Making and Evaluating Point Forecasts

Tilmann GNEITING

Typically, point forecasting methods are compared and assessed by means of an error measure or scoring function, with the absolute error and the squared error being key examples. The individual scores are averaged over forecast cases, to result in a summary measure of the predictive performance, such as the mean absolute error or the mean squared error. I demonstrate that this common practice can lead to grossly misguided inferences, unless the scoring function and the forecasting task are carefully matched. Effective point forecasting requires that the scoring function be specified ex ante, or that the forecaster receives a directive in the form of a statistical functional, such as the mean or a quantile of the predictive distribution. If the scoring function is specified ex ante, the forecaster can issue the optimal point forecast, namely, the Bayes rule. If the forecaster receives a directive in the form of a functional, it is critical that the scoring function be consistent for it, in the sense that the expected score is minimized when following the directive. A functional is elicitable if there exists a scoring function that is strictly consistent for it. Expectations, ratios of expectations and quantiles are elicitable. For example, a scoring function is consistent for the mean functional if and only if it is a Bregman function. It is consistent for a quantile if and only if it is generalized piecewise linear. Similar characterizations apply to ratios of expectations and to expectiles. Weighted scoring functions are consistent for functionals that adapt to the weighting in peculiar ways. Not all functionals are elicitable; for instance, conditional value-at-risk is not, despite its popularity in quantitative finance.

KEY WORDS: Bayes rule; Bregman function; Conditional value-at-risk (CVaR); Decision theory; Elicitability; Expectile; Mean; Median; Mode; Proper scoring rule; Quantile; Statistical functional.

1. INTRODUCTION

In many aspects of human activity, a major desire is to make forecasts for an uncertain future. Consequently, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities or events (Dawid 1984; Gneiting 2008a). Still, many practical situations require single-valued point forecasts, for reasons of decision making, market mechanisms, reporting requirements, communication, or tradition, among others.

1.1 Using Scoring Functions to Evaluate Point Forecasts

In this type of situation, competing point forecasters or forecasting procedures are compared and assessed by means of an error measure, such as the absolute error or the squared error, which is averaged over forecast cases. Thus, the performance criterion takes the form

$$\bar{\mathrm{S}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(x_i, y_i), \tag{1}$$

where there are n forecast cases with corresponding point forecasts, $x_1, \ldots, x_n$, and verifying observations, $y_1, \ldots, y_n$. The function S depends both on the forecast and the realization, and we refer to it as a *scoring function*.

Table 1 lists some commonly used scoring functions. We generally take scoring functions to be *negatively oriented*, that is, the smaller, the better. The absolute error and the squared error are of the *prediction error* form, in that they depend on the forecast error, $x - y$, only, and they are *symmetric*, in that $\mathrm{S}(x, y) = \mathrm{S}(y, x)$. The absolute percentage error and the relative error are used for strictly positive quantities only; they are neither of the prediction error form nor symmetric. Patton (2011) discusses these as well as many other scoring functions that have been used to assess point forecasts for a strictly positive quantity, such as an asset value or a volatility proxy.

Our next two tables summarize the use of scoring functions in academia, the public and the private sector. Table 2 surveys the 2008 volumes of peer-reviewed journals in forecasting (Group I) and statistics (Group II), along with premier journals in two prominent application areas, namely econometrics (Group III) and meteorology (Group IV). We call an article a *forecasting article* if it contains a table or a figure in which the predictive performance of a forecaster or forecasting method is summarized in the form of the mean score (1), or a monotone transformation thereof, such as the root mean squared error. Not surprisingly, the majority of the Group I articles are forecasting articles, and many of them employ several scoring functions simultaneously. Overall, the squared error is the most popular scoring function in academia, particularly in Groups III and IV, followed by the absolute error and the absolute percentage error.

Table 3 reports the use of scoring functions in businesses and organizations, according to surveys conducted or summarized by Carbone and Armstrong (1982), Mentzner and Kahn (1995), McCarthy et al. (2006) and Fildes and Goodwin (2007). In addition to the squared error and the absolute error, the absolute percentage error has been very widely used in practice, presumably because business forecasts focus on demand, sales, or costs, all of which are nonnegative quantities.

There are many options and considerations in choosing a scoring function. What scoring function ought to be used in

Tilmann Gneiting is Professor, Institute of Applied Mathematics, University of Heidelberg, 69120 Heidelberg, Germany (E-mail: *t.gneiting@uni-heidelberg.de*). The author thanks Philip Dawid, Werner Ehm, Marc G. Genton, Kyrill Grant, Peter Guttorp, Leonhard Held, Jorgen Hilden, Hajo Holzmann, Peter J. Huber, Ian T. Jolliffe, Alex Lenkoski, Charles F. Manski, Caren Marzban, Kent Osband, Pierre Pinson, Adrian E. Raftery, Ken Rice, R. Tyrrell Rockafellar, Havard Rue, Paul D. Sampson, J. McLean Sloughter, Stephen Stigler, Adam Szpiro, Thordis L. Thorarinsdottir, Jon A. Wellner, Robert L. Winkler, four anonymous reviewers and the editor for comments, references and preprints. Financial support was provided by the Alfried Krupp von Bohlen und Halbach Foundation, and by the National Science Foundation under Awards ATM-0724721 and DMS-0706745 to the University of Washington. Special thanks go to University of Washington librarians Martha Tucker and Saundra Martin for their unfailing support of the literature survey in Table 2. Of course, the opinions expressed in this article as well as any errors are solely the responsibility of the author.

Journal of the American Statistical Association
June 2011, Vol. 106, No. 494, Review Article
DOI: 10.1198/jasa.2011.r10138

Table 1. Some commonly used scoring functions

$S(x, y) = (x - y)^2$	squared error (SE)
$S(x, y) = \|x - y\|$	absolute error (AE)
$S(x, y) = \|(x - y)/y\|$	absolute percentage error (APE)
$S(x, y) = \|(x - y)/x\|$	relative error (RE)

practice? Do the standard choices have theoretical support? Arguably, there is considerable contention in the scientific community, along with a critical need for theoretically principled guidance. Some 20 years ago, Murphy and Winkler (1987, p. 1330) commented on the state of the art in forecast evaluation, noting that

"[. . .] verification measures have tended to proliferate, with relatively little effort being made to develop general concepts and principles [. . .] This state of affairs has impacted the development of a science of forecast verification."

Nothing much has changed since. Armstrong (2001) called for further research, while Moskaitis and Hansen (2006) asked

"Deterministic forecasting and verification: A busted system?"

Similarly, the recent review by Fildes et al. (2008, p. 1158) states that

"Defining the basic requirements of a good error measure is still a controversial issue."

1.2 Simulation Study

To focus on issues and ideas, we consider a simulation study in which we seek point forecasts for a highly volatile daily asset value, y_t. The data-generating process is such that y_t is a realization of the random variable

$$Y_t = Z_t^2, \tag{2}$$

where Z_t follows a conditionally heteroscedastic Gaussian time series model (Engle 1982; Bollerslev 1986), with the parameter values proposed by Christoffersen and Diebold (1996), in that

$$Z_t \sim \mathcal{N}(0, \sigma_t^2), \quad \text{where } \sigma_t^2 = 0.20Z_{t-1}^2 + 0.75\sigma_{t-1}^2 + 0.05.$$

We consider three forecasters, each of whom issues a one-day-ahead point forecast for the asset value. The *statistician* has knowledge of the data-generating process and the actual value of the conditional variance σ_t^2, and thus predicts the true conditional mean,

$$\hat{x}_t = \mathbb{E}(Y_t|\sigma_t^2) = \sigma_t^2,$$

as her point forecast. The *optimist* always predicts $\hat{x}_t = 5$. The *pessimist* always issues the point forecast $\hat{x}_t = 0.05$. Figure 1 shows these point forecasts along with the realizing asset value for 200 successive trading days. There ought to be little contention as to the predictive performance, in that the statistician is more skilled than the optimist or the pessimist.

Table 4 provides a formal evaluation of the three forecasters for a sequence of $n = 100{,}000$ sequential forecasts, using the mean score (1) and the scoring functions listed in Table 1. The results are counterintuitive and disconcerting, in that the pessimist has the best (lowest) score both under the absolute error and the absolute percentage error scoring functions. In terms of relative error, the optimist performs best. Yet, what we have done here is common practice in academia and businesses, in that point forecasts are evaluated by means of these scoring functions.

1.3 Discussion

The source of these disconcerting results is aptly explained in a recent article by Engelberg, Manski, and Williams (2009, p. 30):

Table 2. Use of scoring functions in the 2008 volumes of leading peer-reviewed journals in forecasting (Group I), statistics (Group II), econometrics (Group III) and meteorology (Group IV). Column 2 shows the total number of articles published in 2008 under Web of Science document type article, note or review. Column 3 shows the number of forecasting articles (FP), that is, the number of articles with a table or figure that summarizes predictive performance in the form of the mean score (1) or a monotone transformation thereof. Columns 4 through 7 show the number of articles employing the squared error (SE), absolute error (AE), absolute percentage error (APE), or miscellaneous (MSC) other scoring functions. The sum of columns 4 through 7 may exceed the number in column 3, because of the simultaneous use of multiple scoring functions in some articles. Articles that apply error measures to evaluate estimation methods, rather than forecasting methods, have not been considered in this study

	Total	FP	SE	AE	APE	MSC
Group I: *Forecasting*						
International Journal of Forecasting	41	32	21	10	8	4
Journal of Forecasting	39	25	23	13	5	3
Group II: *Statistics*						
The Annals of Applied Statistics	62	8	6	3	1	0
The Annals of Statistics	100	5	3	2	0	0
Journal of the American Statistical Association	129	10	9	1	0	0
Journal of the Royal Statistical Society, Ser. B	49	5	4	1	0	0
Group III: *Econometrics*						
Journal of Business & Economic Statistics	26	9	8	2	1	0
Journal of Econometrics	118	5	5	0	0	0
Group IV: *Meteorology*						
Bulletin of the American Meteorological Society	73	1	1	0	0	0
Monthly Weather Review	300	63	58	8	2	0
Quarterly Journal of the Royal Meteorological Society	148	19	19	0	0	0
Weather and Forecasting	79	26	20	11	0	1

Table 3. Use of scoring functions in the evaluation of point forecasts in businesses and organizations. Columns 2 through 4 show the percentage of survey respondents using the squared error (SE), absolute error (AE) and absolute percentage error (APE), with the source of the survey listed in column 1

Source	SE	AE	APE
Carbone and Armstrong (1982, table 1)	27%	19%	9%
Mentzner and Kahn (1995, table VIII)	10%	25%	52%
McCarthy et al. (2006, table VIII)	6%	20%	45%
Fildes and Goodwin (2007, table 5)	9%	36%	44%

"Our concern is prediction of real-valued outcomes such as firm profit, GDP, growth, or temperature. In these cases, the users of point predictions sometimes presume that forecasters report the means of their subjective probability distributions; that is, their best point predictions under square loss. However, forecasters are not specifically asked to report subjective means. Nor are they asked to report subjective medians or modes, which are best predictors under other loss functions. Instead, they are simply asked to 'predict' the outcome or to provide their 'best prediction,' without definition of the word 'best.' In the absence of explicit guidance, forecasters may report different distributional features as their point predictions. Some may report subjective means, others subjective medians or modes, and still others, applying asymmetric loss functions, may report various quantiles of their subjective probability distributions."

Similarly, Murphy and Daan (1985, p. 391) noted that

"It will be assumed here that the forecasters receive a 'directive' concerning the procedure to be followed [...] and that it is desirable to choose an evaluation measure that is consistent with this concept. An example may help to illustrate this concept. Consider a continuous [...] predictand, and suppose that the directive states 'forecast the expected (or mean) value of the variable.' In this situation, the mean square error measure would be an appropriate scoring rule, since it is minimized by forecasting the mean of the (judgemental) probability distribution. Measures that correspond with a directive in this sense will be referred to as consistent scoring rules (for that directive)."

Despite these well-argued perspectives, there has been little recognition that the common practice of requesting "some" point forecast, and then evaluating the forecasters by using "some" (set of) scoring function(s), is not a meaningful endeavor. In this article, we develop the perspectives of Murphy and Daan (1985) and Engelberg, Manski, and Williams (2009) and argue that effective point forecasting depends on "guidance" or "directives," which can be given in one of two complementary ways, namely, by disclosing the *scoring function* ex ante to the forecaster, or by requesting a specific *functional* of the forecaster's predictive distribution, such as the mean or a quantile.

As to the first option, the a priori disclosure of the scoring function allows the forecaster to tailor the point predictor to the scoring function at hand. In particular, this permits our statistician forecaster to mutate into *Mr. Bayes*, who issues the optimal point forecast, namely the Bayes rule,

$$\hat{x} = \arg\min_x \mathbb{E}_F \mathrm{S}(x, Y), \tag{3}$$

where the random variable Y is distributed according to the forecaster's subjective or objective predictive distribution, F. For example, if the scoring function S is the squared error, the optimal point forecast is the mean of the predictive distribution. In the case of the absolute error, the Bayes rule is any median of the predictive distribution. The class

$$\mathrm{S}_\beta(x, y) = \left| 1 - \left(\frac{y}{x} \right)^\beta \right| \qquad (\beta \neq 0) \tag{4}$$

of scoring functions nests both the absolute percentage error ($\beta = -1$) and the relative error ($\beta = 1$) scoring functions. If the predictive distribution F has density f on the positive half-axis and a finite fractional moment of order β, the optimal point forecast under the loss or scoring function (4) is the median of a random variable whose density is proportional to $y^\beta f(y)$. We call this the *β-median* of the probability distribution F and write $\mathrm{med}^{(\beta)}(F)$. The traditional median arises in the limit as $\beta \to 0$.

Table 5 summarizes our discussion, by showing the optimal point forecast, or Bayes rule, under the scoring functions

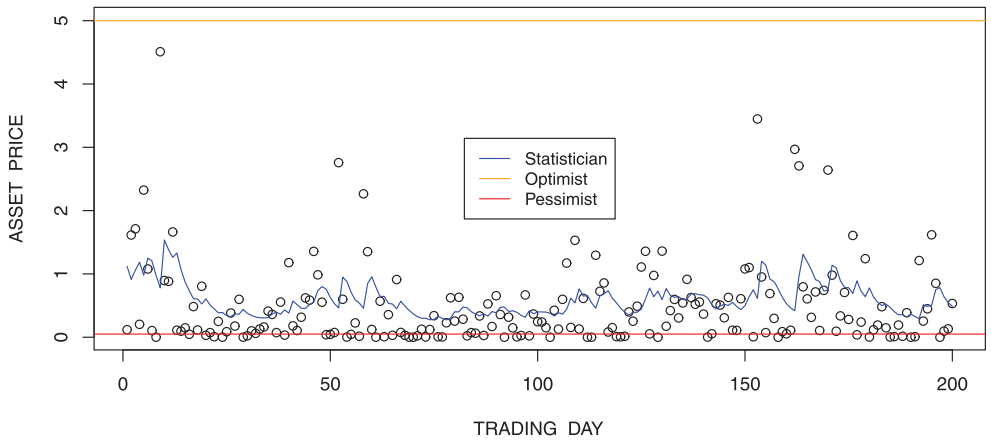

Figure 1. A realized series of volatile daily asset prices under the data-generating process (2), shown by circles, along with the one-day-ahead point forecasts by the statistician (blue curve), the optimist (orange line at top) and the pessimist (red line at bottom). The online version of this figure is in color.

Table 4. The mean error measure (1) for the three point forecasters in the simulation study, using the squared error (SE), absolute error (AE), absolute percentage error (APE) and relative error (RE) scoring functions

Forecaster	SE	AE	APE	RE
Statistician	5.07	0.97	2.58×10^5	0.97
Optimist	22.73	4.35	13.96×10^5	0.87
Pessimist	7.61	0.96	0.14×10^5	19.24

in Table 1, both in full generality and in the special case of the true predictive distribution under the data generating process (2). Table 6 shows the mean score (1) for the new competitor Mr. Bayes in the simulation study, who issues the optimal point forecast. As expected, Mr. Bayes outperforms his colleagues.

An alternative to disclosing the scoring function is to request a specific functional of the forecaster's predictive distribution, such as the mean or a quantile, and to apply any scoring function that is consistent with the functional, roughly in the following sense. Let the interval I be the potential range of the outcomes, such as $\mathrm{I} = \mathbb{R}$ for a real-valued quantity, or $\mathrm{I} = (0, \infty)$ for a strictly positive quantity, and let the probability distribution F be concentrated on I. Then a *scoring function* is any mapping $\mathrm{S} : \mathrm{I} \times \mathrm{I} \to [0, \infty)$. A *functional* is a potentially set-valued mapping $F \mapsto \mathrm{T}(F) \subseteq \mathrm{I}$. A scoring function S is *consistent* for the functional T if

$$\mathbb{E}_F[\mathrm{S}(t, Y)] \le \mathbb{E}_F[\mathrm{S}(x, Y)]$$

for all F, all $t \in \mathrm{T}(F)$ and all $x \in \mathrm{I}$. It is *strictly consistent* if it is consistent and equality of the expectations implies that $x \in \mathrm{T}(F)$. Following Osband (1985) and Lambert, Pennock, and Shoham (2008), a functional is *elicitable* if there exists a scoring function that is strictly consistent for it.

1.4 Plan of the Article

The remainder of the article is organized as follows. Section 2 develops the notions of consistency and elicitability in a comprehensive way, including discussions of weighted scoring functions and of a construction which we call Osband's principle. Section 3 turns to examples. The mean functional, ratios of expectations, quantiles and expectiles are elicitable. Subject to weak regularity conditions, a scoring function for a real-valued predictand is consistent for the mean functional if and only if it is a Bregman function, that is, of the form

$$\mathrm{S}(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x),$$

where ϕ is a convex function with subgradient ϕ' (Savage 1971). More general and novel results apply to ratios of expectations and expectiles. A scoring function is consistent for the α-quantile if and only if it is generalized piecewise linear (GPL) of order $\alpha \in (0, 1)$, that is, of the form

$$\mathrm{S}(x, y) = (\mathbb{1}(x \ge y) - \alpha)(g(x) - g(y)),$$

where $\mathbb{1}(\cdot)$ denotes an indicator function and g is nondecreasing (Thomson 1979; Saerens 2000). However, not all functionals are elicitable. Notably, the conditional value-at-risk (CVaR) functional is not elicitable, despite its popularity as a risk measure in financial applications. In these core sections, Theorems 2, 5, 8, 10, and 11 are new results, with proofs supplied in Appendix A. Section 4 then turns to point forecasts of multivariate quantities.

The article closes with a discussion in Section 5, which makes a plea for change in the practice of point forecasting. I contend that in issuing and evaluating point forecasts, it is essential that either the scoring function be specified ex ante, or an elicitable target functional be named, such as an expectation or a quantile, and scoring functions be used that are consistent for the target functional.

Table 5. Bayes rules under the scoring functions in Table 1 as a functional of the forecaster's predictive distribution, F. The functional $\mathrm{med}^{(\beta)}(F)$ is defined in the text. The final column specializes to the true predictive distribution under the data-generating process (2) in the simulation study. The entry for the absolute percentage error (APE) is to be understood as follows. The predictive distribution F has infinite fractional moment of order -1, and thus $\mathrm{med}^{(-1)}(F)$ does not exist. However, the smaller the (strictly positive) point forecast, the smaller the expected APE. Thus, a prudent forecaster will issue some very small $\epsilon > 0$ as point predictor

Scoring function	Bayes rule	Point forecast in simulation study
SE	$\hat{x} = \mathrm{mean}(F)$	σ_t^2
AE	$\hat{x} = \mathrm{median}(F)$	$0.455\sigma_t^2$
APE	$\hat{x} = \mathrm{med}^{(-1)}(F)$	ε
RE	$\hat{x} = \mathrm{med}^{(1)}(F)$	$2.366\sigma_t^2$

Table 6. Continuation of Table 4, showing the corresponding mean scores for the new competitor, Mr. Bayes, who issues the Bayes rule, as described in Table 5. In the case of the APE, Mr. Bayes issues the point forecast $\hat{x} = \epsilon = 10^{-10}$

	SE	AE	APE	RE
Mr. Bayes	5.07	0.86	1.00	0.75

2. A DECISION–THEORETIC APPROACH TO THE EVALUATION OF POINT FORECASTS

We now develop a theoretical framework for the evaluation of point forecasts. Towards this end, we review the more general, classical decision-theoretic setting whose basic ingredients are as follows:

(a) An *observation domain*, O, which comprises the potential outcomes of a future observation.
(b) A class $\mathcal{F}$ of probability measures on the observation domain O (equipped with a suitable σ-algebra), which constitutes a family of probability distributions for the future observation.
(c) An *action domain*, A, which comprises the potential actions of a decision maker.
(d) A *loss function* $\mathrm{L} : \mathrm{A} \times \mathrm{O} \to [0, \infty)$, where $\mathrm{L}(a, o)$ represents the loss incurred when the decision maker takes the action $a \in \mathrm{A}$ and the observation $o \in \mathrm{O}$ materializes.

Given a probability distribution $F \in \mathcal{F}$ for the future observation, the *Bayes rule* is any decision $\hat{a} \in \mathrm{A}$ such that

$$\hat{a} = \arg\min_a \mathbb{E}_F \mathrm{L}(a, Y), \tag{5}$$

where Y is a random variable with distribution F. Thus, if the decision maker's assessment of the uncertain future is represented by the probability measure F, and she wishes to minimize the expected loss, her optimal decision is the Bayes rule, $\hat{a}$. In general, Bayes acts need not exist nor be unique, but in most cases of practical interest Bayes rules exist, and frequently they are unique (Ferguson 1967).

2.1 Decision-Theoretic Setting

Point forecasting falls into the general decision-theoretic setting, if we follow Granger and Machina (2006) and assume that the observation domain and the action domain coincide. In what follows we assume, for simplicity, that this common domain,

$$\mathrm{D} = \mathrm{O} = \mathrm{A} \subseteq \mathbb{R}^d,$$

is a subset of the Euclidean space $\mathbb{R}^d$ and equipped with the corresponding Borel σ-algebra. Furthermore, we refer to the loss function as a scoring function. With these adaptations, the basic components of our decision-theoretic framework are as follows:

(a) A *prediction-observation* (PO) *domain*, $\mathcal{D} = \mathrm{D} \times \mathrm{D}$, which is the Cartesian product of the domain $\mathrm{D} \subseteq \mathbb{R}^d$ with itself.
(b) A family $\mathcal{F}$ of potential probability distributions for the future observation $\mathbf{Y}$ that takes values in D.
(c) A *scoring function* $\mathrm{S}: \mathcal{D} = \mathrm{D} \times \mathrm{D} \to [0, \infty)$, where $\mathrm{S}(\mathbf{x}, \mathbf{y})$ represents the loss or penalty when the point forecast $\mathbf{x} \in \mathrm{D}$ is issued and the observation $\mathbf{y} \in \mathrm{D}$ materializes.

In this setting, the *optimal point forecast* under the probability distribution $F \in \mathcal{F}$ for the future observation, $\mathbf{Y}$, is the Bayes rule (5), which can now be written as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathbb{E}_F \mathrm{S}(\mathbf{x}, \mathbf{Y}). \tag{6}$$

We will mostly work in dimension $d = 1$, in which any connected domain D is simply an interval, I. The cases of prime interest then are the real line, $\mathrm{I} = \mathbb{R}$, and the nonnegative or positive halfaxis, $\mathrm{I} = [0, \infty)$ or $\mathrm{I} = (0, \infty)$.

Table 7 summarizes assumptions which some of our subsequent results impose on scoring functions on intervals. The nonnegativity condition (S0) is standard and not restrictive. Generally, a loss function can be multiplied by a strictly positive constant, and any function that depends on y only can be added, without changing the nature of the optimal point forecast (Granger and Machina 2006). Furthermore, the optimization problem in (6) is posed in terms of the point predictor, x. In this light, it is natural that assumptions (S1) and (S2) concern continuity and differentiability with respect to the first argument, the point forecast x.

Table 7. Assumptions on a scoring function S on a PO domain $\mathcal{D} = \mathrm{I} \times \mathrm{I}$, where $\mathrm{I} \subseteq \mathbb{R}$ is an interval, $x \in \mathrm{I}$ denotes the point forecast and $y \in \mathrm{I}$ the realizing observation. A tacit technical assumption is that for each $x \in \mathrm{I}$, the function $\mathrm{S}(x, \cdot)$ is measurable

(S0)	$\mathrm{S}(x, y) \geq 0$ with equality if $x = y$
(S1)	$\mathrm{S}(x, y)$ is continuous in x
(S2)	The partial derivative $\partial_x \mathrm{S}(x, y)$ exists and is continuous in x whenever $x \neq y$

Efron (1991) and Patton (2011) argue that *homogeneity* or *scale invariance* is a desirable property of a scoring function. We adopt this notion and call a scoring function S on a general PO domain $\mathcal{D} = \mathrm{D} \times \mathrm{D}$ *homogeneous* of *order b* if

$$\mathrm{S}(c\mathbf{x}, c\mathbf{y}) = |c|^b \mathrm{S}(\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathrm{D} \text{ and } c \in \mathbb{R}$$

which are such that $c\mathbf{x} \in \mathrm{D}$ and $c\mathbf{y} \in \mathrm{D}$. Evidently, the underlying quest is that for equivariance in the decision problem. The scoring function S on the PO domain $\mathcal{D} = \mathrm{D} \times \mathrm{D}$ is *equivariant* with respect to some class $\mathcal{H}$ of injections $h: \mathrm{D} \to \mathrm{D}$ if

$$\arg\min_{\mathbf{x}} \mathbb{E}_F[\mathrm{S}(\mathbf{x}, h(\mathbf{Y}))] = h\Big(\arg\min_{\mathbf{x}} \mathbb{E}_F[\mathrm{S}(\mathbf{x}, \mathbf{Y})]\Big)$$

for all $h \in \mathcal{H}$ and all probability distributions F that are concentrated on D. For instance, if S is homogeneous on $\mathrm{D} = \mathbb{R}^d$ or $\mathrm{D} = (0, \infty)^d$ then it is equivariant with respect to the multiplicative group of the linear transformations $\{\mathbf{x} \mapsto c\mathbf{x} : c > 0\}$. If the scoring function is of the prediction error form on $\mathrm{D} = \mathbb{R}^d$, then it is equivariant with respect to the translation group $\{\mathbf{x} \mapsto \mathbf{x} + \mathbf{b} : \mathbf{b} \in \mathbb{R}^d\}$.

While our decision-theoretic setting is standard and follows those of Osband (1985) and Lambert, Pennock, and Shoham (2008), and the subsequent development owes much to these pioneering works, there are distinctions in technique. For example, Osband (1985) assumes a bounded domain D, while Lambert, Pennock, and Shoham (2008) consider D to be a finite set. The work of Granger and Pesaran (2000a, 2000b), which argues in favor of closer links between decision theory and forecast evaluation, focuses on probability forecasts for a dichotomous event.

2.2 Consistency

In the decision-theoretic framework, we think of the aforementioned "distributional feature" or "directive" for the forecaster as a statistical functional. Formally, a *statistical functional*, or simply a *functional*, is a potentially set-valued mapping from a class of probability distributions, $\mathcal{F}$, to a Euclidean space (Horowitz and Manski 2006; Huber and Ronchetti 2009; Wellner 2009). In the current context of point forecasting, we require that the functional

$$\mathrm{T}: \mathcal{F} \longrightarrow \mathcal{P}(\mathrm{D}), \qquad F \longmapsto \mathrm{T}(F) \subseteq \mathrm{D},$$

maps each distribution $F \in \mathcal{F}$ to a subset $\mathrm{T}(F)$ of the domain $\mathrm{D} \subseteq \mathbb{R}^d$, that is, an element of the power set $\mathcal{P}(\mathrm{D})$. Frequently, we take $\mathcal{F}$ to be the class of all probability measures on D, or the class of the probability measures with compact support in D. Often but not always $\mathrm{T}(F)$ is single valued, with the set-valued quantile functionals being a major exception.

To facilitate the presentation, the following definitions and results suppress the dependence of the scoring function S, the functional T and the class $\mathcal{F}$ on the domain D.

Definition 1. The scoring function S is *consistent* for the functional T relative to the class $\mathcal{F}$ if

$$\mathbb{E}_F \mathrm{S}(\mathbf{t}, \mathbf{Y}) \le \mathbb{E}_F \mathrm{S}(\mathbf{x}, \mathbf{Y}) \tag{7}$$

for all probability distributions $F \in \mathcal{F}$, all $\mathbf{t} \in \mathrm{T}(F)$ and all $\mathbf{x} \in \mathrm{D}$. It is *strictly consistent* if it is consistent and equality in (7) implies that $\mathbf{x} \in \mathrm{T}(F)$.

As noted, the term consistent was coined by Murphy and Daan (1985, p. 391), who stressed that it is critically important to define consistency for a fixed, given functional, as opposed to a generic notion of consistency, which was, correctly, refuted by Jolliffe (2008). For example, the squared error scoring function, $\mathrm{S}(x, y) = (x - y)^2$, is consistent, but not strictly consistent, for the mean functional relative to the class of the probability measures on the real line with finite first moment. It is strictly consistent relative to the class of the probability measures with finite second moment.

In a parametric context, Lehmann (1951) and Noorbaloochi and Meeden (1983) refer to a related property as *decision-theoretic unbiasedness*. The following result notes that consistency is the dual of the optimal point forecast property, just as decision-theoretic unbiasedness is the dual of being Bayes (Noorbaloochi and Meeden 1983). It thus connects the problems of finding optimal point forecasts, and of evaluating point predictions.

Theorem 1. The scoring function S is consistent for the functional T relative to the class $\mathcal{F}$ if and only if, given any $F \in \mathcal{F}$, any $\mathbf{x} \in \mathrm{T}(F)$ is an optimal point forecast under S.

Stated differently, the class of the scoring functions that are consistent for a certain functional is identical to the class of the loss functions under which the functional is an optimal point forecast. Despite its simplicity, and the proof being immediate from the defining properties, this duality does not appear to be widely appreciated.

Our next result shows that the class of the consistent scoring functions is a convex cone, and thus suggests the existence of Choquet representations (Phelps 1966).

Theorem 2. Let λ be a σ-finite measure on a measurable space $(\Omega, \mathcal{A})$. Suppose that for all $\omega \in \Omega$, the scoring function S_ω satisfies (S0) and is consistent for the functional T relative to the class $\mathcal{F}$. Then the scoring function

$$\mathrm{S}(\mathbf{x}, \mathbf{y}) = \int \mathrm{S}_\omega(\mathbf{x}, \mathbf{y}) \lambda(\mathrm{d}\omega)$$

is consistent for T relative to $\mathcal{F}$.

At this point, it will be useful to distinguish the notions of a proper scoring rule (Winkler 1996; Gneiting and Raftery 2007) and a consistent scoring function. I believe that this distinction is useful, even though the extant literature has failed to make it. For example, in referring to proper scoring rules for quantile forecasts, Cervera and Muñoz (1996), Gneiting and Raftery (2007), Hilden (2008), and Jose and Winkler (2009) discuss scoring functions that are consistent for a quantile.

Within our decision-theoretic framework, a *proper scoring rule* is a function $\mathrm{R} : \mathcal{F} \times \mathrm{D} \to \mathbb{R}$ such that

$$\mathbb{E}_F \mathrm{R}(F, \mathbf{Y}) \le \mathbb{E}_F \mathrm{R}(G, \mathbf{Y}) \tag{8}$$

for all probability distributions $F, G \in \mathcal{F}$, where we assume that the expectations are well defined. Note that R is defined on the Cartesian product of the class $\mathcal{F}$ and the domain D. The loss or penalty $\mathrm{R}(F, \mathbf{y})$ arises when a probabilistic forecaster issues the predictive distribution F while $\mathbf{y} \in \mathrm{D}$ materializes. The expectation inequality (8) then implies that the forecaster minimizes the expected loss by following her true beliefs. Thus, the use of proper scoring rules encourages sincerity and candor among probabilistic forecasters.

In contrast, a scoring function S acts on the PO domain, $\mathcal{D} = \mathrm{D} \times \mathrm{D}$, that is, the Cartesian product of D with itself. This is a much simpler domain than that for a scoring rule. However, any consistent scoring function induces a proper scoring rule in a straightforward and natural construction, as follows.

Theorem 3. Suppose that the scoring function S is consistent for the functional T relative to the class $\mathcal{F}$. For each $F \in \mathcal{F}$, let $\mathbf{t}_F \in \mathrm{T}(F)$. Then the function

$$\mathrm{R} : \mathcal{F} \times \mathrm{D} \longrightarrow [0, \infty), \qquad (F, \mathbf{y}) \longmapsto \mathrm{R}(F, \mathbf{y}) = \mathrm{S}(\mathbf{t}_F, \mathbf{y}),$$

is a proper scoring rule.

A more general decision-theoretic approach to the construction of proper scoring rules is described by Dawid (2007, p. 78) and Gneiting and Raftery (2007, p. 361). Indeed, as Pesaran and Skouras (2002) note, decision-theoretic approaches provide a unifying framework for the evaluation of both probabilistic and point forecasts.

2.3 Elicitability

We turn to the notion of elicitability, which is a critically important concept in the evaluation of point forecasts. While the general notion dates back to the pioneering work of Osband (1985), the term elicitable was coined only recently by Lambert, Pennock, and Shoham (2008).

Definition 2. The functional T is *elicitable* relative to the class $\mathcal{F}$ if there exists a scoring function S that is strictly consistent for T relative to $\mathcal{F}$.

Evidently, if T is elicitable relative to the class $\mathcal{F}$, then it is elicitable relative to any subclass $\mathcal{F}_0 \subseteq \mathcal{F}$. The following result then is a version of Osband's (1985, p. 9) *revelation principle*.

Theorem 4 (Osband). Suppose that the class $\mathcal{F}$ is concentrated on the domain D, and let $g : \mathrm{D} \to \mathrm{D}$ be a one-to-one mapping. Then the following holds:

(a) If T is elicitable, then $\mathrm{T}_g = g \circ \mathrm{T}$ is elicitable.
(b) If S is consistent for T, then the scoring function

$$\mathrm{S}_g(\mathbf{x}, \mathbf{y}) = \mathrm{S}(g^{-1}(\mathbf{x}), \mathbf{y})$$

is consistent for T_g.
(c) If S is strictly consistent for T, then S_g is strictly consistent for T_g.

The next theorem concerns weighted scoring functions, where the weight function depends on the realizing observation, $\mathbf{y}$, only.

Theorem 5. Let the functional T be defined on a class $\mathcal{F}$ of probability distributions which admit a density, f, with respect to some dominating measure on the domain D. Consider a measurable weight function

$$w : \mathrm{D} \to [0, \infty).$$

Let $\mathcal{F}^{(w)} \subseteq \mathcal{F}$ denote the subclass of the probability distributions in $\mathcal{F}$ which are such that $w(\mathbf{y})f(\mathbf{y})$ has finite integral over D, and the probability measure $F^{(w)}$ with density proportional to $w(\mathbf{y})f(\mathbf{y})$ belongs to $\mathcal{F}$. Define the functional

$$\mathrm{T}^{(w)} : \mathcal{F}^{(w)} \longrightarrow \mathcal{P}(\mathrm{I}), \qquad F \longmapsto \mathrm{T}^{(w)}(F) = \mathrm{T}(F^{(w)}), \tag{9}$$

on this subclass $\mathcal{F}^{(w)}$. Then the following holds:

(a) If T is elicitable, then $\mathrm{T}^{(w)}$ is elicitable.
(b) If S is consistent for T relative to $\mathcal{F}$, then

$$\mathrm{S}^{(w)}(\mathbf{x}, \mathbf{y}) = w(\mathbf{y})\mathrm{S}(\mathbf{x}, \mathbf{y}) \tag{10}$$

is consistent for $\mathrm{T}^{(w)}$ relative to $\mathcal{F}^{(w)}$.
(c) If S is strictly consistent for T relative to $\mathcal{F}$, then $\mathrm{S}^{(w)}$ is strictly consistent for $\mathrm{T}^{(w)}$ relative to $\mathcal{F}^{(w)}$.

In other words, a weighted scoring function is consistent for the functional $\mathrm{T}^{(w)}$, which acts on the predictive distribution in a peculiar way, in that it applies the original functional, T, to the probability measure whose density is proportional to the product of the weight function and the original density. Propositions 20 and 21 of French and Ríos Insua (2000, pp. 149 and 151) concern the special cases in which S is the squared error and the absolute error scoring function, respectively.

Theorem 5 is a very general result with a wealth of applications, both in forecast evaluation and in the derivation of optimal point forecasts. In particular, the functional (9) is the optimal point forecast under the weighted scoring function (10), which allows us to unify and extend scattered prior results. For example, the scoring function S_β of equation (4),

$$\mathrm{S}_\beta(x, y) = \left| 1 - \left(\frac{y}{x} \right)^\beta \right|,$$

is of the form (10) with the original scoring function $\mathrm{S}(x, y) = |x^{-\beta} - y^{-\beta}|$ and the weight function $w(y) = y^\beta$ on the positive halfaxis, $\mathrm{D} = (0, \infty)$. The scoring function S is consistent for the median functional. Thus, as noted in the introduction, the scoring function S_β is consistent for the *β-median* functional, $\mathrm{med}^{(\beta)}(F)$, that is, the median of a probability distribution whose density is proportional to $y^\beta f(y)$, where f is the density of F. If $\beta = -1$, we recover the absolute percentage error, $\mathrm{S}_{-1}(x, y) = |(x - y)/y|$. The case $\beta = 1$ corresponds to the relative error, $\mathrm{S}_1(x, y) = |(x - y)/x|$, which Patton (2011) refers to as the MAE-prop function. Table 1 of Patton (2011) shows Monte Carlo based approximate values for optimal point forecasts under this scoring function. Theorem 5 permits us to give exact results; these are summarized in Table 8 and differ notably from the approximations. For details see Appendix B.

Another interesting case arises when the original scoring function S is the squared error, $\mathrm{S}(x, y) = (x - y)^2$, which is consistent for the mean or expectation functional. If T is the mean functional, the functional $\mathrm{T}^{(w)}$ of equation (9) becomes

$$\mathrm{T}^{(w)}(F) = \mathrm{T}(F^{(w)}) = \mathbb{E}_{F^{(w)}}[Z] = \frac{\mathbb{E}_F[Yw(Y)]}{\mathbb{E}_F[w(Y)]}. \tag{11}$$

Park and Stefanski (1998) studied optimal point forecasts in the special case in which $\mathrm{D} = (0, \infty)$ is the positive half-axis and $w(y) = 1/y^2$, so that $\mathrm{S}^{(w)}(x, y) = (x - y)^2/y^2$ is the squared percentage error. By Equation (11), the scoring function $\mathrm{S}^{(w)}$ is consistent for the functional $\mathrm{T}^{(w)}(F) = \mathbb{E}_F[Y^{-1}]/\mathbb{E}_F[Y^{-2}]$. By Theorem 1, this latter quantity is the optimal point forecast under the squared percentage error scoring function, which is the result derived by Park and Stefanski (1998).

Situations in which the weight function depends on the point forecast, x, rather than the realization, y, need to be handled on a case-by-case basis. For example, routine calculus shows that the squared relative error scoring function, $\mathrm{S}(x, y) = (x - y)^2/x^2$, is consistent for the functional

$$\mathrm{T}(F) = \frac{\mathbb{E}_F[Y^2]}{\mathbb{E}_F[Y]}. \tag{12}$$

Incidentally, as a special case of (11), the observation-weighted scoring function $\mathrm{S}(x, y) = y(x - y)^2$ is also consistent for the functional (12). Later on in Equation (23) we characterize the class of the scoring functions that are consistent for this functional.

While Theorems 4 and 5 suggest that general classes of functionals are elicitable, not all functionals are such. The following result, which is a variant of proposition 2.5 of Osband (1985) and lemma 1 of Lambert, Pennock, and Shoham (2008), states a necessary condition.

Theorem 6 (Osband). If a functional is elicitable then its level sets are convex in the following sense: If $F_0 \in \mathcal{F}$, $F_1 \in \mathcal{F}$ and $p \in (0, 1)$ are such that $F_p = (1 - p)F_0 + pF_1 \in \mathcal{F}$, then $\mathbf{t} \in \mathrm{T}(F_0)$ and $\mathbf{t} \in \mathrm{T}(F_1)$ imply $\mathbf{t} \in \mathrm{T}(F_p)$.

For example, the sum of two distinct quantiles generally does not have convex level sets and thus is not an elicitable functional. Interesting open questions include those for a converse of Theorem 6 and, more generally, for a characterization of elicitability.

Table 8. The optimal point forecast or Bayes rule (6) when the scoring function is relative error, $\mathrm{S}(x, y) = |(x - y)/x|$, and the future quantity Y can be represented as $Y = Z^2$, where Z has a t-distribution with mean 0, variance 1 and $\nu > 2$ degrees of freedom. In the limiting case as $\nu \to \infty$, we take Z to be standard normal. If Z has variance σ^2 the entries need to be multiplied by this factor. As opposed to the approximations in table 1 of Patton (2011), which stem from numerical and Monte Carlo methods and are reproduced below, our results derive from Theorem 5 and are exact. For details see Appendix B

	$\nu = 4$	$\nu = 6$	$\nu = 8$	$\nu = 10$	$\nu \to \infty$
Exact optimal point forecast	3.4048	2.8216	2.6573	2.5801	2.3660
Patton's approximation	3.0962	2.7300	2.6067	2.5500	2.3600

2.4 Osband's Principle

If a scoring function S exists that is consistent for the functional T, a natural question is whether it might be unique, up to the inessential transformations described by Granger and Machina (2006). The following approach, which originates in the pioneering work of Osband (1985), provides a practical way of describing and characterizing the class of the scoring functions that are consistent for a given functional. Frequently, the resulting characterizations link to monotonicity and convexity.

Suppose that the functional T is defined for a class of probability measures on the domain D which includes the two-point distributions. Assume that there exists an *identification function* $\mathrm{V}:\mathrm{D}\times\mathrm{D}\to\mathbb{R}$ such that

$$\mathbb{E}_F[\mathrm{V}(\mathbf{x},\mathbf{Y})]=0 \qquad \Longleftrightarrow \qquad \mathbf{x}\in\mathrm{T}(F) \tag{13}$$

and $\mathrm{V}(\mathbf{x},\mathbf{y})\neq 0$ unless $\mathbf{x}=\mathbf{y}$. If a consistent scoring function is available, which is smooth in its first argument, we can take $\mathrm{V}(\mathbf{x},\mathbf{y})$ to be the corresponding partial derivative or gradient. For example, if T is the mean or expectation functional on an interval $\mathrm{D}=\mathrm{I}\subseteq\mathbb{R}$, we can pick $\mathrm{V}(x,y)=x-y$, which derives from the squared error scoring function, $\mathrm{S}(x,y)=(x-y)^2$. Table 9 provides further examples.

The function

$$\epsilon(\mathbf{c})=p\mathrm{S}(\mathbf{c},\mathbf{a})+(1-p)\mathrm{S}(\mathbf{c},\mathbf{b}) \tag{14}$$

represents the expected score when we issue the point forecast $\mathbf{c}$ for a random vector $\mathbf{Y}$ such that $\mathbf{Y}=\mathbf{a}$ with probability p and $\mathbf{Y}=\mathbf{b}$ with probability $1-p$. Since S is consistent for the functional T, the identification function property (13) implies that $\epsilon(\mathbf{c})$ has a minimum at $\mathbf{c}=\mathbf{x}$, where

$$p\mathrm{V}(\mathbf{x},\mathbf{a})+(1-p)\mathrm{V}(\mathbf{x},\mathbf{b})=0. \tag{15}$$

If S is smooth in its first argument, we can combine (14) and (15) to result in

$$\partial_{\mathbf{x}}\mathrm{S}(\mathbf{x},\mathbf{a})/\mathrm{V}(\mathbf{x},\mathbf{a})=\partial_{\mathbf{x}}\mathrm{S}(\mathbf{x},\mathbf{b})/\mathrm{V}(\mathbf{x},\mathbf{b}), \tag{16}$$

where $\partial_{\mathbf{x}}\mathrm{S}(\mathbf{x},\mathbf{y})$ denotes a partial derivative or gradient with respect to the first argument. If this latter equality holds for all pairwise distinct $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{x}\in\mathrm{D}$, the function $\partial_{\mathbf{x}}\mathrm{S}(\mathbf{x},\mathbf{y})/\mathrm{V}(\mathbf{x},\mathbf{y})$ is independent of $\mathbf{y}\in\mathrm{D}$ and we can write

$$\partial_{\mathbf{x}}\mathrm{S}(\mathbf{x},\mathbf{y})=h(\mathbf{x})\mathrm{V}(\mathbf{x},\mathbf{y}) \tag{17}$$

for $\mathbf{x},\mathbf{y}\in\mathrm{D}$ and some function $h:\mathrm{D}\to\mathrm{D}$. Frequently, we can integrate (17) to obtain the general form of a scoring rule that is consistent for the functional T.

In recognition of Osband's (1985) fundamental yet unpublished work, we refer to this general approach as *Osband's principle*. The examples in the subsequent section give various instances in which the principle can be successfully put to work. For a general technical result, see theorem 2.1 of Osband (1985).

Table 9. Possible choices for the identification function V with the property (13) in the case in which $\mathrm{D}=\mathrm{I}\subseteq\mathbb{R}$ is an interval

Functional	Identification function
Mean	$\mathrm{V}(x,y)=x-y$
Ratio $\mathbb{E}_F[r(Y)]/\mathbb{E}_F[s(Y)]$	$\mathrm{V}(x,y)=xs(y)-r(y)$
α-Quantile	$\mathrm{V}(x,y)=\mathbb{1}(x\geq y)-\alpha$
τ-Expectile	$\mathrm{V}(x,y)=2\lvert\mathbb{1}(x\geq y)-\tau\rvert(x-y)$

3. EXAMPLES

We now give examples in the case of a univariate predictand, in which any connected domain $\mathrm{D}=\mathrm{I}\subseteq\mathbb{R}$ is an interval. Some of the results are classical, such as the characterizations for expectations (Savage 1971) and quantiles (Thomson 1979), and some are novel, including those for ratios of expectations, expectiles and conditional value-at-risk. In most examples, the technical arguments rely on the properties of convex functions and subgradients, for which we refer to Rockafellar (1970).

3.1 Expectations

It is well known that the squared error scoring function, $\mathrm{S}(x,y)=(x-y)^2$, is strictly consistent for the mean functional relative to the class of the probability distributions on $\mathbb{R}$ whose second moment is finite. Thus, means or expectations are elicitable. Before turning to more general settings in subsequent sections, we review a classical result of Savage (1971) which identifies the class of scoring functions that are consistent for the mean functional as that of Bregman functions. Closely related results have been obtained by Reichelstein and Osband (1984), Saerens (2000), Banerjee, Guo, and Wang (2005) and Patton (2011).

Theorem 7 (Savage). Let $\mathcal{F}$ be the class of probability measures on the interval $\mathrm{I}\subseteq\mathbb{R}$ with finite first moment. Then the following holds:

(a) The mean functional is elicitable relative to the class $\mathcal{F}$.
(b) Suppose that the scoring function S satisfies assumptions (S0), (S1), and (S2) on the PO domain $\mathcal{D}=\mathrm{I}\times\mathrm{I}$. Then S is consistent for the mean functional relative to the class of the compactly supported probability measures on I if, and only if, it is of the form
$$\mathrm{S}(x,y)=\phi(y)-\phi(x)-\phi'(x)(y-x), \tag{18}$$
where ϕ is a convex function with subgradient ϕ' on I.
(c) If ϕ is strictly convex, the scoring function (18) is strictly consistent for the mean functional relative to the class of the probability measures F on I for which both $\mathbb{E}_F Y$ and $\mathbb{E}_F\phi(Y)$ exist and are finite.

Banerjee, Guo, and Wang (2005) refer to a function of the form (18) as a *Bregman function*. For example, if $\mathrm{I}=\mathbb{R}$ and $\phi(x)=|x|^a$, where $a>1$ to ensure strict convexity, the Bregman representation yields the scoring function

$$\mathrm{S}_a(x,y)=|y|^a-|x|^a-a\,\mathrm{sign}(x)|x|^{a-1}(y-x), \tag{19}$$

which is homogeneous of order a and nests the squared error that arises when $a=2$. Savage (1971) showed that, up to a multiplicative constant, the squared error is the unique Bregman function of the prediction error form, as well as the unique symmetric Bregman function. Patton (2011) introduced a rich and flexible family of homogeneous Bregman functions on the PO domain $\mathcal{D}=(0,\infty)\times(0,\infty)$, namely

$$\mathrm{S}_b(x,y)=\begin{cases}\dfrac{1}{b(b-1)}(y^b-x^b) & \\ \quad -\dfrac{1}{b-1}x^{b-1}(y-x) & \text{if } b\in\mathbb{R}\setminus\{0,1\}\\ \dfrac{y}{x}-\log\dfrac{y}{x}-1 & \text{if } b=0\\ y\log\dfrac{y}{x}-y+x & \text{if } b=1.\end{cases} \tag{20}$$

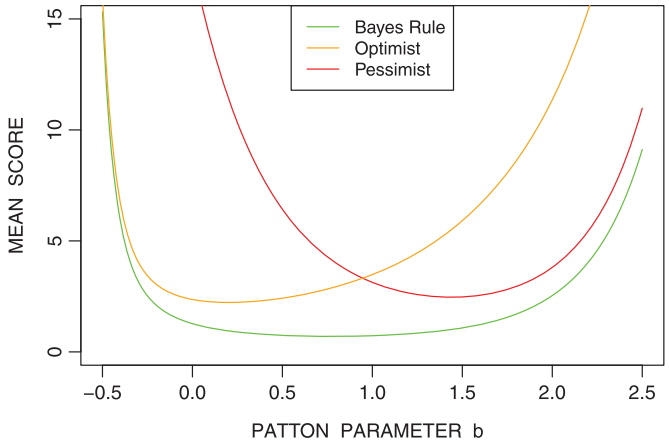

Figure 2. The mean score (1) under the Patton scoring function (20) for Mr. Bayes (green), who uses the Bayes rule, the optimist (orange) and the pessimist (red) in the simulation study of Section 1.2. The online version of this figure is in color.

Up to a multiplicative constant, these are the only homogeneous Bregman functions on this PO domain. The squared error scoring function emerges when $b = 2$ and the QLIKE function (Patton 2011) when $b = 0$. If $b = a > 1$ the Patton function (20) coincides with the corresponding restriction of the power function (19), up to a multiplicative constant.

Finally, it is worth noting that proper scoring rules for probability forecasts of a dichotomous event are also of the Bregman form, because the probability of a binary event equals the expectation of the corresponding indicator variable. Compare Savage (1971), DeGroot and Fienberg (1983), Schervish (1989), Winkler (1996), Buja, Stuetzle, and Shen (2005), and Gneiting and Raftery (2007), among others.

Figure 2 returns to the initial simulation study of Section 1.2 and shows the mean score (1) under the Patton scoring function (20) for Mr. Bayes, the optimist and the pessimist. The optimal point forecast under a Bregman scoring function is the mean of the predictive distribution, so that the statistician forecaster fuses with Mr. Bayes.

3.2 Ratios of Expectations

We now consider statistical functionals which can be represented as ratios of expectations, and arise naturally in the context of weighted scoring functions, as exemplified in Section 2.3. The mean functional emerges in the special case in which $r(y) = y$ and $s(y) = 1$.

Theorem 8. Let $\mathrm{I} \subseteq \mathbb{R}$ be an interval, and suppose that $r : \mathrm{I} \to \mathbb{R}$ and $s : \mathrm{I} \to (0, \infty)$ are measurable functions. Then the following holds:

(a) The functional

$$\mathrm{T}(F) = \frac{\mathbb{E}_F[r(Y)]}{\mathbb{E}_F[s(Y)]} \tag{21}$$

is elicitable relative to the class of the probability measures on I for which $\mathbb{E}_F[r(Y)]$, $\mathbb{E}_F[s(Y)]$ and $\mathbb{E}_F[Ys(Y)]$ exist and are finite.

(b) If S is of the form

$$\mathrm{S}(x, y) = s(y)(\phi(y) - \phi(x)) - \phi'(x)(r(y) - xs(y)) + \phi'(y)(r(y) - ys(y)), \tag{22}$$

where ϕ is a convex function with subgradient ϕ', then it is consistent for the functional (21) relative to the class of the probability measures F on I for which $\mathbb{E}_F[r(Y)]$, $\mathbb{E}_F[s(Y)]$, $\mathbb{E}_F[r(Y)\phi'(Y)]$, $\mathbb{E}_F[s(Y)\phi(Y)]$, and $\mathbb{E}_F[Ys(Y)\phi'(Y)]$ exist and are finite. If ϕ is strictly convex, then S is strictly consistent.

(c) Suppose that the scoring function S satisfies assumptions (S0), (S1), and (S2) on the PO domain $\mathcal{D} = \mathrm{I} \times \mathrm{I}$. If s is continuous and $r(y) = ys(y)$ for $y \in \mathrm{I}$, then S is consistent for the functional (21) relative to the class of the compactly supported probability measures on I if, and only if, it is of the form (22), where ϕ is a convex function with subgradient ϕ'.

In the case in which $s(y) = w(y)$ and $r(y) = yw(y)$ for a strictly positive, continuous weight function w, the ratio (21) coincides with the functional (11). If $\mathrm{I} = (0, \infty)$ and $w(y) = y$, the special case $\mathrm{T}(F) = \mathbb{E}_F[Y^2]/\mathbb{E}_F[Y]$ of Equation (12) arises. In Section 2.3 we saw that both the squared relative error scoring function, $\mathrm{S}(x, y) = (x - y)^2/x^2$, and the observation-weighted scoring function $\mathrm{S}(x, y) = y(x - y)^2$ are consistent for this functional. By part (c) of Theorem 8, the general form of a scoring function that is consistent for the functional (12) is

$$\mathrm{S}(x, y) = y(\phi(y) - \phi(x)) - y(y - x)\phi'(x), \tag{23}$$

where ϕ is convex with subgradient ϕ'. The above scoring functions emerge when $\phi(y) = 1/y$ and $\phi(y) = y^2$, respectively.

3.3 Quantiles and Expectiles

An α-quantile $(0 < \alpha < 1)$ of the cumulative distribution function F is any number x for which $\lim_{y \uparrow x} F(y) \le \alpha \le F(x)$. In finance, quantiles are often referred to as *value-at-risk* (VaR; Duffie and Pan 1997). The literature on the evaluation of quantile forecasts generally recommends the use of the *asymmetric piecewise linear* scoring function,

$$\mathrm{S}_\alpha(x, y) = (\mathbb{1}(x \ge y) - \alpha)(x - y) = \begin{cases} \alpha|x - y|, & x \le y \\ (1 - \alpha)|x - y|, & x \ge y, \end{cases} \tag{24}$$

which is strictly consistent for the α-quantile relative to the class of the probability measures with finite first moment (Raiffa and Schlaifer 1961, p. 196; Ferguson 1967, p. 51). This well-known property lies at the heart of quantile regression (Koenker and Bassett 1978).

As regards the characterization of the scoring functions that are consistent for a quantile, results of Thomson (1979) and Saerens (2000) can be summarized as follows. For a discussion of their equivalence and historical comments, see Gneiting (2011).

Theorem 9 (Thomson, Saerens). Let $\mathcal{F}$ be the class of the probability measures on the interval $\mathrm{I} \subseteq \mathbb{R}$, and let $\alpha \in (0, 1)$. Then the following holds:

(a) The α-quantile functional is elicitable relative to the class $\mathcal{F}$.

(b) Suppose that the scoring function S satisfies assumptions (S0), (S1), and (S2) on the PO domain $\mathcal{D} = \mathrm{I} \times \mathrm{I}$. Then S is consistent for the α-quantile relative to the class of the compactly supported probability measures on I if, and only if, it is of the form

$$\mathrm{S}(x, y) = (\mathbb{1}(x \geq y) - \alpha)(g(x) - g(y)), \tag{25}$$

where g is a nondecreasing function on I.

(c) If g is strictly increasing, the scoring function (25) is strictly consistent for the α-quantile relative to the class of the probability measures F on I for which $\mathbb{E}_F g(Y)$ exists and is finite.

Gneiting (2008b) refers to a function of the form (25) as *generalized piecewise linear* (GPL) of order $\alpha \in (0, 1)$, because it is piecewise linear after applying a nondecreasing transformation. Any GPL function is equivariant with respect to the class of the nondecreasing transformations, just as the quantile functional is equivariant under monotone mappings (Koenker 2005, p. 39). If $\mathrm{I} = (0, \infty)$ and $g(x) = x^b/|b|$ for $b \in \mathbb{R} \setminus \{0\}$, and taking the corresponding limit as $b \to 0$, we obtain the family

$$\mathrm{S}_{\alpha,b}(x, y) = \begin{cases} (\mathbb{1}(x \geq y) - \alpha)\dfrac{1}{|b|}(x^b - y^b) & \text{if } b \in \mathbb{R} \setminus \{0\} \\ (\mathbb{1}(x \geq y) - \alpha)\log\dfrac{x}{y} & \text{if } b = 0, \end{cases} \tag{26}$$

of the GPL power scoring functions, which are homogeneous of order b. The asymmetric piecewise linear function (24) arises when $b = 1$, and the MAE-LOG and MAE-SD functions described by Patton (2011) emerge when $\alpha = \frac{1}{2}$, and $b = 0$ and $b = \frac{1}{2}$, respectively. The loss functions proposed by Lopez (1999) and Caporin (2008) for evaluating VaR forecasts are not of the GPL form. Thus, Theorem 9 provides a theoretical explanation of the empirical finding that these loss or scoring functions "in many cases are not able to properly distinguish between the true data generating process and alternative VaR models" (Caporin 2008, p. 80). Figure 3 returns to the simulation study in Section 1.2 and shows the mean score (1) under the GPL power function (26), where $\alpha = \frac{1}{2}$, for Mr. Bayes, the statistician, the optimist and the pessimist. Once again, Mr. Bayes dominates his competitors.

Newey and Powell (1987) introduced the τ-*expectile* functional ($0 < \tau < 1$) of a probability measure F with finite mean as the unique solution $x = \mu_\tau$ to the equation

$$\tau \int_x^\infty (y - x)\,\mathrm{d}F(y) = (1 - \tau)\int_{-\infty}^x (x - y)\,\mathrm{d}F(y).$$

If the second moment of F is finite, the τ-expectile equals the Bayes rule or optimal point forecast (6) under the *asymmetric piecewise quadratic* scoring function,

$$\mathrm{S}_\tau(x, y) = |\mathbb{1}(x \geq y) - \tau|(x - y)^2, \tag{27}$$

similar to the α-quantile being the Bayes rule under the asymmetric piecewise linear function (24). Not surprisingly, expectiles have properties that resemble those of quantiles.

The following original result characterizes the class of the scoring functions that are consistent for expectiles. It is interesting to observe the ways in which the corresponding class (28) combines key characteristics of the Bregman and GPL families.

Theorem 10. Let $\mathcal{F}$ be the class of the probability measures on the interval $\mathrm{I} \subseteq \mathbb{R}$ with finite first moment, and let $\tau \in (0, 1)$. Then the following holds:

(a) The τ-expectile functional is elicitable relative to the class $\mathcal{F}$.

(b) Suppose that the scoring function S satisfies assumptions (S0), (S1), and (S2) on the PO domain $\mathcal{D} = \mathrm{I} \times \mathrm{I}$. Then S is consistent for the τ-expectile relative to the class of the compactly supported probability measures on I if, and only if, it is of the form

$$\mathrm{S}(x, y) = |\mathbb{1}(x \geq y) - \tau| \times (\phi(y) - \phi(x) - \phi'(x)(y - x)), \tag{28}$$

where ϕ is a convex function with subgradient ϕ' on I.

(c) If ϕ is strictly convex, the scoring function (28) is strictly consistent for the τ-expectile relative to the class of the probability measures F on I for which both $\mathbb{E}_F Y$ and $\mathbb{E}_F \phi(Y)$ exist and are finite.

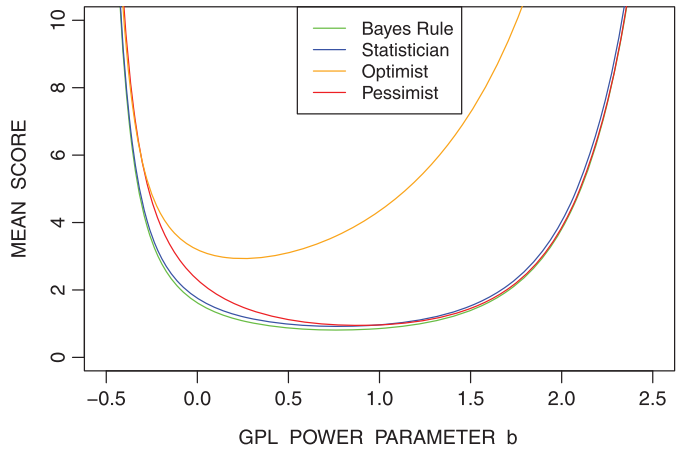

Figure 3. The mean score (1) under the GPL power scoring function (26) with $\alpha = \frac{1}{2}$ for Mr. Bayes (green), who uses the Bayes rule, the statistician (blue), the optimist (orange) and the pessimist (red) in the simulation study of Section 1.2. The online version of this figure is in color.

3.4 Conditional Value-at-Risk

The *α-conditional value-at-risk* functional (CVaR_α, $0 < \alpha < 1$) equals the expectation of a random variable with distribution F conditional on it taking values in its upper $(1-\alpha)$-tail (Rockafellar and Uryasev 2000, 2002). An often convenient, equivalent definition is

$$\mathrm{CVaR}_\alpha(F) = \frac{1}{1-\alpha}\int_\alpha^1 q_\beta(F)\,\mathrm{d}\beta,$$

where q_β denotes the β-quantile (Acerbi 2002). The CVaR functional is a popular risk measure in quantitative finance. Its varied, elegant and appealing properties include coherency in the sense of Artzner et al. (1999), who consider functionals defined in terms of random variables, rather than the corresponding probability measures. However, it is not elicitable.

Theorem 11. The CVaR_α functional is not elicitable relative to any class $\mathcal{F}$ of probability distributions on the interval $\mathrm{I} \subseteq \mathbb{R}$ that contains the measures with finite support, or the finite mixtures of the absolutely continuous distributions with compact support.

This negative result may challenge the use of the CVaR functional as a predictive measure of risk, and may provide a partial explanation for the lack of literature on the evaluation of CVaR forecasts, as opposed to quantile or VaR forecasts, for which we refer to Berkowitz and O'Brien (2002), Giacomini and Komunjer (2005), and Bao, Lee, and Saltoğlu (2006), among others. Christoffersen (2003) suggests a way of evaluating CVaR forecasts that depends on external covariates and allows for tests, but not for a direct comparison and ranking of the predictive performance of competing forecasting methods. As Theorem 11 shows, scoring functions that are consistent for the CVaR functional are not available, whence it remains unclear how such comparisons could be done. For further discussion, see Osband (2011, p. 246).

3.5 Mode

Let $\mathcal{F}$ be a class of probability measures on the real line, each of which has a well-defined, unique mode. It is sometimes stated informally that the mode is an optimal point forecast under the zero-one scoring function,

$$\mathrm{S}_c(x, y) = \mathbb{1}(|x-y| > c),$$

where $c > 0$. A rigorous statement is that the optimal point forecast or Bayes rule (6) under the scoring function S_c is the *midpoint*

$$\hat{x} = \arg\max_x \Big(F(x+c) - \lim_{y\uparrow x-c} F(y)\Big)$$

of the *modal interval* of length $2c$ of the probability measure $F \in \mathcal{F}$ (Ferguson 1967, p. 51). Example 7.20 of Lehmann and Casella (1998) explores this argument in more detail.

Expressed differently, the zero-one scoring function S_c is consistent for the midpoint functional, which we denote by T_c. If c is sufficiently small, then $\mathrm{T}_c(F)$ is well defined and single-valued for all $F \in \mathcal{F}$. We can then define the *mode functional* on $\mathcal{F}$ as the limit

$$\mathrm{T}_0(F) = \lim_{c\downarrow 0} \mathrm{T}_c(F),$$

with the corresponding limit of the (suitably scaled) scoring function S_c being a Dirac delta function. Essentially identical arguments have been used by Bernardo and Smith (2000, p. 257) and French and Ríos Insua (2000, pp. 151–152).

Stronger results become available if one puts conditions on both the scoring function S and the family $\mathcal{F}$ of probability distributions. Theorem 2 of Granger (1969) is a result of this type. Consider the PO domain $\mathcal{D} = \mathbb{R} \times \mathbb{R}$. If the scoring function S is an even function of the prediction error that attains a minimum at the origin, and each $F \in \mathcal{F}$ admits a Lebesgue density, f, which is symmetric, continuous and unimodal, so that mean, median and mode coincide, then S is consistent for this common functional. Theorem 1 of Granger (1969) and theorem 7.15 of Lehmann and Casella (1998) trade the continuity and unimodality conditions on f for an additional assumption of convexity on the scoring function.

Henderson, Jones, and Stare (2001, p. 3087) posit that in survival analysis a loss function of the form

$$\mathrm{S}_k^*(x, y) = \begin{cases} 0, & \frac{x}{k} \le y \le kx \\ 1, & \text{otherwise} \end{cases} = \mathbb{1}(|\log x - \log y| > \log k)$$

is reasonable, with a choice of $k = 2$ often being adequate, arguing that "most people for example would accept that a lifetime prediction of, say, 2 months, was reasonably accurate if death occurs between about 1 and 4 months." The optimal point forecast or Bayes rule under S_k^* is the midpoint functional $\mathrm{T}_{\log(k)}$ applied to the predictive distribution of the logarithm of the lifetime, rather than the lifetime itself.

4. MULTIVARIATE PREDICTANDS

While thus far we have restricted attention to point forecasts of a univariate quantity, the general case of a multivariate predictand that takes values in a domain $\mathrm{D} \subseteq \mathbb{R}^d$ is of considerable interest. Applications include those of Gneiting et al. (2008) and Hering and Genton (2010) to predictions of wind vectors, or that of Laurent, Rombouts, and Violante (2009) to forecasts of multivariate volatility, to name but a few. We turn to the decision-theoretic setting of Section 2.1 and assume, for simplicity, that the point forecast, the observation and the target functional take values in $\mathrm{D} = \mathbb{R}^d$.

We first discuss the mean functional. Assuming that $\mathrm{S}(\mathbf{x}, \mathbf{y}) \ge 0$ with equality if $\mathbf{x} = \mathbf{y}$, Savage (1971), Osband and Reichelstein (1985), and Banerjee, Guo, and Wang (2005) showed that a scoring function under which the (component-wise) expectation of the predictive distribution is an optimal point forecast, is of the Bregman form

$$\mathrm{S}(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{y}) - \phi(\mathbf{x}) - \langle \nabla\phi(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle, \tag{29}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex with gradient $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ and $\langle\cdot,\cdot\rangle$ denotes a scalar product, subject to smoothness conditions. Expressed differently, a sufficiently smooth scoring function is consistent for the mean functional if and only if it is of the form (29). When $\phi(\mathbf{x}) = \|\mathbf{x}\|^2$ is the squared Euclidean norm, we obtain the squared error scoring function, and similarly its ramifications, such as the weighted squared error and the pseudo Mahalanobis error (Bernardo and Smith 2000, p. 257; Laurent, Rombouts, and Violante 2009).

It is of interest to note that rigorous versions of the Bregman characterization depend on restrictive smoothness conditions. Osband and Reichelstein (1985) assume that the scoring function is continuously differentiable with respect to its first argument, the point forecast; Banerjee, Guo, and Wang (2005) assume the existence of continuous second partial derivatives with respect to the observation. A challenging, nontrivial problem is to unify and strengthen these results, both in univariate and multivariate settings, and relative to the notion of k-elicitability of integer order $k \geq 2$, as defined and studied by Lambert, Pennock, and Shoham (2008). For example, while the univariate variance functional is not elicitable, it is two-elicitable, because $\mathrm{var}_F(X) = \mathbb{E}_F[X^2] - (\mathbb{E}_F[X])^2$, and both $\mathbb{E}_F[X]$ and $\mathbb{E}_F[X^2]$ are elicitable.

Laurent, Rombouts, and Violante (2009) consider point forecasts of multivariate stochastic volatility, where the predictand is a symmetric and positive definite matrix in $\mathbb{R}^{q\times q}$. If the matrix is vectorized, the above results for the mean functional apply, thereby leading to the Bregman representation (29) for the respective consistent scoring functions, which is hidden in proposition 3 of Laurent, Rombouts, and Violante (2009). Corollary 1 of Laurent, Rombouts, and Violante (2009) supplies a version that applies directly to point forecasts, say $\mathbf{\Sigma}_\mathbf{x} \in \mathbb{R}^{q\times q}$, of a matrix-valued, symmetric and positive definite quantity, say $\mathbf{\Sigma}_\mathbf{y} \in \mathbb{R}^{q\times q}$, without any need to resort to vectorization. Specifically, any scoring function of the form

$$\mathrm{S}(\mathbf{\Sigma}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{y}) = \phi(\mathbf{\Sigma}_\mathbf{y}) - \phi(\mathbf{\Sigma}_\mathbf{x}) - \mathrm{tr}(\nabla_0\phi(\mathbf{\Sigma}_\mathbf{x})(\mathbf{\Sigma}_\mathbf{y} - \mathbf{\Sigma}_\mathbf{x})) \quad (30)$$

is consistent for the (component-wise) mean functional, where ϕ is convex and smooth, and $\nabla_0\phi$ denotes a symmetric matrix of first partial derivatives, with the off-diagonal elements multiplied by a factor of one half.

Dawid and Sebastiani (1999) and Pukelsheim (2006) give various examples of convex functions ϕ whose domain is the convex cone of the symmetric and positive definite elements of $\mathbb{R}^{q\times q}$, with the matrix norm

$$\phi(\mathbf{\Sigma}) = \left(\frac{1}{q}\mathrm{tr}(\mathbf{\Sigma}^s)\right)^{1/s} \quad (31)$$

for integers $s \geq 1$ being one such instance. The matrix norm is nonnegative, nondecreasing in the Loewner order, continuous, convex, standardized and homogeneous of order one. With simple adaptations, the construction extends to any real or extended real-valued exponent s and to general, not necessarily positive definite symmetric matrices (Pukelsheim 2006, pp. 141 and 151). In the limit as $s \to 0$ in (31) the log determinant $\phi(\mathbf{\Sigma}) = \log\det(\mathbf{\Sigma})$ emerges. When used in the Bregman representation (30), the log determinant function gives rise to a well known homogeneous scoring function for point predictions of a positive definite symmetrically matrix-valued quantity in $\mathbb{R}^{q\times q}$, namely,

$$\mathrm{S}(\mathbf{\Sigma}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{y}) = \mathrm{tr}(\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{\Sigma}_\mathbf{y}) - \log\det(\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{\Sigma}_\mathbf{y}) - q, \quad (32)$$

which was introduced by James and Stein (1961, section 5).

In the case of quantiles, the passage from the univariate functional to multivariate analogues is much less straightforward. Notions of quantiles for multivariate distributions based on loss or scoring functions have been studied by Abdous and Theodorescu (1992), Chaudhuri (1996), Koltchinskii (1997), Serfling (2002), and Hallin, Paindaveine, and Šiman (2010), among others. In particular, it is customary to define the *median* of a probability distribution F on $\mathbb{R}^d$ as

$$\hat{\mathbf{x}} = \arg\min_\mathbf{x} \mathbb{E}_F(\|\mathbf{x} - \mathbf{Y}\| - \|\mathbf{Y}\|),$$

where $\|\cdot\|$ denotes the Euclidean norm (Small 1990). If $d = 1$, this yields the traditional median on the real line, with the final term eliminating the need for moment conditions on the predictive distribution (Kemperman 1987). Of course, norms and distances other than the Euclidean could be considered. In this more general type of situation, Koenker (2006) proposed that a functional based on minimizing the square of a distance be called a *Fréchet mean*, and a functional based on minimizing a distance a *Fréchet median*, just as in the traditional case of the Euclidean distance.

5. DISCUSSION

Ideally, forecasts ought to be probabilistic, taking the form of predictive distributions over future quantities and events (Dawid 1984; Diebold, Gunther, and Tay 1998; Granger and Pesaran 2000a, 2000b; Geweke and Whiteman 2006; Gneiting 2008a). 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.

Our plea for the use of consistent scoring functions supplements and qualifies, but does not contradict, extant recommendations in the forecasting literature, such as those of West and Harrison (1997), Armstrong (2001), Jolliffe and Stephenson (2003), and Fildes and Goodwin (2007). For example, Fildes and Goodwin (2007) propose forecasting principles for organizations, the eleventh of which suggests that "multiple measures of forecast accuracy" be employed. I agree, with the qualification that the scoring functions to be used be consistent for the target functional.

We have developed theory for the notions of consistency and elicitability, and have characterized the classes of the loss or scoring functions that result in expectations, ratios of expectations, quantiles or expectiles as optimal point forecasts. Some of these results are classical, such as those for means and quantiles (Savage 1971; Thomson 1979), while others are original, including a negative result, in that scoring functions which are consistent for the CVaR functional do not exist.

In the case of the mean functional, the consistent scoring functions are the Bregman functions of the form (18). Among these, a particularly attractive choice is the Patton family (20) of homogeneous scoring functions, which nests the squared error (SE) and QLIKE functions. In evaluating volatility forecasts, Patton and Sheppard (2009) recommend the use of the latter because of its superior power in Diebold and Mariano (1995) and West (1996) tests of predictive ability, which depend on differences between mean scores of the form (1) as test statistics. Further work in this direction is desirable, both empirically and theoretically. If quantile forecasts are to be assessed, the consistent scoring functions are the GPL functions of the form (25), with the homogeneous power functions in (26) being appealing

examples. Interestingly, the scoring functions that are consistent for expectiles combine key elements of the Bregman and GPL families.

As regards the most commonly used scoring functions in academia, businesses and organizations, the squared error scoring function is consistent for the mean, and the absolute error scoring function for the median. The absolute percentage error scoring function, which is commonly used by businesses and organizations, and occasionally in academia, is consistent for a nonstandard functional, namely, the median of order -1, $\text{med}^{(-1)}$, which tends to support severe underforecasts, as compared to the mean or median. It thus seems prudent that businesses and organizations consider the intended or unintended consequences and reassess its suitability as a scoring function, similarly to the practices described by Yelland, Kim, and Stratulate (2010). In this context, the recent proposal of Chen et al. (2010), who suggest the use of the sum of the absolute percentage error and the relative error as a readily interpretable, scale free loss or scoring function, merits further study.

Pers et al. (2009) propose a game of prediction for a fair comparison between competing predictive models, which employs proper scoring rules to assess full predictive distributions. As Theorem 3 shows, consistent scoring functions can also be interpreted as proper scoring rules. Hence, the protocol of Pers et al. (2009) applies to the evaluation of point forecasting methods as well. Their focus is on the comparison of custom-built predictive models for a specific purpose, as opposed to the M-competitions in the forecasting literature (Makridakis and Hibon 1979, 2000; Makridakis et al. 1982, 1993), which compare the predictive performance of point forecasting methods across multiple, unrelated time series. In this latter context, additional considerations arise, such as the comparability of scores across time series with realizations of differing magnitude and volatility, and commonly used evaluation methods remain controversial (Armstrong and Collopy 1992; Fildes 1992; Ahlburg et al. 1992; Hyndman and Koehler 2006; Athanasopoulos et al. 2011).

The notions of consistency and elicitability apply to point forecast competitions, where participants ought to be advised ex ante about the scoring function(s) to be employed, or, alternatively, target functional(s) ought to be named. If multiple target functionals are named, participants can enter possibly distinct point forecasts for distinct functionals. Similarly, if multiple scoring functions are to be used in the evaluation, and the scoring functions are consistent for distinct functionals, participants ought to be allowed to submit possibly distinct point forecasts.

Sometimes, skill scores of the form

$$\bar{\mathrm{S}}_{\text{skill}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{S}(x_i, y_i)\right) \Big/ \left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{S}(r_i, y_i)\right), \qquad (33)$$

or linear transformations thereof, are considered, where S is a fixed scoring function, $x_1, \ldots, x_n$ refer to the point forecast at hand, and $r_1, \ldots, r_n$ to a simple reference technique, such as a no-change or persistence forecast (Jolliffe and Stephenson 2003, sections 1.4.1 and 2.7). For instance, the mean absolute scaled error (MASE) of Hyndman and Koehler (2006) is of the form (33), where S is the absolute error scoring function, and the reference forecast r_i is a one-step-ahead persistence forecast in a time series setting. While it is difficult to design expectation optimizing strategies under skill scores (Murphy 1973), the original score and the skill score incur the same ranking of the forecasting methods, which suggests that the notions of consistency and elicitability continue to apply, at least when the test sample size n is large. To give an example, the MASE incurs the same ranking as the mean absolute error (MAE), and thus suggests medians as point forecasts. The choice of the reference method requires care, to avoid spurious beliefs of predictability, as illustrated by Gneiting and Thorarinsdottir (2010) in the context of inflation forecasts.

Medians or trimmed or Winsorized means of individual scores have also been used to assess predictive performance (Armstrong 2001; Hyndman and Koehler 2006). This type of performance measure ignores or downweighs individual catastrophic forecast failures, which may or may not be desirable in practice, depending on the application at hand. There is no easy way to translate our analysis to such a setting, and theoretical approaches remain scarce, with the recent work of Yu (2009) being an exception.

While thus far we have addressed forecasting or prediction problems, similar issues arise when the goal is estimation. Bayesian approaches lead naturally to posterior predictive distributions, from which any desired functional can readily be computed, and be used as a point estimate (Zellner 1986). In contrast, it might be hard to reconcile non-Bayesian estimation procedures with the decision-theoretic duality of loss functions and functionals, particularly when penalization and cross-validation interact. However, non-Bayesian approaches may allow for Bayesian interpretations under suitable prior specifications, thus alleviating these concerns. For example, lasso estimates for regression parameters can be interpreted as posterior mode estimates under independent Laplace priors, and slight variations lead to Bayesian versions of related estimation methods, such as ridge regression (Hoerl and Kennard 1970; Park and Casella 2008).

Technically, our discussion connects to M-estimation (Huber 1964; Huber and Ronchetti 2009) and quasi-likelihood (Wedderburn 1974; Dawid and Lauritzen 2006). A century ago Keynes (1911, p. 325) derived the Bregman representation (18) in characterizing the probability density functions for which the "most probable value" is the arithmetic mean. For a contemporary perspective in terms of maximum likelihood and M-estimation, see Klein and Grottke (2008). Komunjer (2005) applied the GPL class (25) in conditional quantile estimation, in generalization of the traditional approach to quantile regression, which is based on the asymmetric piecewise linear scoring function (Koenker and Bassett 1978). Similarly, Bregman functions of the original form (18) and the asymmetric variant (28) could be employed in generalizing symmetric and asymmetric least squares regression, both with and without penalization (Zhang, Jiang, and Chai 2010).

In applied settings, the distinction between prediction and estimation is frequently blurred. For example, Shipp and Cohen (2009) report on U.S. Census Bureau plans for evaluating population estimates against the results of the 2010 Census. Five measures of accuracy are to be used to assess the Census Bureau estimates, including the root mean squared error (SE) and the mean absolute percentage error (APE). Our results demonstrate that Census Bureau scientists face an impossible task in

designing procedures and point estimates aimed at minimizing both measures simultaneously, because the SE and the APE are consistent for distinct statistical functionals. In this light, it may be desirable for administrative or political leadership to provide a directive or target functional to Census Bureau scientists, much in the way that Murphy and Daan (1985) and Engelberg, Manski, and Williams (2009) requested guidance for point forecasters, in the quotes that open and motivate this article.

Alternatively, and often preferably, our imperfect knowledge of parameters or variables can be expressed in the form of probability distributions. In a general context that comprises both prediction and estimation, Pratt, Raiffa, and Schlaifer (1995, pp. 265 and 747) make this argument forcefully, thereby referring to optimal point forecasts or Bayes predictors as certainty equivalents. While their reasoning is well founded in Bayesian tradition, as noted, it applies in non-Bayesian settings as well. In the assessment of point forecasts, the duality of loss or scoring functions and functionals along with the corresponding notions of consistency and elicitability provide guidance in the choice of performance measures, alleviate much extant controversy, and merit broad attention in academia, businesses and organizations.

APPENDIX A: PROOFS

We omit the proofs of Theorems 1, 3, 4, and 6, which are well-known results.

As regards Theorem 7, the statements in parts (b) and (c) are immediate from the arguments in section 6.3 of Savage (1971) and form special cases of the more general result in Theorem 8. To prove the necessity of the representation (18), Savage essentially applied Osband's principle with the identification function $\mathrm{V}(x, y) = x - y$. Concerning Theorem 9, see Gneiting (2008b) for concise yet full-fledged proofs of parts (b) and (c), using Osband's principle with the identification function $\mathrm{V}(x, y) = \mathbb{1}(x \geq y) - \alpha$. To prove part (a), we may apply part (c) with any strictly increasing, bounded function $g : \mathrm{I} \to \mathrm{I}$, with $g(x) = \exp(-x)/(1 + \exp(-x))$ being one such example.

Proof of Theorem 2

Given $F \in \mathcal{F}$, let $t \in \mathrm{T}(F)$ and $x \in \mathrm{D}$. Then

$$\begin{aligned}\mathbb{E}_F \mathrm{S}(t, Y) &= \mathbb{E}_F\left[\int \mathrm{S}_\omega(t, y)\lambda(\mathrm{d}\omega)\right] \\ &= \int [\mathbb{E}_F \mathrm{S}_\omega(t, y)]\lambda(\mathrm{d}\omega) \\ &\leq \int [\mathbb{E}_F \mathrm{S}_\omega(x, y)]\lambda(\mathrm{d}\omega) = \mathbb{E}_F \mathrm{S}(x, Y),\end{aligned}$$

where the interchange of the expectation and the integration is allowable, because each S_ω is a nonnegative scoring function.

Proof of Theorem 5

We first prove part (b). Let $F \in \mathcal{F}^{(w)}$, $\mathbf{t} \in \mathrm{T}^{(w)}(F)$, and $\mathbf{x} \in \mathrm{D}$. Then

$$\begin{aligned}\mathbb{E}_F \mathrm{S}^{(w)}(\mathbf{t}, \mathbf{Y}) &= \mathbb{E}_F[w(\mathbf{Y})\mathrm{S}(\mathbf{t}, \mathbf{Y})] \\ &= \int \mathrm{S}(\mathbf{t}, \mathbf{y}) w(\mathbf{y}) f(\mathbf{y}) \mu(\mathrm{d}\mathbf{y}) \\ &= \left[\int \mathrm{S}(\mathbf{t}, \mathbf{y})\, \mathrm{d}F^{(w)}(\mathbf{y})\right] \cdot \left[\int w(\mathbf{y}) f(\mathbf{y}) \mu(\mathrm{d}\mathbf{y})\right]^{-1} \\ &\leq \left[\int \mathrm{S}(\mathbf{x}, \mathbf{y})\, \mathrm{d}F^{(w)}(\mathbf{y})\right] \cdot \left[\int w(\mathbf{y}) f(\mathbf{y}) \mu(\mathrm{d}\mathbf{y})\right]^{-1} \\ &= \mathbb{E}_F[w(\mathbf{Y})\mathrm{S}(\mathbf{x}, \mathbf{Y})] \\ &= \mathbb{E}_F\big[\mathrm{S}^{(w)}(\mathbf{x}, \mathbf{Y})\big],\end{aligned}$$

where μ is a dominating measure. The critical inequality holds because $F^{(w)} \in \mathcal{F}^{(w)} \subseteq \mathcal{F}$ and $\mathbf{t} \in \mathrm{T}^{(w)}(F) = \mathrm{T}(F^{(w)})$. To prove parts (c) and (a), we note that the inequality is strict if S is strictly consistent for S, unless $\mathbf{x} \in \mathrm{T}(F^{(w)}) = \mathrm{T}^{(w)}(F)$.

Proof of Theorem 8

We first prove part (b). To show the sufficiency of the representation (22), let $x \in \mathrm{I}$ and let F be a probability measure on I for which $\mathbb{E}_F[r(Y)]$, $\mathbb{E}_F[s(Y)]$, $\mathbb{E}_F[r(Y)\phi'(Y)]$, $\mathbb{E}_F[s(Y)\phi(Y)]$, and $\mathbb{E}_F[Ys(Y)\phi'(Y)]$ exist and are finite. Then

$$\begin{aligned}&\mathbb{E}_F \mathrm{S}(x, Y) - \mathbb{E}_F \mathrm{S}\left(\frac{\mathbb{E}_F[r(Y)]}{\mathbb{E}_F[s(Y)]}, Y\right) \\ &\quad = \mathbb{E}_F[s(Y)]\left\{\phi\left(\frac{\mathbb{E}_F[r(Y)]}{\mathbb{E}_F[s(Y)]}\right) - \phi(x) - \phi'(x)\left(\frac{\mathbb{E}_F[r(Y)]}{\mathbb{E}_F[s(Y)]} - x\right)\right\}\end{aligned}$$

is nonnegative, and is strictly positive if ϕ is strictly convex and $x \neq \mathbb{E}_F[r(Y)]/\mathbb{E}_F[s(Y)]$.

As regards part (c), it remains to show the necessity of the representation (22). We apply Osband's principle with the identification function $\mathrm{V}(x, y) = xs(y) - r(y)$, as proposed by Osband (1985, p. 14). Arguing in the same way as in Section 2.4, we see that

$$\partial_x \mathrm{S}(x, a)/(xs(a) - r(a)) = \partial_x \mathrm{S}(x, b)/(xs(b) - r(b))$$

for all pairwise distinct a, b and $x \in \mathrm{I}$. Hence,

$$\partial_x \mathrm{S}(x, y) = h(x)(xs(y) - r(y))$$

for $x, y \in \mathrm{I}$ and some function $h : \mathrm{I} \to \mathrm{I}$. Partial integration yields the representation (22), where

$$\phi(x) = \int_{x_0}^{x} \int_{x_0}^{s} h(u)\, \mathrm{d}u\, \mathrm{d}s \tag{A.1}$$

for some $x_0 \in \mathrm{I}$. Finally, ϕ is convex, because the scoring function S is nonnegative, which implies the validity of the subgradient inequality.

To prove part (a), we consider the scoring function (22) with $\phi(y) = y^2/(1 + |y|)$, for which the expectations in part (b) exist and are finite if, and only if, $\mathbb{E}_F[r(Y)]$, $\mathbb{E}_F[s(Y)]$, and $\mathbb{E}_F[Ys(Y)]$ exist and are finite.

Proof of Theorem 10

To show the sufficiency of the representation (28), let $x \in \mathrm{I}$ where $x < \mu_\tau$, and let F be a probability measure with compact support in I. A tedious but straightforward calculation shows that if S is of the form (28) then

$$\begin{aligned}&\mathbb{E}_F \mathrm{S}(x, Y) - \mathbb{E}_F \mathrm{S}(\mu_\tau, Y) \\ &\quad = (1 - \tau) \int_{(-\infty, x)} (\phi(\mu_\tau) - \phi(x) - \phi'(x)(\mu_\tau - x))\, \mathrm{d}F(y) \\ &\qquad + \tau \int_{[x, \mu_\tau)} (\phi(y) - \phi(x) - \phi'(x)(y - x))\, \mathrm{d}F(y) \\ &\qquad + \tau \int_{[\mu_\tau, \infty)} (\phi(\mu_\tau) - \phi(x) - \phi'(x)(\mu_\tau - x))\, \mathrm{d}F(y) \\ &\qquad + (1 - \tau) \int_{[x, \mu_\tau)} \underbrace{(\phi(\mu_\tau) - \phi(y) - \phi'(x)(\mu_\tau - y))}_{\geq \phi(\mu_\tau) - \phi(y) - \phi'(y)(\mu_\tau - y) \geq 0}\, \mathrm{d}F(y)\end{aligned}$$

is nonnegative, and is strictly positive if ϕ is strictly convex. An analogous argument applies when $x > \mu_\tau$. This proves sufficiency in part (b) as well as the claim in part (c).

To prove the necessity of the representation (28) in part (b), we apply Osband's principle with the identification function $\mathrm{V}(x, y) = |\mathbb{1}(x \geq y) - \tau|(x - y)$. Arguing in the usual way, we see that

$$\partial_x \mathrm{S}(x, y) = h(x) V(x, y)$$

for $x, y \in \mathrm{I}$ and some function $h : \mathrm{I} \to \mathrm{I}$. Partial integration yields the representation (28), where ϕ is defined as in (A.1) and is convex, because S is nonnegative.

To prove part (a), we apply part (c) with the convex function $\phi(y) = y^2/(1+|y|)$, for which $\mathbb{E}_F\phi(Y)$ exists and is finite if, and only if, $\mathbb{E}_F Y$ exists and is finite.

Proof of Theorem 11

Suppose first that $\mathcal{F}$ contains the measures with finite support. Let $a, b, c, d \in \mathrm{I}$ be such that $a < b < c < \frac{1}{2}(b+d)$, which implies $b < d$, and consider the probability measures

$$F_1 = \alpha\delta_a + \frac{1}{2}(1-\alpha)(\delta_b + \delta_d), \qquad F_2 = \alpha\delta_c + (1-\alpha)\delta_{(b+d)/2},$$

where δ_x denotes the point measure in $x \in \mathbb{R}$. If $\alpha \geq \frac{1}{3}$ then $\mathrm{CVaR}_\alpha(F_1) = \mathrm{CVaR}_\alpha(F_2) = \frac{1}{2}(b+d)$, while $\mathrm{CVaR}_\alpha(\frac{1}{2}(F_1+F_2)) = \frac{1}{4}(b+c+2d) > \frac{1}{2}(b+d)$. Thus, the level sets of the functional are not convex. By Theorem 6, the CVaR functional is not elicitable relative to the class $\mathcal{F}$. Analogous examples emerges for $\alpha < \frac{1}{3}$, or when the point measures are replaced by appropriately focused and centered absolutely continuous distributions with compact support.

APPENDIX B: OPTIMAL POINT FORECASTS UNDER THE RELATIVE ERROR SCORING FUNCTION (TABLE 8)

Here we address a problem posited by Patton (2011), in that we find the optimal point forecast or Bayes rule

$$\hat{x} = \arg\min_x \mathbb{E}_F \mathrm{S}(x, y) \quad \text{under } \mathrm{S}(x, y) = |(x-y)/x|, \tag{B.1}$$

where $Y = Z^2$ and Z has a t-distribution with mean 0, variance 1 and $\nu > 2$ degrees of freedom. In the limiting case as $\nu \to \infty$, we take Z to be standard normal.

To find the optimal point forecast, we apply Theorem 1 and part (b) of Theorem 5 with the original scoring function $\mathrm{S}(x, y) = |x^{-1} - y^{-1}|$, the weight function $w(y) = y$ and the domain $\mathrm{D} = (0, \infty)$, so that $\mathrm{S}^{(w)}(x, y) = |(x-y)/x|$. By Theorem 9, the scoring function S is consistent for the median functional. Therefore, by Theorem 5 the optimal point forecast under the weighted scoring function $\mathrm{S}^{(w)}$ is the median of the probability distribution whose density is proportional to $yf(y)$, where f is the density of Y, or equivalently, proportional to $y^{1/2}g(y^{1/2})$, where g is the density of Z.

Hence, if Z has a t-distribution with mean 0, variance 1 and $\nu > 2$ degrees of freedom, the optimal point forecast under the relative error scoring function is the median of the probability distribution whose density is proportional to

$$y^{1/2}\left(1 + \frac{y}{\nu - 2}\right)^{-(\nu+1)/2}$$

on the positive halfaxis. Using any computer algebra system, this median can readily be computed symbolically or numerically, to any desired degree of accuracy. For example, if $\nu = 4$ the optimal point forecast (B.1) is

$$\hat{x} = \frac{2}{2^{2/3} - 1} = 3.4048\ldots.$$

Table 8 provides numerical values along with the approximations in table 1 of Patton (2011), which were obtained by Monte Carlo methods, and thus are less accurate. If Z has variance σ^2, the entries in the table continue to apply, if they are multiplied by this constant.

[Received March 2010. Revised April 2011.]

REFERENCES

Abdous, B., and Theodorescu, R. (1992), "Note on the Spatial Quantile of a Random Vector," *Statistics & Probability Letters*, 13, 333–336. [757]

Acerbi, C. (2002), "Spectral Measures of Risk: A Coherent Representation of Subjective Risk Aversion," *Journal of Banking and Finance*, 26, 1505–1518. [756]

Ahlburg, D. A., Chatfield, C., Taylor, S. J., Thompson, P. H., Murphy, A. H., Winkler, R. L., Collopy, F., Armstrong, J. S., and Fildes, R. (1992), "A Commentary on Error Measures," *International Journal of Forecasting*, 8, 99–111. [758]

Armstrong, J. S. (2001), "Evaluating Forecasting Methods," in *Principles of Forecasting*, ed. J. S. Armstrong, Norwell MA: Kluwer, pp. 443–471. [747,757,758]

Armstrong, J. S., and Collopy, F. (1992), "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons," *International Journal of Forecasting*, 8, 69–80. [758]

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999), "Coherent Measures of Risk," *Mathematical Finance*, 9, 203–228. [756]

Athanasopoulos, G., Hyndman, R. J., Song, H., and Wu, D. C. (2011), "The Tourism Forecasting Competition," *International Journal of Forecasting*, 27, 822–844. [758]

Banerjee, A., Guo, X., and Wang, H. (2005), "On the Optimality of Conditional Expectation as a Bregman Predictor," *IEEE Transactions on Information Theory*, 51, 2664–2669. [753,756,757]

Bao, Y., Lee, T.-H., and Saltoğlu, B. (2006), "Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check," *Journal of Forecasting*, 25, 101–128. [756]

Berkowitz, J., and O'Brien, J. (2002), "How Accurate Are Value-at-Risk Models at Commercial Banks?" *Journal of Finance*, 57, 1093–1111. [756]

Bernardo, J. M., and Smith, A. F. M. (2000), *Bayesian Theory*, Chichester: Wiley. [756]

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroscedasticity," *Journal of Econometrics*, 31, 307–327. [747]

Buja, A., Stuetzle, W., and Shen, Y. (2005), "Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications," working paper, University of Pennsylvania, available at *http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf*. [754]

Caporin, M. (2008), "Evaluating Value-at-Risk Measures in the Presence of Long Memory Conditional Volatility," *Journal of Risk*, 10, 79–110. [755]

Carbone, R., and Armstrong, J. S. (1982), "Evaluation of Extrapolative Forecasting Methods: Results of a Survey of Academicians and Practitioners," *Journal of Forecasting*, 1, 215–217. [746,748]

Cervera, J. L., and Muñoz, J. (1996), "Proper Scoring Rules for Fractiles," in *Bayesian Statistics*, Vol. 5, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 513–519. [751]

Chaudhuri, P. (1996), "On a Geometric Notion of Quantiles for Multivariate Data," *Journal of the American Statistical Association*, 91, 862–872. [757]

Chen, K., Guo, S., Lin, Y., and Ying, Z. (2010), "Least Absolute Relative Error Estimation," *Journal of the American Statistical Association*, 105, 1104–1112. [758]

Christoffersen, P. F. (2003), *Elements of Financial Risk Management*, San Diego: Academic Press. [756]

Christoffersen, P. F., and Diebold, F. X. (1996), "Further Results on Forecasting and Model Selection Under Asymmetric Loss," *Journal of Applied Econometrics*, 11, 561–571. [747]

Dawid, A. P. (1984), "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society, Ser. A*, 147, 278–292. [746,757]

——— (2007), "The Geometry of Proper Scoring Rules," *Annals of the Institute of Statistical Mathematics*, 59, 77–93. [751]

Dawid, A. P., and Lauritzen, S. L. (2006), "The Geometry of Decision Theory," in *Proceedings of the Second International Symposium on Information Geometry and Its Applications, University of Tokyo, December 12–16, 2005*, pp. 22–28. [758]

Dawid, A. P., and Sebastiani, P. (1999), "Coherent Dispersion Criteria for Optimal Experimental Design," *The Annals of Statistics*, 27, 65–81. [757]

DeGroot, M. H., and Fienberg, S. E. (1983), "The Comparison and Evaluation of Probability Forecasters," *The Statistician: Journal of the Institute of Statisticians*, 12, 12–22. [754]

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [757]

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998), "Evaluating Density Forecasts With Applications to Financial Risk Management," *International Economic Review*, 39, 863–883. [757]

Duffie, D., and Pan, J. (1997), "An Overview of Value at Risk," *Journal of Derivatives*, 4, 7–49. [754]

Efron, B. (1991), "Regression Percentiles Using Asymmetric Squared Error Loss," *Statistica Sinica*, 1, 93–125. [750]

Engelberg, J., Manski, C. F., and Williams, J. (2009), "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters," *Journal of Business & Economic Statistics*, 27, 30–41. [747,748,759]

Engle, R. F. (1982), "Autoregressive Conditional Heteroscedasticity With Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 45, 987–1007. [747]

Ferguson, T. S. (1967), *Mathematical Statistics: A Decision-Theoretic Approach*, New York: Academic. [750,754,756]

Fildes, R. (1992), "The Evaluation of Extrapolative Forecasting Methods," *International Journal of Forecasting*, 8, 81–98. [758]

Fildes, R., and Goodwin, P. (2007), "Against Your Better Judgement? How Organizations Can Improve Their Use of Management Judgement in Forecasting," *Interfaces*, 37, 570–576. [746,748,757]

Fildes, R., Nikolopoulos, K., Crone, S. F., and Syntetos, A. A. (2008), "Forecasting and Operational Research: A Review," *Journal of the Operational Research Society*, 59, 1150–1172. [747]

French, S., and Ríos Insua, D. (2000), *Statistical Decision Theory*, London: Arnold. [752,756]

Geweke, J., and Whiteman, C. (2006), "Bayesian Forecasting," in *Handbook of Economic Forecasting*, Vol. 1, eds. G. Elliott, C. W. J. Granger, and A. Timmermann, Amsterdam: Elsevier, pp. 3–80. [757]

Giacomini, R., and Komunjer, I. (2005), "Evaluation and Combination of Conditional Quantile Forecasts," *Journal of Business & Economic Statistics*, 23, 416–431. [756]

Gneiting, T. (2008a), "Editorial: Probabilistic Forecasting," *Journal of the Royal Statistical Society, Ser. A*, 171, 319–321. [746,757]

——— (2008b), "Quantiles as Optimal Point Forecasts," Technical Report 538, University of Washington, Dept. Statistics, available at *http://www.stat.washington.edu/research/reports/2008/tr538.pdf*. [755]

——— (2011), "Quantiles as Optimal Point Forecasts," *International Journal of Forecasting*, 27, 197–207. [754]

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [751,754]

Gneiting, T., and Thorarinsdottir, T. L. (2010), "Predicting Inflation: Professional Experts versus No-Change Forecasts," unpublished manuscript, University of Heidelberg, available at *http://arxiv.org/abs/1010.2318*. [758]

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008), "Assessing Probabilistic Forecasts of Multivariate Quantities, With Applications to Ensemble Predictions of Surface Winds," *Test*, 17, 211–264. [756]

Granger, C. W. J. (1969), "Prediction With a Generalized Cost of Error Function," *Operational Research Quarterly*, 20, 199–207. [756]

Granger, C. W. J., and Machina, M. J. (2006), "Forecasting and Decision Theory," in *Handbook of Economic Forecasting*, Vol. 1, eds. G. Elliott, C. W. J. Granger, and A. Timmermann, Amsterdam: Elsevier, pp. 82–98. [750,753]

Granger, C. W. J., and Pesaran, M. H. (2000a), "Economic and Statistical Measures of Forecast Accuracy," *Journal of Forecasting*, 19, 537–560. [750,757]

——— (2000b), "A Decision Theoretic Approach to Forecast Evaluation," in *Statistics and Finance: An Interface*, eds. W.-S. Chan, W. K. Li, and H. Tong, London: Imperial College Press, pp. 261–278. [750,757]

Hallin, M., Paindaveine, D., and Šiman, M. (2010), "Regression Quantiles: From L_1 Optimization to Halfspace Depth" (with discussion), *The Annals of Statistics*, 38, 635–703. [757]

Henderson, R., Jones, M., and Stare, J. (2001), "Accuracy of Point Predictions in Survival Analysis," *Statistics in Medicine*, 20, 3083–3096. [756]

Hering, A. S., and Genton, M. G. (2010), "Powering Up With Space-Time Wind Forecasting," *Journal of the American Statistical Association*, 105, 92–104. [756]

Hilden, J. (2008), "Scoring Rules for Evaluation of Prognosticians and Prognostic Rules," workshop notes, University of Copenhagen, available at *http://biostat.ku.dk/~jh/*. [751]

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67. [758]

Horowitz, J. L., and Manski, C. F. (2006): "Identification and Estimation of Statistical Functionals Using Incomplete Data," *Journal of Econometrics*, 132, 445–459. [750]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101. [758]

Huber, P. J., and Ronchetti, P. M. (2009), *Robust Statistics* (2nd ed.), Hoboken, NJ: Wiley. [750,758]

Hyndman, R. J., and Koehler, A. B. (2006), "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting*, 22, 679–688. [758]

James, W., and Stein, C. (1961), "Estimation With Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, ed. J. Neyman, Berkeley: University of California Press, pp. 361–379. [757]

Jolliffe, I. T. (2008), "The Impenetrable Hedge: A Note on Propriety, Equitability and Consistency," *Meteorological Applications*, 15, 25–29. [751]

Jolliffe, I. T., and Stephenson, D. B. (eds.) (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Chichester: Wiley. [757,758]

Jose, V. R. R., and Winkler, R. L. (2009), "Evaluating Quantile Assessments," *Operations Research*, 57, 1287–1297. [751]

Kemperman, J. H. B. (1987), "The Median of a Finite Measure on a Banach Space," in *Statistical Data Analysis Based on the L_1 Norm and Related Methods*, ed. Y. Dodge, Amsterdam: North Holland, pp. 217–230. [757]

Keynes, J. M. (1911), "The Principal Averages and the Laws of Error Which Lead to Them," *Journal of the Royal Statistical Society*, 74, 322–331. [758]

Klein, I., and Grottke, M. (2008), "On J. M. Keynes' 'The Principal Averages and the Laws of Error Which Lead to Them'—Refinement and Generalisation," discussion paper, University of Erlangen, available at *http://www.iwqw.wiso.uni-erlangen.de/forschung/07-2008.pdf*. [758]

Koenker, R. (2005), *Quantile Regression*, Cambridge: Cambridge University Press. [755]

——— (2006), "The Median Is the Message: Toward the Fréchet Mean," *Journal de la Société Française de Statistique*, 147, 61–64. [757]

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [754,758]

Koltchinskii, V. I. (1997), "M-Estimation, Convexity and Quantiles," *The Annals of Statistics*, 25, 435–477. [757]

Komunjer, I. (2005), "Quasi Maximum-Likelihood Estimation for Conditional Quantiles," *Journal of Econometrics*, 128, 137–164. [758]

Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008), "Eliciting Properties of Probability Distributions," in *Proceedings of the 9th ACM Conference on Electronic Commerce, July 8–12, 2008, Chicago, Illinois*, extended abstract. [749-752,757]

Laurent, S., Rombouts, J. V. K., and Violante, F. (2009), "On Loss Functions and Ranking Forecasting Performances of Multivariate Volatility Models," discussion paper, Catholic University of Louvain, available at *http://www.cirpee.org/fileadmin/documents/Cahiers_2009/CIRPEE09-48.pdf*. [756,757]

Lehmann, E., and Casella, G. (1998), *Theory of Point Estimation* (2nd ed.), New York: Springer. [756]

Lehmann, E. L. (1951), "A General Concept of Unbiasedness," *Annals of Mathematical Statistics*, 22, 587–592. [751]

Lopez, J. A. (1999), "Regulatory Evaluation of Value-at-Risk Models," *Journal of Risk*, 1, 37–64. [755]

Makridakis, S., and Hibon, M. (1979), "Accuracy of Forecasting: An Empirical Investigation" (with discussion), *Journal of the Royal Statistical Society, Ser. A*, 142, 97–145. [758]

——— (2000), "The M3-Competition: Results, Conclusions and Implications," *International Journal of Forecasting*, 16, 451–476. [758]

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1982), "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," *Journal of Forecasting*, 1, 111–153. [758]

Makridakis, S., Chatfield, C., Hibon, M., Lawrance, M., Mills, T., Ord, K., and Simmons, L. F. (1993), "The M2-Competition: A Real-Time Judgementally Based Forecasting Study," *International Journal of Forecasting*, 9, 5–22. [758]

McCarthy, T. M., Davis, D. F., Golicic, S. L., and Mentzner, J. T. (2006), "The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practice," *Journal of Forecasting*, 25, 303–324. [746,748]

Mentzner, J. T., and Kahn, K. B. (1995), "Forecasting Technique Familiarity, Satisfaction, Usage, and Application," *Journal of Forecasting*, 14, 465–476. [746,748]

Moskaitis, J. R., and Hansen, J. A. (2006), "Deterministic Forecasting and Verification: A Busted System?" working paper, Massachusetts Institute of Technology, available at *http://wind.mit.edu/~hansen/papers/MoskaitisHansenWAF2006.pdf*. [747]

Murphy, A. H. (1973), "Hedging and Skill Scores for Probability Forecasts," *Journal of Applied Meteorology*, 12, 215–223. [758]

Murphy, A. H., and Daan, H. (1985), "Forecast Evaluation," in *Probability, Statistics and Decision Making in the Atmospheric Sciences*, eds. A. H. Murphy, and R. W. Katz, Boulder, CO: Westview Press, pp. 379–437. [748,751,759]

Murphy, A. H., and Winkler, R. L. (1987), "A General Framework for Forecast Verification," *Monthly Weather Review*, 115, 1330–1338. [747]

Noorbaloochi, S., and Meeden, G. (1983), "Unbiasedness as the Dual of Being Bayes," *Journal of the American Statistical Association*, 78, 619–623. [751]

Newey, W. K., and Powell, J. L. (1987), "Asymmetric Least Squares Estimation and Testing," *Econometrica*, 55, 819–847. [755]

Osband, K. (2011), *Pandora's Risk: Uncertainty at the Core of Finance*, New York: Columbia University Press. [756]

Osband, K., and Reichelstein, S. (1985), "Information-Eliciting Compensation Schemes," *Journal of Public Economics*, 27, 107–115. [756,757]
Osband, K. H. (1985), "Providing Incentives for Better Cost Forecasting," Ph.D. thesis, University of California, Berkeley. [749-753,759]
Park, H., and Stefanski, L. A. (1998), "Relative-Error Prediction," *Statistics & Probability Letters*, 40, 227–236. [752]
Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [758]
Patton, A. J. (2011), "Volatility Forecast Comparison Using Imperfect Volatility Proxies," *Journal of Econometrics*, 160, 246–256. [746,750,752-755,760]
Patton, A. J., and Sheppard, K. (2009), "Evaluating Volatility and Correlation Forecasts," in *Handbook of Financial Time Series*, eds. T. G. Anderson, R. A. Davis, J.-P. Kreiss, and T. Mikosch, Berlin: Springer, pp. 801–838. [757]
Pers, T. H., Albrechtsen, A., Holst, C., Sørensen, T. I. A., and Gerds, T. A. (2009), "The Validation and Assessment of Machine Learning: A Game of Prediction From High-Dimensional Data," *PLoS ONE*, 4, e6287, *doi:10.1371/journal.pone.0006287*. [758]
Pesaran, M. H., and Skouras, S. (2002), "Decision-Based Methods for Forecast Evaluation," in *A Companion to Economic Forecasting*, eds. M. H. Pesaran and S. Skouras, Malden: Blackwell Publishers, pp. 241–267. [751]
Phelps, R. R. (1966), *Lectures on Choquet's Theorem*, Princeton: D. Van Nostrand. [751]
Pratt, J. W., Raiffa, H., and Schlaifer, R. (1995), *Introduction to Statistical Decision Theory*, Boston: Massachusetts Institute of Technology. [759]
Pukelsheim, F. (2006), *Optimal Design of Experiments* (SIAM Classics ed.), Philadelphia: SIAM. [757]
Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Clinton: Colonial Press. [754]
Reichelstein, S., and Osband, K. (1984), "Incentives in Government Contracts," *Journal of Public Economics*, 24, 257–270. [753]
Rockafellar, R. T. (1970), *Convex Analysis*, Princeton: Princeton University Press. [753]
Rockafellar, R. T., and Uryasev, S. (2000), "Optimization of Conditional Value-at-Risk," *Journal of Risk*, 2, 21–42. [756]
——— (2002), "Conditional Value-at-Risk for General Loss Distributions," *Journal of Banking and Finance*, 26, 1443–1471. [756]
Saerens, M. (2000), "Building Cost Functions Minimizing to Some Summary Statistics," *IEEE Transactions on Neural Networks*, 11, 1263–1271. [749,753,754]
Savage, L. J. (1971), "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66, 783–810. [749,753,754,756,757,759]
Schervish, M. J. (1989), "A General Method for Comparing Probability Assessors," *The Annals of Statistics*, 17, 1856–1879. [754]
Serfling, R. (2002), "Quantile Functions for Multivariate Analysis: Approaches and Applications," *Statistica Neerlandica*, 56, 214–232. [757]
Shipp, S., and Cohen, S. (2009), "COPAFS Focuses on Statistical Activities," *Amstat News*, August 2009, 15–18. [758]
Small, C. G. (1990), "A Survey of Multidimensional Medians," *International Statistical Review*, 58, 263–277. [757]
Thomson, W. (1979), "Eliciting Production Possibilities From a Well-Informed Manager," *Journal of Economic Theory*, 20, 360–380. [749,753,754,757]
Wedderburn, R. W. M. (1974), "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss–Newton Method," *Biometrika*, 61, 439–447. [758]
Wellner, J. A. (2009), "Statistical Functionals and the Delta Method," lecture notes, University of Washington, available at *http://www.stat.washington.edu/people/jaw/COURSES/580s/581/LECTNOTES/ch7.pdf*. [750]
West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084. [757]
West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models* (2nd ed.), New York: Springer. [757]
Winkler, R. L. (1996), "Scoring Rules and the Evaluation of Probabilities" (with discussion), *Test*, 5, 1–60. [751,754]
Yelland, P. M., Kim, S., and Stratulate, R. (2010), "A Bayesian Model for Sales Forecasting at Sun Microsystems," *Interfaces*, 40, 118–129. [758]
Yu, C. W. (2009), "Median Loss Analysis and Its Application to Model Selection," Ph.D. thesis, Dept. of Statistics, University of British Columbia, Vancouver. [758]
Zellner, A. (1986), "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," *Journal of the American Statistical Association*, 81, 446–451. [758]
Zhang, C., Jiang, Y., and Chai, Y. (2010), "Penalized Bregman Divergence for Large-Dimensional Regression and Classification," *Biometrika*, 97, 551–566. [758]