Discussion of the paper: "Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings"

Hajo Holzmann[1] **and Bernhard Klar**[2]

[1]: *Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany.*
[2]: *Institut für Stochastik, Karlsruher Institut für Technologie (KIT), Germany.*

The authors have provided us with an interesting paper on the comparison of distinct forecasts of quantiles and expectiles. Recently, Gneiting (2011) made a strong point that such forecast comparisons should be based on consistent scoring functions. However, there are a variety of consistent scoring functions for both quantiles and expectiles, and empirical forecast rankings may well depend on the particular choice.
Therefore, the authors introduce the notion of forecast dominance for all consistent scoring functions, and show that in case of quantiles or expectiles, this can be checked by considering appropriate one-parameter families of elementary scoring functions.
We feel that this notion of forecast dominance is very strong, and can often not be expected to hold even in situations where one forecast should clearly be favoured over another. Consider the example given in Table 2 in the paper, and let us focus on forecasting the median. Introduce another forecaster, call her extreme, who issues $10\,\mu$. Murphy diagrams for the expected score as well as for empirical scores for samples of sizes $n = 30$ are given in Figure 1. The curves of the perfect and the extreme forecaster all touch each other at $\vartheta = 0$, and the empirical curves in Figure 1 (a)-(c) actually all intersect, even in the "reasonable" subinterval $[-1, 1]$ of values for the parameter ϑ. We simulated the probability that the Murphy diagrams intersect (not only touch) in the interval $[-1, 1]$ for distinct sample sizes, the results are in Table 1. While the probability converges to zero for the climatological and the sign-reversed forecasters, it actually comes close to one for the extreme forecaster. In contrast, the probability that the classical quantile score of the perfect forecast exceeds that of one of the other forecasters quickly goes to zero, as shown in Table 2.
Thus, while Murphy diagrams are useful descriptive tools for comparing forecasts, the formal notion of forecast dominance which involves a large family of consistent but not strictly consistent scoring functions should be used with care.

[1]Address for correspondence: Prof. Dr. Hajo Holzmann, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerwein-Straße. D-35032 Marburg, Germany, Email: holzmann@mathematik.uni-marburg.de, Fon: + 49 6421 2825454

n	Clim.	Sign	Extreme
10	63.5	21.5	85.2
20	58.4	8.2	91.9
30	51.2	3.0	94.4
50	39.1	0	96.6
100	20.8	0	98.1
200	5.2	0	98.9

Table 1: Relative frequency of an intersection of the Murphy curve with that of the perfect forecaster in the interval $[-1, 1]$ for different sample sizes in 10000 simulations.

n	Clim.	Sign	Extreme
10	8.7	0.6	0
20	2.9	0	0
30	0.8	0	0
50	0	0	0

Table 2: Relative frequency of exceedance of the quantile score of the perfect forecaster over that of the respective forecaster for different sample sizes in 10000 simulations.

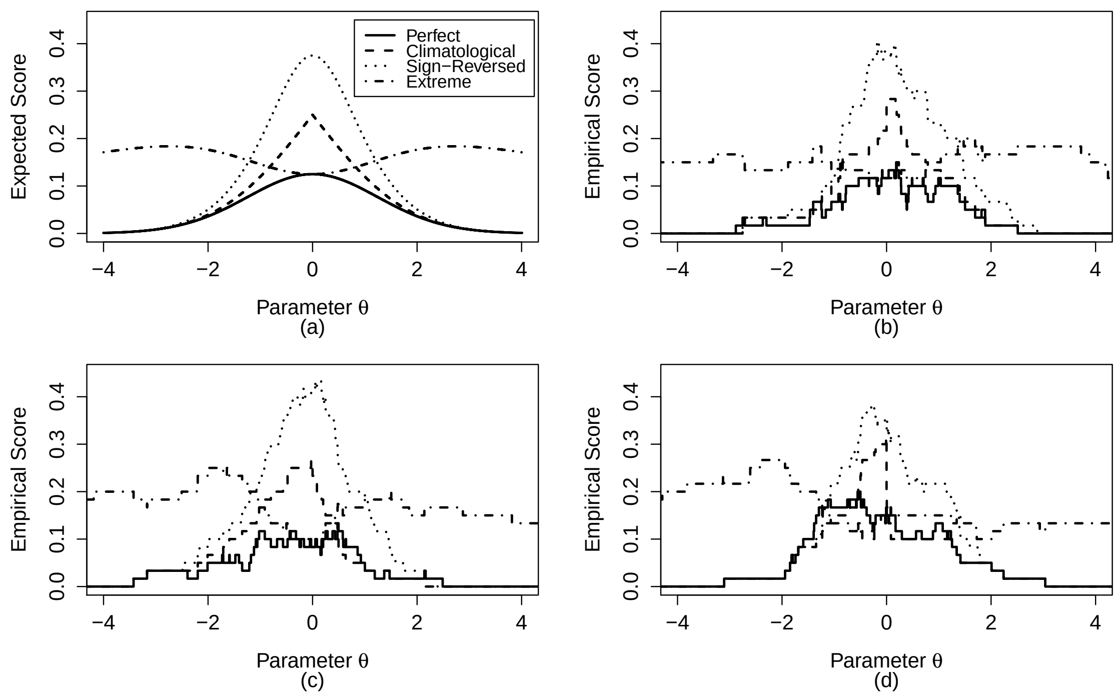

Figure 1: Murphy diagrams for perfect, climatological, sign-reversed and extreme forecasters. (a): Theoretical scores (b)-(d): scores for distinct samples of size $n = 30$.