The Elicitation Problem

Tobias Fissler[*1]

[1] *University of Bern*

Competing point forecasts for functionals such as the mean, a quantile, or a certain risk measure are commonly compared in terms of loss functions. These should be incentive compatible, i.e., the expected score should be minimized by the correctly specified functional of interest. A functional is called *elicitable* if it possesses such an incentive compatible loss function. With the squared loss and the absolute loss, the mean and the median possess such incentive compatible loss functions, which means they are elicitable. In contrast, variance or Expected Shortfall are not elicitable. Besides investigating the elicitability of a functional, it is important to determine the whole class of incentive compatible loss functions as well as to give recommendations which loss function to use in practice, taking into regard secondary quality criteria of loss functions such as order-sensitivity, convexity, or homogeneity.

Keywords: Consistency, Elicitability, Expected Shortfall, Scoring functions, Value at Risk

1 Evaluating and comparing forecasts

"From the cradle to the grave, human life is full of decisions. Due to the inherent nature of time, decisions have to be made today, but at the same time, they are supposed to account for unknown and uncertain future events. However, since these future events cannot be *known* today, the best thing to do is to base the decisions on *predictions* for these unknown and uncertain events. The call for and the usage of predictions for future events is literally ubiquitous and even dates back to ancient times." [2] Today, elaborated forecasts are present in a variety of different disciplines: government, business, finance, the energy market, agriculture, or everyday life.

Assume we have $m \in \mathbb{N}$ competing experts issuing their forecasts for time $t = 1, \ldots, N$. Then, one has *prediction-observation-sequences*

$$\big(x_t^{(i)}, y_t\big)_{t=1,\ldots,N} \qquad i \in \{1, \ldots, m\}. \tag{1}$$

*Corresponding author: tobias.fissler@stat.unibe.ch

The values y_t are *ex post* realizations of a time series $(Y_t)_{t\in\mathbb{N}}$, taking values in an *observation domain* O, whereas $x_t^{(i)}$ are *ex ante* forecasts taking values in an *action domain* A. Assessing the quality of the forecasts, one can ask two main questions: (i) How good is the forecast at hand in *absolute* terms? And (ii) How good is the forecast at hand in *relative* terms? Question (i) deals with *forecast validation*, whereas question (ii) is concerned with *forecast selection*, *forecast comparison*, or *forecast ranking.* The concept of elicitability – and the elicitation problem in particular – focuses on question (ii).

1.1 Consistent scoring functions and elicitability

To introduce the abstract decision-theoretic framework of forecast comparison, there is no need to specify the observation domain O and the action domain A. In particular, the observations can be real-valued, vector-valued, but also functional-valued or even set-valued. Acknowledging the uncertainty of future outcomes, the forecasts can be probabilistic in nature, taking the form of probability distributions or densities. In this case, the action domain A coincides with a class of probability distributions $\mathcal{F}$ where one assumes that $\mathcal{F}$ contains the (conditional) distributions F_t of Y_t. On the other hand, one is often interested in certain statistical properties of the underlying distribution $F_t \in \mathcal{F}$ of Y_t such as the mean, the median, or a certain risk measure. Mathematically speaking, such a property can be specified in terms of a *functional* $T\colon \mathcal{F} \to \mathsf{A}$. In this situation, one speaks about *point forecasts*, and typically, A coincides with O (e.g. in case of the mean) where $\mathsf{A} = \mathbb{R}^k$, but might also be functional-valued or set-valued. Interestingly, the concept of probabilistic forecasts can be covered by the latter upon considering the identity map on $\mathcal{F}$ as the functional T. For most of the forthcoming results, we focus on vector-valued point forecasts, meaning $\mathsf{A} = \mathbb{R}^k$, and $\mathsf{O} = \mathbb{R}^d$.

Commonly, competing forecasts are assessed in terms of loss or *scoring functions* $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$, with the most popular choices $S(x, y) = |x - y|$, or $S(x, y) = (x - y)^2$ when $\mathsf{A} = \mathsf{O} = \mathbb{R}$. Thus, if a forecaster reports the quantity $x \in \mathsf{A}$ and $y \in \mathsf{O}$ materializes, she is *penalized* by $S(x, y) \in \mathbb{R}$. Given the competing prediction-observation-sequences at (1), the ranking is done in terms of the *realized scores* $\bar{\mathbf{S}}_N^{(i)} = \frac{1}{N}\sum_{t=1}^N S(x_t^{(i)}, y_t)$, $i \in \{1, \dots, m\}$. That is, a forecaster is deemed to be the better the lower her realized score is. However, this ranking depends on the choice of the scoring function S. To incentivize truthful and hones forecasts, the *Bayes act* $\arg\min_{x\in\mathsf{A}} \mathbf{E}_F[S(x, Y)]$ should coincide with the correctly specified forecast $T(F)$, hence, the scoring function must be chosen *in line* with the functional T. If $T(F) = \arg\min_{x\in\mathsf{A}} \mathbf{E}_F[S(x, Y)]$ for all $F \in \mathcal{F}$, S is called *strictly $\mathcal{F}$-consistent* for $T\colon \mathcal{F} \to \mathsf{A}$. Following the terminology of [5, 8], a functional $T\colon \mathcal{F} \to \mathsf{A}$ is called *elicitable* if it possesses a strictly $\mathcal{F}$-consistent scoring function S. Besides meaningful forecast comparison and ranking, the elicitability of a functional opens the possibility to do M-estimation. That is, under certain regularity conditions on the sequence $(Y_t)_{t\in\mathbb{N}}$ detailed e.g. in [6], $\widehat{T}_n = \arg\min_{x\in\mathsf{A}} \frac{1}{n}\sum_{t=1}^{n} S(x, Y_t)$ is a consistent estimator for T, if S is strictly consistent for T. Similarly, elicitability leads the way to generalized regression such as quantile regression or expectile regression; see [7, 9].

2 The elicitation problem

Having settled the basic definitions, one can formulate a threefold *elicitation problem* with respect to a fixed functional $T\colon \mathcal{F} \to \mathsf{A}$.

(i) Is T elicitable?

(ii) What is the class of strictly $\mathcal{F}$-consistent scoring functions for T?

(iii) What are *good choices* of strictly $\mathcal{F}$-consistent scoring functions?

The rest of this abstract summarizes some important ideas, contributions, and results concerning the elicitation problem.

2.1 Which functionals are elicitable?

One natural way to show the elicitability of a functional is by directly providing a strictly consistent scoring function. In particular, one can show that under certain regularity assumptions, the piecewise linear loss $S_\alpha(x, y) = (\mathbf{1}\{y \le x\} - \alpha)(x - y)$ is strictly consistent for the α-quantile, and that the piecewise squared loss $S_\tau(x, y) = |\mathbf{1}\{y \le x\} - \tau|(x - y)^2$ is strictly consistent for the τ-expectile (in particular, the mean and the median, as well as all moments, are elicitable, subject to mild regularity assumptions). [11] has provided a powerful necessary condition in terms of the level sets of the functional at hand, which is often relatively easy to check in practice.

Proposition 1 (Convex level sets [11])**.** *Let $T\colon \mathcal{F} \to \mathsf{A}$ be elicitable. Then, for any $F_0, F_1 \in \mathcal{F}$ such that $T(F_0) = T(F_1) = t$ and for any $\lambda \in (0,1)$ such that $F_\lambda = (1-\lambda)F_0 + \lambda F_1 \in \mathcal{F}$ it holds that $T(F_\lambda) = t$.*

Remarkably, the proof works independently of the specific choice of A. The result shows that variance and Expected Shortfall (ES) are generally not elicitable [5]. If $\mathsf{A} = \mathbb{R}$ and if the functional T fulfills some continuity conditions, [12] showed the sufficiency of convex level sets for elicitability. Similar results for sufficiency lack for the case $\mathsf{A} = \mathbb{R}^k$ when $k > 1$.

In case of vector-valued functionals, a functional $T = (T_1, \dots, T_k)$ consisting of elicitable components is again elicitable. If S_m is strictly consistent for T_m, then $S(x_1, \dots, x_k, y) = \sum_{m=1}^k S_m(x_m, y)$ is a strictly consistent scoring function for T. This observation provokes the questions (a) whether strictly consistent scoring functions must be necessarily of this form, and (b) whether functionals consisting only of elicitable components are the only vector-valued functionals. The *revelation principle* [11] gives a negative answer to question (b). It asserts that any bijection of an elicitable functional is elicitable. Since the pair (mean, variance) is a bijection of the first two moments, which are elicitable, this shows the elicitability of the pair (mean, variance), even though variance itself is not elicitable. This somehow unexpected result leads to the natural question: *Are bijections of functionals with elicitable components the only elicitable functionals?* It turns out that this is not the case: The two risk measures Expected Shortfall (ES) and Value at Risk (VaR) are, as a pair, jointly elicitable even though ES itself is not elicitable; see Theorem 1. Moreover, there is generally no (known) bijection between (VaR, ES) and a vector consisting only of elicitable components.

2.2 Determine the class of strictly consistent scoring functions

Interestingly, strictly consistent scoring functions for a functional T are not unique. E.g., if S is strictly consistent for T, then $(x, y) \mapsto \lambda S(x, y) + a(y)$ is also strictly consistent for T for any $\lambda > 0$ and any 'offset-function' $a\colon \mathsf{O} \to \mathbb{R}$. Moreover, the class of strictly consistent scoring functions is convex. However, there is far more flexibility in the class. A powerful tool is the so-called *Osband's principle* [11, 3]. It connects the gradient of an expected score with the expectation of an *identification function*. An identification function for a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is a function $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ such that $\mathbf{E}_F[V(x, Y)] = 0$ if and only if $x = T(F)$ for all $F \in \mathcal{F}$. Examples are $V(x, y) = x - y$ for the mean and $V(x, y) = \mathbf{1}\{y \le x\} - \alpha$ for the α-quantile. If a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is elicitable and possesses an identification function, then, under some richness conditions on the class $\mathcal{F}$, there exists a matrix-valued function $h\colon \mathsf{A} \to \mathbb{R}^{k\times k}$ such that

$$\nabla_x \mathbf{E}_F[S(x, Y)] = h(x)\, \mathbf{E}_F[V(x, Y)] \qquad \forall x \in \mathsf{A},\ \forall F \in \mathcal{F}. \tag{2}$$

One can also derive a *second order* Osband's principle considering the Hessian $\nabla_x^2 \mathbf{E}_F[S(x, Y)]$ of the expected score. Under appropriate smoothness conditions, the Hessian must be symmetric for all $F \in \mathcal{F}$ and positive semi-definite at $x = T(F)$. This implies further necessary conditions on the matrix-function

h often even leading to sufficient conditions for strict consistency.[1] Exploiting Osband's principle, one can show that – under some regularity conditions – $S\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is a strictly consistent scoring function for the mean if and only if S is of *Bregman type*, that is,

$$S(x,y) = \phi'(x)(x-y) - \phi(x) + a(y), \tag{3}$$

where $\phi\colon \mathbb{R}\to\mathbb{R}$ is strictly convex. Similarly, S is strictly consistent for the α-quantile if and only if

$$S(x,y) = \big(\mathbf{1}\{y\le x_1\} - \alpha\big)g(x_1) - \mathbf{1}\{y\le x_1\}g(y) + a(y), \tag{4}$$

where $g\colon \mathbb{R}\to\mathbb{R}$ is strictly increasing. Indeed, taking derivatives of the expected score, (3) becomes $\partial_x\,\mathbf{E}_F\left[S(x,Y)\right] = \phi''(x)(x-\mathbf{E}_F\left[Y\right])$ such that ϕ'' plays the role of h in (2). For (4), one obtains $\partial_x\,\mathbf{E}_F\left[S(x,Y)\right] = g'(x)(F(x)-\alpha)$, such that $g' = h$ in (2).

Expected Shortfall is jointly elicitable with Value at Risk

VaR and ES are the most popular risk measures in practice. For a financial position Y with distribution F and a level $\alpha\in(0,1)$, they are defined as

$$\begin{aligned}
\mathrm{VaR}_\alpha(F) &:= F^{-1}(\alpha) = \inf\{x\in\mathbb{R} : F(x)\ge\alpha\},\\
\mathrm{ES}_\alpha(F) &:= \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\beta(F)\,\mathrm{d}\beta = \mathbf{E}_F\left[Y\,|\,Y\le \mathrm{VaR}_\alpha(F)\right].
\end{aligned}$$

That means risky positions yield large *negative* values of VaR_α or ES_α. Intuitively, VaR_α gives the worst loss out of the best $(1-\alpha)\times 100\%$ of all cases, whereas ES_α gives the average loss given one exceeds VaR_α. There is an ongoing debate in academia and industry which risk measure to use. The debate mainly concentrates on ES_α and VaR_α. The latter, as a quantile, is elicitable under mild regularity conditions, it fails to be superadditive, thus violating the coherence property of risk measures. Moreover, it fails to take into account the size of losses beyond the level α. Conversely, ES_α considers the whole tail of the distribution beyond the level α, it fulfills the coherence property, but fails to be elicitable. In this light, the following result is crucial and opens the possibility to meaningful forecast comparison of joint (VaR, ES)-forecasts which is of particular importance in the context of quantitative risk management and especially the question of backtestability [4, 10].

[1] Using second order Osband's principle, one can show for example, that any vector of different quantiles and / or expectiles only possesses strictly consistent scoring functions that are additively separable. On the other hand, vectors of expectations allow for a more flexible structure similar to (3). This gives answers to the previous question (a).

Theorem 1 ([3]). *Let $\alpha \in (0,1)$. Let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ with finite first moments and unique α-quantiles.*
(i) If $\phi\colon \mathbb{R} \to \mathbb{R}$ is strictly convex and if for any $x_2 \in \mathbb{R}$, the function

$$[x_2, \infty) \to \mathbb{R}, \quad x_1 \mapsto g(x_1) + \phi'(x_2)\frac{x_1}{\alpha} \tag{5}$$

is strictly increasing, then the scoring function $S\colon \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$, where $\mathsf{A}_0 := \{(x_1, x_2) \in \mathbb{R}^2 \colon x_1 \geq x_2\}$, of the form

$$\begin{aligned} S(x_1, x_2, y) = {} & \big(\mathbf{1}\{y \leq x_1\} - \alpha\big) g(x_1) - \mathbf{1}\{y \leq x_1\} g(y) + a(y) \\ & + \phi'(x_2)\Big(x_2 + \big(\mathbf{1}\{y \leq x_1\} - \alpha\big)\frac{x_1}{\alpha} - \mathbf{1}\{y \leq x_1\}\frac{y}{\alpha}\Big) - \phi(x_2), \end{aligned} \tag{6}$$

is strictly $\mathcal{F}$-consistent for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$.
(ii) Conversely, under some regularity conditions, all strictly consistent scoring functions for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ are of the form given at (6).

Part (ii) of Theorem 1 asserting the necessity of the form at (6) can be shown using Osband's principle with the joint two-dimensional identification function $V(x_1, x_2, y) = \big(\mathbf{1}\{y \leq x_1\} - \alpha, x_2 + \big(\mathbf{1}\{y \leq x_1\} - \alpha\big)\frac{x_1}{\alpha} - \mathbf{1}\{y \leq x_1\}\frac{y}{\alpha}\big)'$. Part (i) can be proved by anticipating that for fixed x_1 the function $(x_2, y) \mapsto S(x_1, x_2, y)$ is of Bregman-type with minimum at $V_2(x_1, x_2, y) = 0$. On the other hand, for fixed x_2, due to the condition at (5), the function $(x_1, y) \mapsto S(x_1, x_2, y)$ is a strictly consistent scoring function for the α-quantile.

2.3 Secondary quality criteria besides strict consistency

Facing the multitude of strictly consistent scoring functions illustrated at (3), (4), and (6), this burden of choice calls for new concepts such as the notion of forecast dominance introduced in [1]. Alternatively, it motivates the introduction of secondary quality criteria besides strict consistency giving guidance which scoring function to use. This line of research is pursued in [2]. Generalizations of the concept of order-sensitivity [8] to the higher dimensional setting are introduced, ensuring meaningful forecast comparison of possibly misspecified predictions in particular settings. Convexity of scoring functions can show to be beneficial for optimization purposes, but also shed new light on the paradigm of maximizing the sharpness of a forecast subject to calibration as well as on incentives for cooperation between competing forecasters. Finally, equivariance properties of functionals motivate the notion of order-preserving scoring functions, nesting concepts such as homogeneity or translation invariance of scoring functions.

References

[1] W. Ehm, T. Gneiting, A. Jordan, and F. Krüger. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *JRSSB*, 78(3):505–562, 2016.

[2] T. Fissler. *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space*. PhD thesis, University of Bern, 2017.

[3] T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. *Ann. Statist.*, 44(4):1680–1707, 2016.

[4] T. Fissler, J. F. Ziegel, and T. Gneiting. Expected shortfall is jointly elicitable with value-at-risk: implications for backtesting. *Risk Magazine*, pages 58–61, January 2016.

[5] T. Gneiting. Making and Evaluating Point Forecasts. *J. Amer. Statist. Assoc.*, 106:746–762, 2011.

[6] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, second edition, 2009.

[7] R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

[8] N. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, Chicago, Il, USA, 2008. ACM.

[9] W. K. Newey and J. L. Powell. Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55:819–847, 1987.

[10] N. Nolde and J. F. Ziegel. Elicitability and backtesting. *Preprint on arXiv*, 2017.

[11] K. H. Osband. *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley, 1985.

[12] I. Steinwart, C. Pasin, R. Williamson, and S. Zhang. Elicitation and Identification of Properties. *JMLR: Workshop and Conference Proceedings*, 35:1–45, 2014.