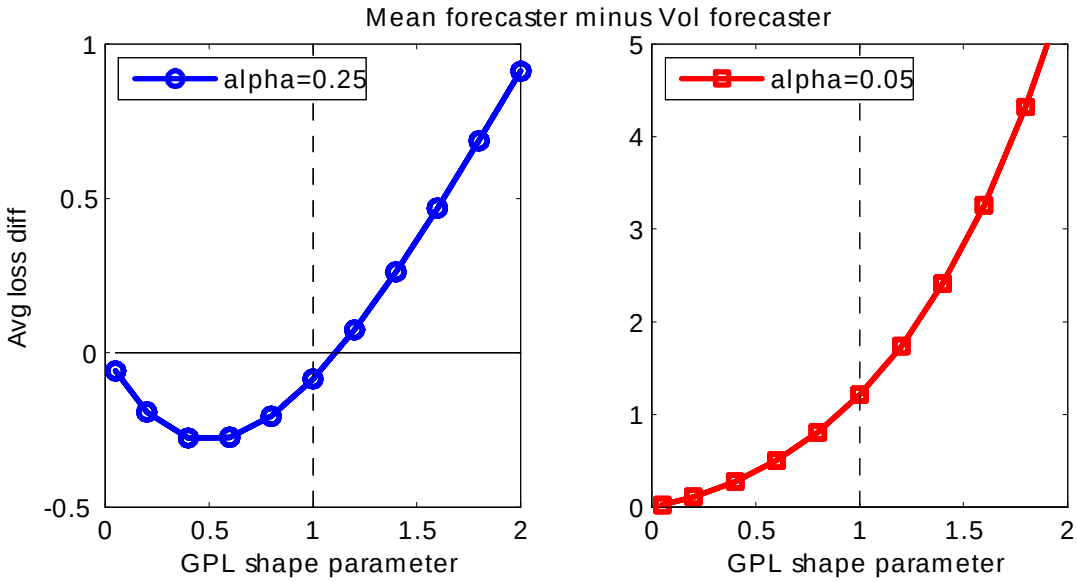

Figure 6: *Average loss from a AR-constant volatility forecsast minus that from a constant mean-GARCH forecast for various GPL loss functions. (Lin-Lin loss is marked with a vertical line at 1.) The left panel is for the 0.25 quantile, and the right panel is for the 0.05 quantile.*

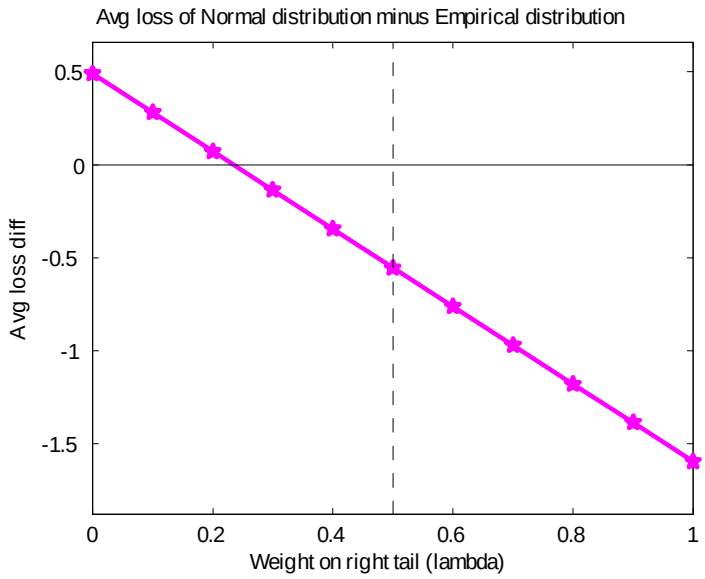

Figure 7: *Average wCRSP loss from a Normal distribution forecast minus that from an empirical distribution forecast based on 100 observations. The x-axis plots different weights on the left/right tail, with equal weight at 0.5, indicated with a vertical line.*

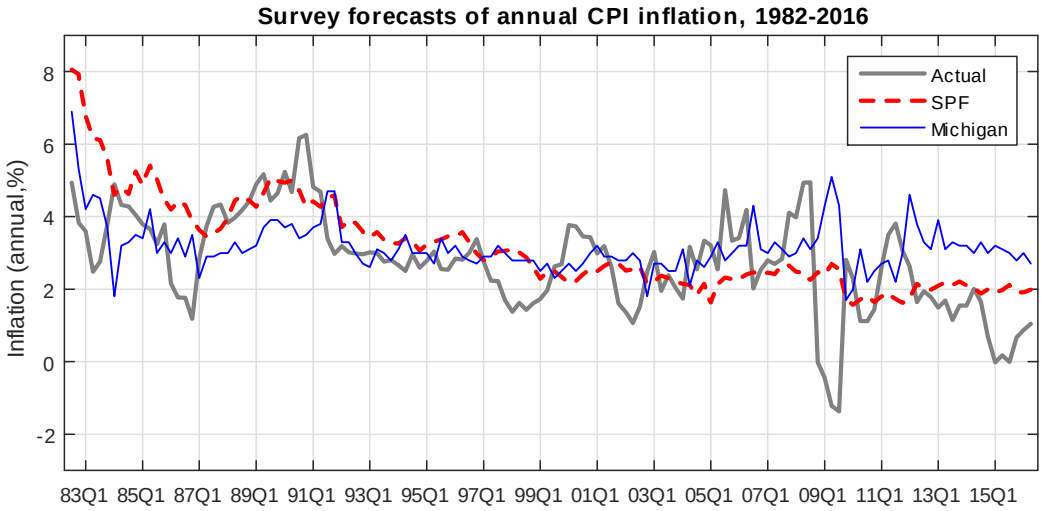

Figure 8: *Time series of actual and predicted annual US CPI inflation, updated quarterly, over the period 1982Q3–2016Q2. The inflation forecasts are from the Survey of Professional Forecasters and the Michigan survey of consumers.*

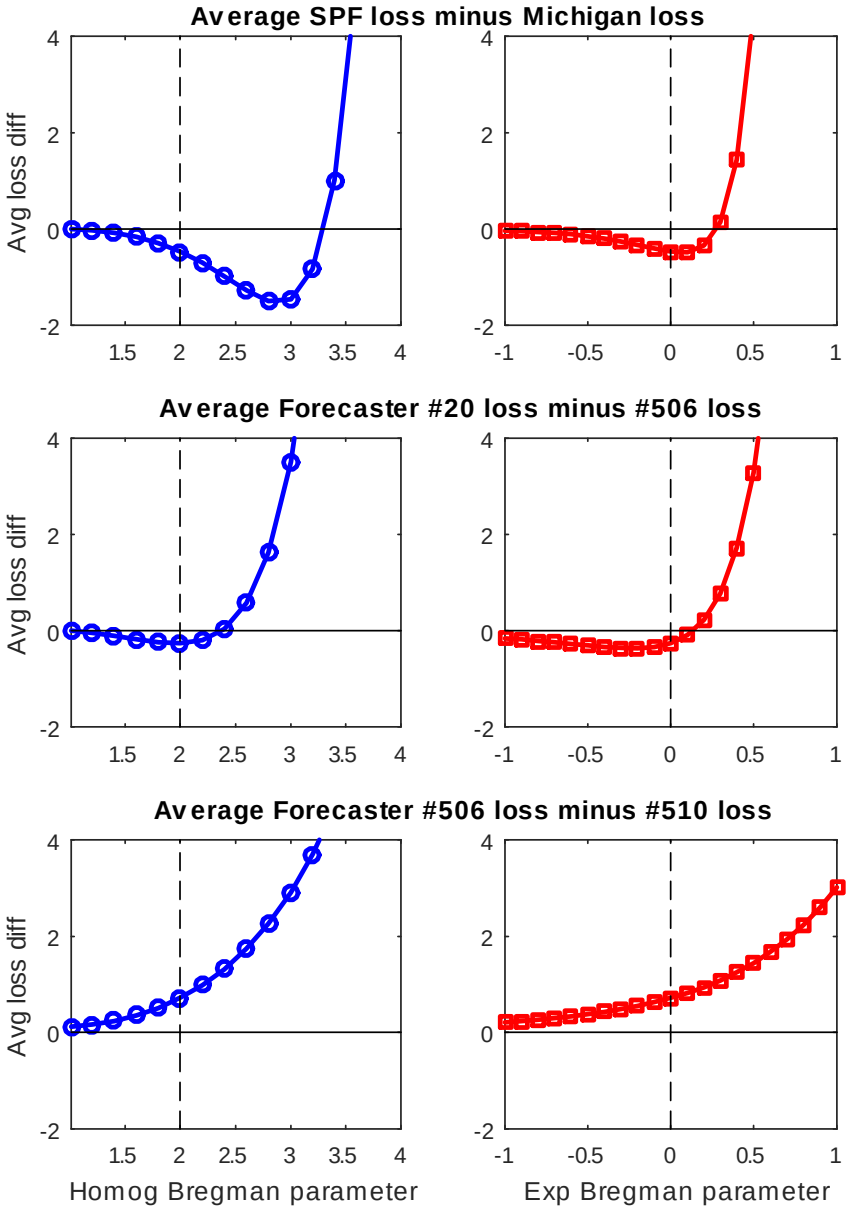

Figure 9: *Differences in average losses between two forecasts, for a range of loss function parameters. The "homogeneous Bregman" loss function is in the left column, and the "exponential Bregman" loss function is in the right column. The squared-error loss function is nested at 2 and 0 for these loss 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.*