International Journal of Forecasting 36 (2020) 208–211

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Discussion

Why the "best" point forecast depends on the error or accuracy measure

Stephan Kolassa

SAP Switzerland, Bahnstrasse 2, 8274 Tägerwilen, Switzerland

1. Introduction

The M4 competition is certain to stimulate a considerable discussion of forecasting methods, as well as of the methodology to be used when running forecasting competitions, whether ones as ambitious as the M competitions or just ones conducted by a company searching for a new forecasting solution. I will focus on only one aspect of such competitions; specifically, on point forecast error or accuracy measures (henceforth "PFEMs", for "point forecast error measures").

In discussing PFEMs, Makridakis, Spiliotis, and Assimakopoulos (2020) cite a lack of "a clear agreement regarding their advantages and drawbacks", and therefore decide to use three different ones: the sMAPE, the MASE and an overall weighted average (OWA) of the two. While it may be true that no such agreement yet exists, my objective is to propose a possibility for it, and to discuss implications for future forecasting competitions.

2. Optimal point forecasts

Let us take a step back. In forecasting, we do not know the future perfectly, and even if we produce a point forecast, this does not imply any certainty that the actual outcome will be exactly equal to the point forecast. We can express our predictive uncertainty using a *predictive density* $\widehat{f}_t$. Our point forecast can now be seen as an attempt to summarize $\widehat{f}_t$ in a single number $\widehat{f}_t \mapsto \widehat{y}_t(\widehat{f}_t)$. For instance, one very common "one-number summary" is the expectation, $\widehat{y}_t(\widehat{f}_t) := E(\widehat{f}_t)$. Quantile forecasts can also be seen as one-number summaries of a predictive density (Gneiting, 2011b). Note that I contend that, while $\widehat{f}_t$ may not be discussed explicitly, considering its presence in the background can help us to understand the behavior of PFEMs.

Similarly, the actual outcome y_t is a realization of an unknown distribution f_t, which we will call the *(actual) future density*. The PFEM is then a function e of the point forecast $\widehat{y}_t$ and the observation y_t, or $e(\widehat{y}_t, y_t)$.

What is a "good" point forecast? Since we typically evaluate our forecasts using e over multiple time points (or over multiple series), a "good" forecast is one that minimizes e "on average", or *in expectation*. That is, we wish to find the $\widehat{y}_t = \widehat{y}_t(\widehat{f}_t)$ that minimizes $E_{y_t \sim f_t}\left(e(\widehat{y}_t, y_t)\right)$.

Now, the key observation is that, given $\widehat{f}_t$, the one-number summary

$$\widehat{f}_t \mapsto \widehat{y}_t(\widehat{f}_t) := \operatorname*{argmin}_x E_{y_t \sim f_t}\left(e(x, y_t)\right) \tag{1}$$

that minimizes this expectation will of course depend on the quantity being minimized in expectation, i.e., on e. In other words, different PFEMs e will be minimized in expectation by *different* ways of summarizing the predictive density:

$$(\widehat{f}_t, e) \mapsto \widehat{y}_t^e(\widehat{f}_t) := \operatorname*{argmin}_x E_{y_t \sim f_t}\left(e(x, y_t)\right). \tag{2}$$

A one-number summary of a given predictive density that minimizes the expected MSE may (and usually *will*) be quite different from a one-number summary of the *same* predictive density that minimizes the expected MAPE.

3. An example

Consider Fig. 1. We simulate an iid series of gamma distributed data with shape parameter 3 and scale parameter 3. Here, we *know* the future density, which does not depend on t, $f_t = f$, so we can calculate the point forecasts that actually minimize the expected errors for various PFEMs (Kolassa, 2019).

E-mail address: Stephan.Kolassa@sap.com.

https://doi.org/10.1016/j.ijforecast.2019.02.017

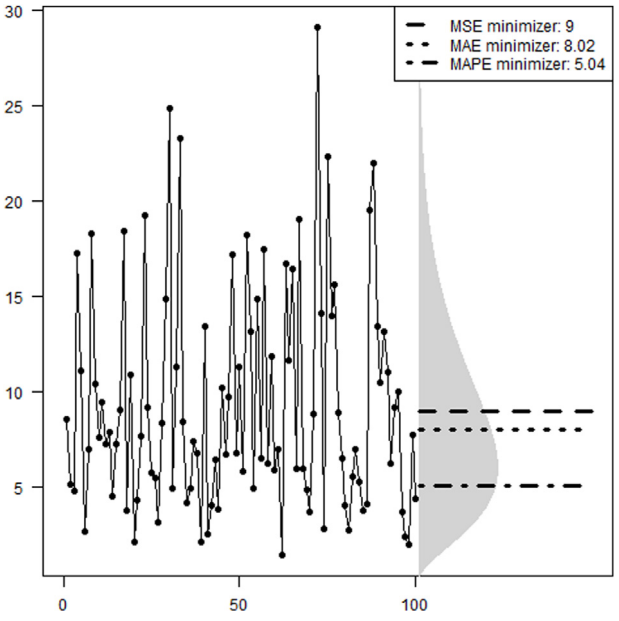

Fig. 1. Simulated time series, iid gamma distributed with shape parameter 3 and scale parameter 3, with density and point forecasts minimizing the expected (M)SE, (M)AE and (M)APE.

- We know that the expected value of f minimizes the expected squared error, so the MSE-minimizing point forecast is $\widehat{y}^{\text{MSE}} = 3 \times 3 = 9$.
- The median of f minimizes the expected absolute error (Hanley, Joseph, Platt, Chung, & Belisle, 2001). There is no closed form for the median of a gamma, so we use the approximation proposed by Choi (1994) to obtain $\widehat{y}^{\text{MAE}} \approx 8.02$. Since the MASE is simply a scaled MAE (Hyndman & Koehler, 2006), this quantity also minimizes the expected MASE.
- The median of a distribution that is proportional to $\frac{1}{y}f(y)$ minimizes the expected absolute percentage error (Gneiting, 2011a, p. 748, with $\beta = -1$). A short calculation shows that this distribution is again a gamma, with shape $3 - 1 = 2$ and scale again 3. We approximate its median by again using the work of Choi (1994) to obtain $\widehat{y}^{\text{MAPE}} \approx 5.04$.

We note that the three point forecasts are quite different, especially the one that minimizes the expected MAPE.

4. Takeaways

What should a forecaster do if he believes that the optimal predictive density is a gamma and knows that the point forecast will be being evaluated on multiple PFEMs, as is widespread practice? As his point forecast, he might choose a functional of the gamma that minimizes a particular PFEM if he believes this one to be the most relevant one; however, this comes at the cost of performing worse on other PFEMs. For instance, reporting the mean of the predictive density will minimize the expected MSE but perform quite badly for the MAPE, whereas reporting the MAPE-minimizing (-1)-median (Gneiting, 2011a) will yield a high MSE.

Clearly, it is problematic to report a *single* point forecast, e.g., the conditional mean, and evaluate it using *different* PFEMs, as there is not a single "best point forecast" which the different PFEMs evaluate "from different angles". Instead, there are different functionals of central tendency of a distribution, such as the mean, the median and the (-1)-median, and different PFEMs measure the agreement of a given point forecast with these functionals of the actual future distribution. Summarizing a predictive distribution in a single functional but then evaluating this using different PFEMs makes no more sense than summarizing someone's health by estimating their height and then evaluating this estimate using first a yardstick and then a bathroom scale.

Note that this conundrum is not a consequence of an error in estimating the predictive density, but arises even if our forecaster *knows* the correct future distribution. Indeed, it reflects the underlying uncertainty in the data generating process itself. Nor is this an artifact of the asymmetry of the gamma: if the predictive density is symmetric, then the MSE-minimizing and MAE-minimizing point forecasts are indeed identical, but the MAPE-minimizing one will still be lower (Kolassa & Martin, 2011).

Thus, we have a simple and statistically rigorous explanation for the well-known fact that the rankings of different forecasting methods using different PFEMs will often exhibit little correlation, as in Table 4 of Makridakis et al. (2020). I believe that the forecasting community should draw multiple conclusions from this effect, both for future competitions such as the planned M5 and for general forecasting research, teaching and practice:

- If forecast competitions request point forecasts and assess these using different PFEMs, they should not evaluate a *single* point forecast using *different* PFEMs. Instead, it should be possible to submit multiple different point forecasts for each time series and future time point, one for each PFEM.
- Similarly, I hope to see future forecasting papers address this issue explicitly and use different point forecasts for different PFEMs. This is already understood well in terms of providing different quantile forecasts for different confidence levels (cf. Gneiting, 2011b), and the same should hold if our PFEMs aim to assess the central tendencies of predictive densities. Newly-proposed forecasting methods regularly improve on benchmarks by only a very little (see many papers in the *IJF*), and I contend that these differences are meaningless as long as (implicit) expectation point forecasts are being evaluated using, e.g., the MAPE.
- Point forecasts for central tendencies will continue to make sense and be studied. What does *not* make sense is for a model to be fitted by minimizing the in-sample MSE, but for holdout forecasts to be evaluated using the MAPE. Let us be consistent: if our aim is solely to minimize the MAPE, models should be fitted using the in-sample MAPE as an optimization criterion.

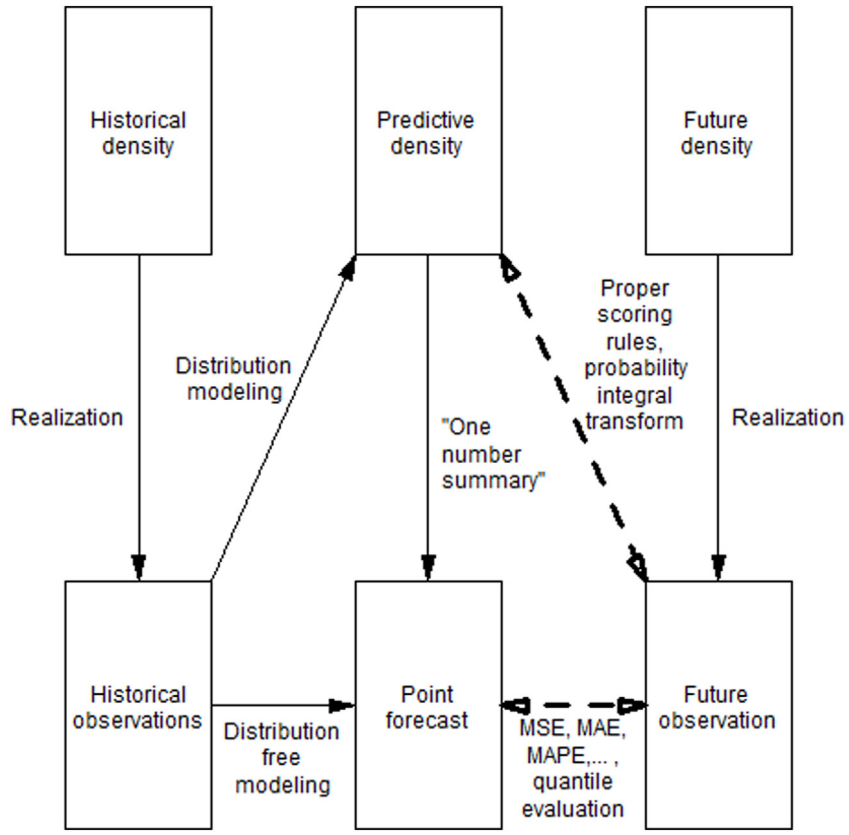

Fig. 2. The relationships among densities, observations, point forecasts and evaluations. I propose shifting our emphasis from the horizontal dashed arrow (assessing point forecasts on actual observations) to the diagonal one (assessing predictive densities against observations), with the ultimate aim of evaluating how close a predictive density comes to the actual density.

- We should move towards assessing full predictive densities and investigating how well they predict the true future density, of course via the actual future observations. Compare Fig. 2: I propose shifting our emphasis from the horizontal dashed arrow to the diagonal one. The M4's solicitation of quantile forecasts is a step in the right direction, but I hope that the M5 will request participants to submit (discretized) full predictive densities, e.g., by submitting the 1st,..., 99th predictive percentiles.
- We should therefore also aim to understand predictive densities in themselves better. The M5 would be an ideal opportunity to advance this research agenda by requesting predictive densities.
- More research into (and use of) methods for assessing the quality of predictive densities, e.g., proper scoring rules or probability integral transforms (Gneiting & Katzfuss, 2014), would also be welcome.
- New PFEMs are still being proposed. In the future, such proposals should always include a discussion of the relationship between the predictive density and the proposed PFEM and explicitly note which functional of the future density the proposed PFEM is aimed at (and why eliciting this functional is worthwhile).
- Although the present discussion is quite an abstract matter, the relationship between predictive densities and PFEMs should also be dealt with in introductory forecasting courses and textbooks.

In conclusion, although there is not yet widespread agreement on the merits and drawbacks of various PFEMs, I believe that viewing PFEMs through the lens of predictive densities provides a way forward, and that agreement is definitely possible.

References

Choi, K. P. (1994). On the medians of gamma distributions and an equation of Ramanujan. *Proceedings of the Americal Mathematical Society*, *121*, 245–251.

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

Gneiting, T. (2011b). Quantiles as optimal point forecasts. *International Journal of Forecasting*, *27*, 197–207.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

Hanley, J. A., Joseph, L., Platt, R. W., Chung, M. K., & Belisle, P. (2001). Visualizing the median as the minimum-deviation location. *The American Statistician*, *55*, 150–152.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688.

Kolassa, S. (2019). What is the best point forecast for gamma distributed data? Cross Validated (version: 2019-02-21). https://stats.stackexchange.com/q/389318.

Kolassa, S., & Martin, R. (2011). Percentage errors can ruin your day (and rolling the dice shows how). *Foresight: The International Journal of Applied Forecasting*, *23*, 21–29.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74.

Stephan Kolassa is a Data Science Expert at SAP Switzerland. His responsibilities include the algorithmic aspects of SAP's retail forecasting and planning solutions. He is an Associate Editor for *Foresight: The International Journal of Applied Forecasting*.