Discussion of "A Stricly Proper Score for Assessing the quality of Prediction Intervals"

Pierre Pinson, *Senior Member, IEEE*, *et al.*

IN A SERIES of recent work published in the *IEEE Transactions on Neural Networks and Learning Systems*, the *IEEE Transactions on Power Systems*, *Electric Power Systems Research* and here in the *IEEE Transactions on Sustainable Energy* (among others), Khosravi and co-authors proposed and utilize a new score for the evaluation of interval forecasts, the so-called Coverage Width-based Criterion (CWC). This score has been used for the tuning (in-sample) and genuine evaluation (out-of-sample) of prediction intervals for various applications, e.g. electric load [1], electricity prices [2], general purpose prediction [3] and wind power generation [4], [5]. After a series of discussion papers on the improper nature of the CWC [10], [11], Khosravi *et al.* [6] proposed a new version of the CWC, which they claim is proper. Unfortunately, it is again not the case. This will be illustrated based on an example below, which is a slightly evolved version of that used in [10].

Probabilistic forecasting is to become a core aspect in modern power systems engineering, with increased penetration of renewable energy sources, and with their inherent variability and lack of predictability, e.g., wind and solar energy. Besides, load patterns are becoming more variable and less predictable due to changes in consumption patterns with the apparition of proactive prosumers. It will overall result in more uncertainty in market-clearing outcomes such as energy volumes and prices. Probabilistic forecasts in the form of quantiles, intervals, predictive densities or more generally trajectories, are optimal inputs to a wide range of decision-making problems defined in a stochastic or robust optimization framework. These forecasts are more difficult to evaluate than the more common single-valued (/deterministic) predictions, owing to their very nature. Scoring rules to be used for the evaluation of probabilistic forecasts are required to be proper [7]–[9]: propriety is the basic property of a score to insure that perfect forecasts should be given the best score value, say, the lowest one if the score is negatively oriented. If not the case, one could then hedge the score, by finding tricks that permit to get better score values without attempting to issue better forecasts. More generally, employing a score that is not proper makes that one can never be sure of the validity of the results from an empirical comparison or benchmarking of rival approaches. Research on the topic of proper evaluation of probabilistic forecasts, in the form of prediction intervals, can at least be traced back to the work of Winkler [12].

Unfortunately in the case of the aforementioned paper proposing a new score for interval forecasts, this version of the CWC score is not proper, as will be illustrated below based on a simple example.

Pierre Pinson is with the Department of Electrical Engineering at the Technical University of Denmark, while ...

Let us first remind the reader about the definition of the new version of the CWC score. For a given lead time and nominal coverage rate $(1-\beta)$, it writes

$$\text{CWC} = \bar{\delta}\mathbf{1}\{\Delta b > 0\}\exp\left(\eta\Delta b\right),\ \eta > 0, \tag{1}$$

with $\mathbf{1}\{.\}$ an indicator function, returning 1 if the condition between brackets realizes, and to 0 otherwise. In parallel, $\Delta b = (1-\beta) - b$ is the difference between nominal $(1-\beta)$ and empirical (b) coverage rates (that is, a form of probabilistic bias), while $\bar{\delta}$ is the average width of the prediction intervals. η is a free parameter that can be set to any positive value. It is argued that based on the above definition, the CWC penalizes intervals that are not probabilistically reliable, while it rewards them for their sharpness (since sharp intervals are intuitively expected to be more informative). The CWC is negatively oriented: lower values indicate prediction intervals of higher quality. This proposal formulation is very close from the original one proposed in [1] only insuring that the term related to the probabilistic bias component is not multiplied by the average interval width.

We now introduce a simple example in order to show how the CWC is not proper and may give a better score value to intervals that should actually be deemed of lower quality. This example was first introduced in [10]. Consider a stochastic process $\{X_t,\ t = 1, \ldots, T\}$ defined as a sequence of T independent and identically distributed (i.i.d.) random variables X_t with probability density function (pdf) defined on a compact support, with

$$g(x) = 12\left(x - \frac{1}{2}\right)^2, \quad x \in [0, 1]. \tag{2}$$

We denote by G the cumulative distribution function (cdf) associated to g, given by

$$G(x) = 4\left(x - \frac{1}{2}\right)^3 + \frac{1}{2}, \quad x \in [0, 1]. \tag{3}$$

One can readily verifies that G is an increasing function, with $G(0) = 0$ and $G(1) = 1$.

For this stochastic process consisting of i.i.d. random variables, it straightforward to define the optimal interval forecasts directly based on the density in (2). For instance, for a nominal coverage rate of 0.9 (to cover observations 90% of the times), optimal central prediction intervals $\mathcal{I}_t^*$ for any time t are defined by the quantiles with nominal levels 0.05 and 0.95:

$$\mathcal{I}^* = \left[G^{-1}(0.05), G^{-1}(0.95)\right]. \tag{4}$$

And, based on the expression for G given in (3),

$$\mathcal{I}^* = [0.017, 0.983]. \tag{5}$$

These intervals are perfectly reliable by definition, and therefore the CWC value assessing their quality is equal to their average width, i.e., $\text{CWC}^* = 0.966$. Since the above prediction intervals are the perfect ones, no other intervals should be given a better score.

Now in order to hedge the score, simply consider generating prediction intervals in a binary manner, although acknowledging that the nominal coverage rate should be respected in practice. Following such a binary approach, it was proposed in [10] that intervals are defined as full intervals [0,1] 90% of the times, and as empty intervals (i.e., any single value in [0,1]) 10% of the times. This writes

$$\mathcal{I} = \begin{cases} [0,1], & \text{if } u_t \geq 0.1 \\ 0.5, & \text{otherwise} \end{cases}, \tag{6}$$

using 0.5 as an example value for the empty intervals, and where u_t is a realization at time t from a sequence of i.i.d. uniform random variables $U_t \sim \mathcal{U}[0,1]$. These intervals are clearly not sophisticated ones, and not informative at all. For the former version of the CWC, since covering the actual observations of the process 90% of the times, by construction, their CWC score values is also given by their average width, that is, $\text{CWC} = 0.9$ (significantly lower than the value obtained for the perfect prediction intervals). With the new version of the CWC, the score is still $\text{CWC} = 0.9$. Note that if it was argued that empty intervals are not allowed, one could replace them with intervals of width 0.1, centered in any way, and the resulting score would still be $\text{CWC} = 0.91$ (since widening the intervals cannot worsen calibration in the CWC definition), much better than for the optimal interval forecasts.

Actually, the two problems with the CWC are *(i)* that it is not based on a scoring rule that would assign a score value for each and every forecast-obervation pair. It is instead based on statistics (for calibration and sharpness) calculated over an evaluation set of considerable length; and *(ii)* it is based on the fallacy such that intervals having empirical coverage higher than the nominal ones are adequately calibrated. This is not true, since lack of coverage, or too much coverage both comprise a lack of probabilistic calibration.

REFERENCES

[1] A. Khosravi, S. Nahavandi, and D. Creighton, "Construction of optimal prediction intervals for load forecasting problems," *IEEE Transactions on Power Systems*, vol. 25, pp. 1496-1503, 2010.

[2] A. Khosravi, S. Nahavandi, and D. Creighton, "A neural-network-GARCH based method for construction of prediction intervals," *Electric Power Systems Research*, vol. 96, pp. 185–193, 2013.

[3] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Comprehensive review of Neural Network-based prediction intervals and new advances," *IEEE Transactions on Neural Networks*, vol. 22, pp. 1341–1356, 2011.

[4] A. Khosravi, S. Nahavandi, and D. Creighton, "Prediction intervals for short-term wind farm generation forecasts," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 3, pp. 602610, Jul. 2013.

[5] A. Khosravi, and S. Nahavandi, "Combined nonparametric prediction intervals for wind power generation," *IEEE Transactions on Sustainable Energy*,vol. 4, no. 4, pp. 849856, Oct. 2013.

[6] A. Khosravi, A.F. Atiya, and S. Nahavandi, "A strictly proper score for assessing the quality of prediction intervals," *IEEE Transactions on Neural Networks and Learning Systems*, submitted.

[7] J. Bröcker, and L. A. Smith, "Scoring probabilistic forecasts: on the importance of being proper," *Weather and Forecasting*, vol. 22, pp. 382–388, 2007.

[8] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society B*, vol. 69, pp. 243–268, 2007.

[9] T. Gneiting, and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation", *Journal of the American Statistical Association*, vol. 102, pp. 359-378, 2007.

[10] P. Pinson, J. Tastu (2014). Discussion of "Prediction intervals for short-term wind farm generation forecasts" and "Combined nonparametric prediction intervals for wind power generation," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 3, pp. 1019-1020.

[11] C. Wan, Z. Xu, J. østergaard, Z.Y. Dong, K.P. Wong (2014). Discussion of "Combined nonparametric prediction intervals for wind power generation," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 3, pp. 1021.

[12] R. L. Winkler, "A decision-theoretic approach to interval estimation," *Journal of the American Statistical Association*, vol. 67, pp. 187191, 1972.