Making and Evaluating Point Forecasts

Tilmann Gneiting

Universität Heidelberg

Eltville, June 2, 2012

Probabilistic forecasts versus point forecasts

there is a growing, **trans-disciplinary concensus** that

forecasts ought to be **probabilistic** in nature, taking the form of **probability distributions** over future quantities and events

however, many applications require a **point forecast**, x, for a future quantity with **realizing value**, y

in a nutshell, I contend that in **making** and **evaluating point forecasts**, it is critical that the point forecasts

- **derive** from **probabilistic forecasts**, and
- are **evaluated** in **decision theoretically** principled ways

indeed, as argued by Pesaran and Skouras (2002), the **decision-theoretic** approach provides a **unifying framework** for the **evaluation** of both probabilistic and point **forecasts**

How point forecasts are commonly assessed

many applications require a **point forecast**, x, for a future real-valued or positive quantity with **realizing value**, y

various **forecasters** or forecasting methods $m = 1, \ldots, M$ compete

they issue point forecasts x_{mn} with realizing values y_n, at a **finite set** of times, locations or instances $n = 1, \ldots, N$

the forecasters are assessed and ranked by the **mean score**

$$\bar{\mathsf{S}}_m = \frac{1}{N} \sum_{n=1}^{N} \mathsf{S}(x_{mn}, y_n)$$

for $m = 1, \ldots, M$, where

$$\mathsf{S} : \mathbb{R} \times \mathbb{R} \to [0, \infty) \quad \text{or} \quad \mathsf{S} : (0, \infty) \times (0, \infty) \to [0, \infty)$$

is a **scoring function**, generally satisfying regularity conditions

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

Some frequently used scoring functions

often, not just one but a whole set of **scoring functions** is used to compare and rank competing forecasting methods

the following are among the **most commonly used** for a **positive quantity**

$\mathrm{S}(x,y) = (x-y)^2$	squared error (SE)
$\mathrm{S}(x,y) = \lvert x-y \rvert$	absolute error (AE)
$\mathrm{S}(x,y) = \lvert (x-y)/y \rvert$	absolute percentage error (APE)
$\mathrm{S}(x,y) = \lvert (x-y)/x \rvert$	relative error (RE)

according to surveys, **organizations** and **businesses** commonly use the **SE**, **AE** and, in particular, the **APE**

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

Use of scoring functions in the journal literature in 2008

	Total	FP	SE	AE	APE	MSC
Group I: **Forecasting**						
Int J Forecasting	41	32	21	10	8	4
J Forecasting	39	25	23	13	5	3
Group II: **Statistics**						
Ann Appl Stat	62	8	6	3	1	0
Ann Stat	100	5	3	2	0	0
J Am Stat Assoc	129	10	9	1	0	0
J Roy Stat Soc Ser B	49	5	4	1	0	0
Group III: **Econometrics**						
J Bus Econ Stat	26	9	8	2	1	0
J Econometrics	118	5	5	0	0	0
Group IV: **Meteorology**						
Bull Am Meteor Soc	73	1	1	0	0	0
Mon Wea Rev	300	63	58	8	2	0
Q J Roy Meteor Soc	148	19	19	0	0	0
Wea Forecasting	79	26	20	11	0	1

What scoring function(s) ought to be used in practice?

arguably, there is considerable **contention** about the **choice** of a **scoring function** or **error measure**

Murphy and Winkler (1987):

> "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."

Fildes (2008):

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

Bowsher and Meeks (2008):

> "It is now widely recognized that when comparing forecasting models [. . .] no close relationship is guaranteed between model evaluations based on conventional error-based measures such as [squared error] and those based on the ex post realized profit (or utility) from using each model's forecasts to solve a given economic decision or trading problem. Leitch and Tanner (1993) made just this point in the context of interest rate forecasting."

Simulation study: Forecasting a highly volatile asset price

we seek to predict a highly **volatile asset price**, y_t

in this **simulation study**, y_t is a realization of the random variable

$$Y_t = Z_t^2,$$

where Z_t follows a **GARCH time series model**, namely,

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

we consider **three competing forecasters** issuing **one-step ahead** point predictions of the asset price

- the **statistician** is aware of the data-generating mechanism and issues the true conditional mean,

  $$\widehat{x}_t = \mathbb{E}(Y_t) = \mathbb{E}(Z_t^2) = \sigma_t^2$$

  as point forecast

Simulation study: Forecasting a highly volatile asset price

we seek to predict a highly **volatile asset price**, y_t

in this **simulation study**, y_t is a realization of the random variable

$$Y_t = Z_t^2,$$

where Z_t follows a **GARCH time series model**, namely,

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

we consider **three competing forecasters** issuing **one-step ahead** point predictions of the asset price

- the **statistician** is aware of the data-generating mechanism and issues the true conditional mean,

  $$\widehat{x}_t = \mathbb{E}(Y_t) = \mathbb{E}(Z_t^2) = \sigma_t^2$$

  as point forecast

- the **optimist** always issues $\widehat{x}_t = 5$

Simulation study: Forecasting a highly volatile asset price

we seek to predict a highly **volatile asset price**, y_t

in this **simulation study**, y_t is a realization of the random variable

$$Y_t = Z_t^2,$$

where Z_t follows a **GARCH time series model**, namely,

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

we consider **three competing forecasters** issuing **one-step ahead** point predictions of the asset price

- the **statistician** is aware of the data-generating mechanism and issues the true conditional mean,

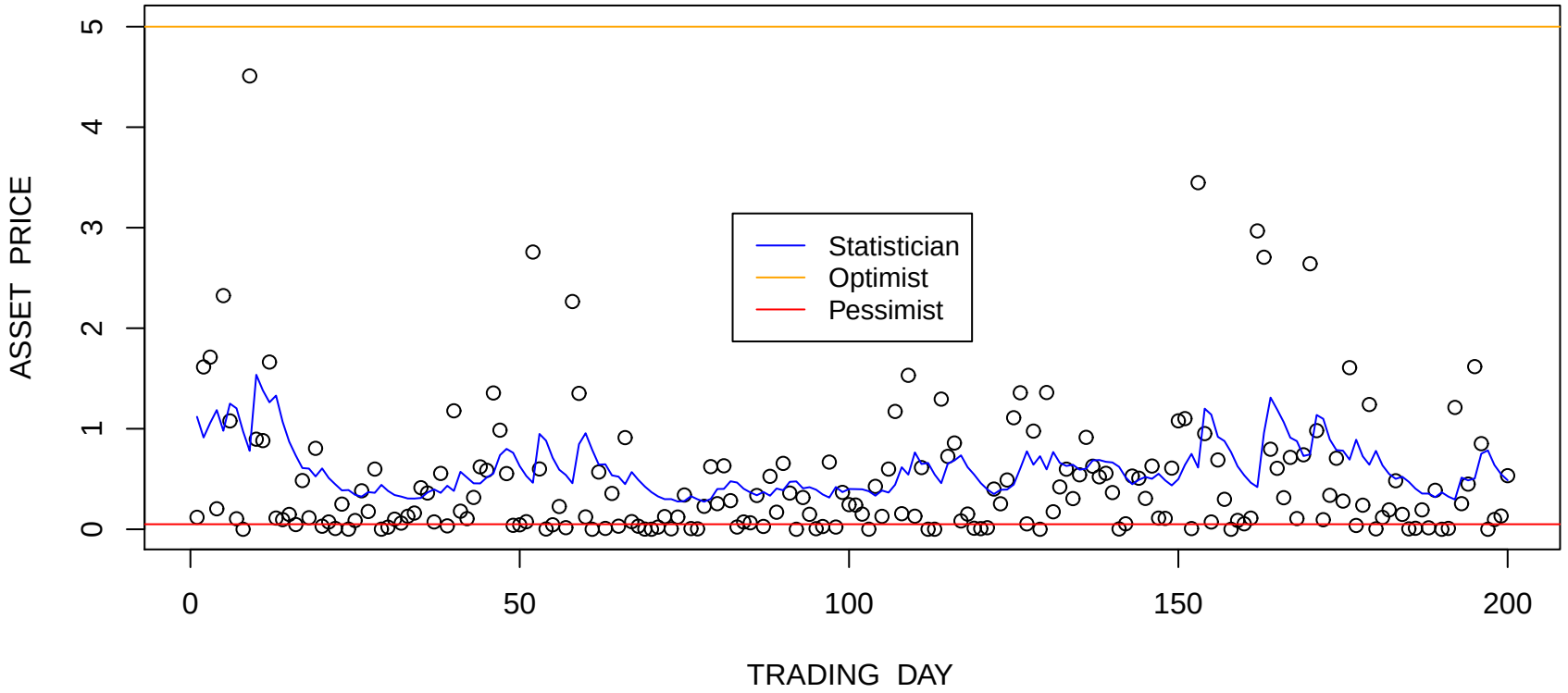
  $$\widehat{x}_t = \mathbb{E}(Y_t) = \mathbb{E}(Z_t^2) = \sigma_t^2$$

  as point forecast

- the **optimist** always issues $\widehat{x}_t = 5$

- the **pessimist** always issues $\widehat{x}_t = 0.05$

Simulation study: Forecasting a highly volatile asset price



Simulation study: Forecasting a highly volatile asset price

we evaluate and rank the three competing forecasters, namely the **statistician**, the **optimist** and the **pessimist**, by their **mean scores**, which are averaged over 100,000 one-step ahead point forecasts

Forecaster	SE	AE	APE	RE
Statistician	**5.07**	0.97	2.58	0.97
Optimist	22.73	4.35	13.96	**0.87**
Pessimist	7.61	**0.96**	**0.14**	19.24

(APE to be multiplied by 10^5)

What does the literature say?

Engelbert, Manski and Williams (2008):

> "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.**"

Murphy and Daan (1985):

> "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."

Resolving the puzzle:
Point forecasters need ‘guidance’ or ‘directives’

requesting ‘some’ point forecast, and then evaluating forecasters by using ‘some’ (set of) scoring functions, as is common practice in the literature, is not a meaningful endeavor

rather, **point forecasters** need **‘guidance’** or **’directives’**

First option

inform forecasters **ex ante** about the **scoring function(s)** to be employed, and allow them to **tailor** the **point forecast** to the **scoring function**

Second option

request a specific **functional** of the forecaster’s predictive distribution, such as the **mean** or a **quantile**

First option: Specify scoring function ex ante

inform forecasters **ex ante** about the **scoring function(s)** to be employed to assess their work, and allow them to **tailor** the **point forecast** to the scoring function

this permits the statistically literate forecaster to mutate into **Mr. Bayes**, that is, to issue the **Bayes predictor**,

$$\widehat{x} = \arg\min_x \mathbb{E}_F \left[\mathsf{S}(x, Y)\right]$$

as her **point forecast**, where the expectation is taken with respect to the forecaster's (subjective or objective) **predictive distribution**, F

for example, if S is the **squared error** scoring function **(SE)**, the Bayes predictor is the **mean** of the predictive distribution

if S is the **absolute error** scoring function **(AE)**, the Bayes predictor is any **median** of the predictive distribution

Simulation study: Forecasting a highly volatile asset price

we consider the aforementioned point forecasters, namely **Mr. Bayes**, the **statistician**, the **optimist**, and the **pessimist**

Mr. Bayes employs the **Bayes rule** or **optimal point forecast**

Scoring Function	Bayes Rule	Simulation Study
SE	$\widehat{x} = \text{mean}(F)$	σ_t^2
AE	$\widehat{x} = \text{median}(F)$	$0.455\,\sigma_t^2$
APE	$\widehat{x} = \text{med}^{(-1)}(F)$	ε
RE	$\widehat{x} = \text{med}^{(1)}(F)$	$2.366\,\sigma_t^2$

Mr. Bayes dominates his competitors

Forecaster	SE	AE	APE	RE
Mr. Bayes	**5.07**	**0.86**	**< 0.01**	**0.75**
Statistician	**5.07**	0.97	2.58	0.97
Optimist	22.73	4.35	13.96	0.87
Pessimist	7.61	0.96	0.14	19.24

Second option: Specify functional ex ante

Consistency and elicitability

request a specific **functional**, $\mathsf{T}(F)$, of the forecaster's predictive distribution, F, such as the **mean** or a **quantile**

and apply any **scoring function** that is **consistent for** the **functional** T, in the sense that

$$\mathbb{E}_F\left[\mathsf{S}(\mathsf{T}(F), Y)\right] \leq \mathbb{E}_F\left[\mathsf{S}(x, Y)\right]$$

for all x, with the natural interpretation when T is set-valued

a **consistent scoring function** is a special case of a **proper scoring rule** for probabilistic forecasts

a functional is **elicitable** if there exists a scoring function that is **strictly consistent** for it, in the sense that equality holds if, and only if, $x = \mathsf{T}(F)$

not all functionals are elicitable, for example, the **variance** and the **conditional value-at-risk (CVaR)** functionals are not

Osband's principle

given an **elicitable functional** T, can we **characterize** the class of the **scoring functions** S that are **consistent** for it?

Osband (1985) argued that if there exists an **identification function** V such that

$$\mathbb{E}_F[\mathsf{V}(x, Y)] = 0 \iff x = \mathsf{T}(F)$$

and the consistent scoring function S is smooth in its first argument, then

$$\mathsf{S}_{(1)}(x, y) = h(x)\,\mathsf{V}(x, y)$$

with some (typically) strictly positive function h

frequently, we can **integrate** with respect to x to obtain the **general form** of a scoring rule S that is **consistent** for T

see the examples below, in which the functional T is a **mean**, a **ratio of expectations**, a **quantile** or an **expectile**

Mean functional

the **mean functional** is **elicitable**, and the scoring functions that are **consistent** for it are of the **Bregman** form

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

where ϕ is **convex** with **subgradient** ϕ'

an important special case is that of a **probability forecast** $x = p = \mathbb{E}[Y]$ for a **binary event** Y that realizes as $y = 0$ or $y = 1$

implying that **proper scoring rules** for **probability forecasts** are also of the **Bregman** form

for example, if $\phi(x) = x^2$ we obtain the **squared error (SE)** or **Brier** scoring function, namely $\mathsf{S}(x, y) = (x - y)^2$

rich history: **Brier (1950)**, McCarthy (1956), Shuford, Albert and Massengil (1966), **Savage (1971)**, Reichelstein and Osband (1984), Banerjee, Guo and Wang (2005)

Ratios of expectations functional

the **ratio of expectations** functional

$$\mathsf{T}(F) = \frac{\mathbb{E}_F[r(Y)]}{\mathbb{E}_F[s(Y)]},$$

where r and s are sufficiently regular and s is strictly positive, is **elicitable**

the scoring functions that are **consistent** for the **ratio of expectations** functional are of the form

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

where ϕ is **convex** with **subgradient** ϕ'

if $r(y) = y$ and $s(y) \equiv 1$, we recover the above classical case of the **mean** functional

Quantiles

the α-**quantile** functional ($0 < \alpha < 1$) is **elicitable**, and the scoring functions that are **consistent** for it are of the **generalized piecewise linear (GPL)** form

$$\mathsf{S}(x,y) = \begin{cases} \alpha\,(g(y) - g(x)) & \text{if} \quad x \le y \\ (1-\alpha)\,(g(x) - g(y)) & \text{if} \quad x \ge y \end{cases}$$

where g is **nondecreasing**

for example, if $g(x) = x$ we obtain the **asymmetric piecewise linear** scoring function,

$$\mathsf{S}(x,y) = \begin{cases} \alpha\,|x-y| & \text{if} \quad x \le y \\ (1-\alpha)\,|x-y| & \text{if} \quad x \ge y \end{cases}$$

which includes the **absolute error** scoring function **(AE)** in the special case $\alpha = \frac{1}{2}$

history: **Thomson (1979)**, Saerens (2000), Cervera and Muñoz (1996), Gneiting and Raftery (2007), Jose and Winkler (2009)

Expectiles

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

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

the **τ-expectile** functional is **elicitable**, and the scoring functions that are **consistent** for it are of the form

$$\mathsf{S}(x, y) = \begin{cases} \tau\,(\phi(y) - \phi(x) - \phi'(x)(y - x)) & \text{if } \; x \le y \\ (1 - \tau)\,(\phi(y) - \phi(x) - \phi'(x)(y - x)) & \text{if } \; x \ge y \end{cases}$$

where ϕ is **convex** with **subgradient** ϕ'

for example, if $\phi(x) = x^2$ we obtain the **asymmetric piecewise quadratic** scoring function

these scoring functions combine key characteristics of the **Bregman** and **GPL** families

Consistent scoring functions as proper scoring rules

at this point, we return to **probabilistic forecasts**

if $\mathcal{F}$ denote a class of probabilistic forecasts on $\mathbb{R}$, a **proper scoring rule** is any function

$$\mathsf{R} : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$$

such that

$$\mathbb{E}_F \, \mathsf{R}(F, Y) \leq \mathbb{E}_F \, \mathsf{R}(G, Y) \quad \text{for all} \quad F, G \in \mathcal{F},$$

with the **logarithmic score** and the **continuous ranked probability score** being key examples

any consistent scoring function induces a proper scoring rule, as follows: if the **scoring function**

$$\mathsf{S} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$$

is **consistent** for the **functional** T, the relationship

$$\mathsf{R} : \mathcal{F} \times \mathbb{R} \longrightarrow [0, \infty), \quad (F, y) \longmapsto \mathsf{R}(F, y) = \mathsf{S}(\mathsf{T}(F), y)$$

defines a **proper scoring rule**

Discussion: Theoretical perspectives

motivating challenge: **methods** for **forecast evaluation** need to be **decision theoretically coherent**, so that we avoid pathologies and paradoxes

in particular, **scoring rules** for **probabilistic forecasts** ought to be **proper**, and **scoring functions** for **point forecasts** ought to be **consistent** for the target functional at hand

here, we have characterized the **loss** (or **scoring**) **functions** that lead to the **mean**, **ratios of expectations**, **quantiles**, and **expectiles** as the **Bayes rule** or **optimal point forecast** and thus are **consistent** for these functionals

pioneering works on the critical notions of **consistency** and **elicitability** include those of Savage (1971), Thomson (1979) and Osband (1985)

Discussion: A puzzle in economic forecast evaluation

Bowsher and Meeks (2008):

> "It is now widely recognized that when comparing forecasting models [. . .] no close relationship is guaranteed between model evaluations based on conventional error-based measures such as [squared error] and those based on the ex post realized profit (or utility) from using each model's forecasts to solve a given economic decision or trading problem. Leitch and Tanner (1993) made just this point in the context of interest rate forecasting."

despite being widely recognized, these are **counterintuitive** and **disconcerting** observations

I contend that they stem from misguided forecast evaluation techniques and disappear when point forecasts are **derived from probabilistic forecasts** and assessed in **decision theoretically** principled ways

Discussion: A puzzle in economic forecast evaluation

our student Sam Dörken has made an attempt to **replicate** and **extend** the study of **Leitch and Tanner (1993)**, using decision theoretically principled methods

	DA	AE	SE	A	B	C	D
DA	—	0.41	0.45	0.75	0.57	0.65	0.69
AE	0.41	—	0.97	0.37	0.41	0.48	0.59
SE	0.45	0.97	—	0.45	0.48	0.56	0.65
Profit Rule A	0.75	0.37	0.45	—	0.82	0.86	0.85
Profit Rule B	0.57	0.41	0.48	0.82	—	0.89	0.80
Profit Rule C	0.65	0.48	0.56	0.86	0.89	–	0.94
Profit Rule D	0.69	0.59	0.65	0.85	0.80	0.94	—

absolute correlation between evaluations using various mean scores for 1982 to 1996, treating each forecast horizon separately

in this experiment, the **correlations** are **substantial**, contrary to what Leitch and Tanner (1993) observed

Discussion: Economic perspectives

principled **decision making** requires full **predictive distributions** rather than just **point forecasts**

in **making** and **evaluating point forecasts**, it is critical that the point forecasts **derive** from **probabilistic forecasts**, and are evaluated in decision theoretically principled ways, using **scoring functions** that are **consistent** for an **elicitable** target functional

not all functionals are elicitable

in particular, **conditional value-at-risk (CVaR)** is **not elicitable** and thus may not be the ideal risk measure in the revision of the **Basel protocol** for banking regulations

Selected references

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**, 746–762.

Osband, K. H. (1985). Providing incentives for better cost forecasting. Ph.D. Thesis, University of California, Berkeley.

Pesaran, M. H. and Skouras, S. (2002). Decision-based methods for forecast evaluation. In *A Companion to Economic Forecasting*, Pesaran, M. H. and Skouras, S., eds., Blackwell Publishers, pp. 241–267.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66**, 783–810.

Thomson, W. (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, **20**, 360–380.

Gneiting, T. and Thorarinsdottir, T. L. (2010). Predicting inflation: Professional experts versus no-change forecasts. Preprint, `arxiv.org/abs/1010.2318`.