Density calibration with consistent scoring functions

Roberto Casarin and and Francesco Ravazzolo

Abstract This contribution studies a calibration approach for predictive densities based on generalized scoring rules. We consider a set of simulated experiments in order to study the effectiveness of the method.

Questo lavoro studia un approccio di calibrazione per densità previsive basato su regole di scoring generalizzate. Considera una serie di esperimenti simulati per studiare l'efficacia del metodo.

Key words: Density calibration, Predictive distributions, Scoring rules.

1 Introduction

When multiple forecasts are available from different models or sources it is possible to combine these in order to make use of all relevant information on the variable to be predicted and, as a consequence, to produce better forecasts. This is particular important when working with large database and selection of relevant information *a priori* is not an easy task. Early papers on forecasting with model combinations are [1], who considered air passenger data, and [17] who introduced a distribution which includes the predictions from two experts (or models). This latter distribution is essentially a weighted average of the posterior distributions of two models and is similar to the result of a Bayesian Model Averaging (BMA) procedure, see [15]. [14] extend the BMA framework by introducing a method for obtaining probabilistic forecasts from ensembles in the form of predictive densities and [12] extend it to Bayesian predictive synthesis.

[3] deal with the combination of predictions from different forecasting models using descriptive regression. [9] extend this and propose to combine forecasts with unrestricted regression coefficients as weights. [18] generalize the problem to a state space with weights that are assumed to follow a random walk process. [11] propose robust time-varying weights and account for both model and parameter uncertainty

Roberto Casarin
University Ca' Foscari of Venice e-mail: r.casarin@unive.it

Francesco Ravazzolo
Free University of Bozen-Bolzano, BI Norwegian Business School and RCEA e-mail: Francesco.Ravazzolo@unibz.it

in model averaging. [13] derive time-varying weights in dynamic model averaging, and speed up computations by applying forgetting factors in the recursive Kalman filter updating.

Combination weights that depend on (optimal) score functions have also been studied. [10] introduce the Kullback-Leibler divergence as a unified measure for the evaluation and suggest weights that maximize such a distance, see also [6] for a comprehensive discussion on how such weights are robust to model incompleteness, that is the true model is not included in the model set. [8] recommend strictly proper scoring rules. [4] develops a general method that can deal with most of issues discussed above, including time-variation in combination weights, learning from past performance, model incompleteness, correlations among weights and joint combined predictions of several variables.

Finally, the last aspect relates to calibration and combinations. [16] and [7] introduces the idea of recalibration density forecasts when the density is not well-calibrated. They introduce a monotone non-decreasing map via a Beta distribution to achieve it. [2] generalize to Beta mixtures, allowing for more flexibility in calibrating and combinations in presence of fat tails, skewness and multiple-modes.

This paper extends the density calibration literature and proposes to apply consistent scoring functions when calibrating models. We follow [5] and consider three different consistent scoring functions. These functions are minimized to compute the parameters of a beta calibration scheme. We study in simulation exercises the effectiveness of the method.

The structure of the paper is organized as follows. Section 2 presents the optimal calibration method. Section 3 provides numerical examples and directions for future research.

2 Optimal calibration

Consider the forecast distribution F_1 from a predictive model and F the distribution of Y, the variable to forecast. One can consider the following map

$$(\theta,\xi) \mapsto D(\theta,\xi) = \mathbb{E}_{\mathbb{Q},\xi}(S_{\alpha,\theta}(X,Y)) \tag{1}$$

where X is a point forecast from F_1, α a quantile level, $\theta \in \Theta \subset \mathbb{R}$ a threshold parameter and $\xi \in \Xi \subset \mathbb{R}^k$ a combination/calibration parameter vector. If the parameter ξ is indexing a family of continuous distributions $H_{\xi,F_1}(X) = (G_\xi \circ F_1)(X)$, with $x \mapsto G_\xi(x) \in (0,1)$ $\xi \in \Xi$ a sequence of non-decreasing functions with $G(0) = 0$ and $G(1) = 1$, then we obtain a calibration scheme. Our optimal calibration scheme can be defined as

$$\theta \mapsto \inf_{\xi \in \Xi} D(\theta,\xi) \tag{2}$$

In this study we follow [2] and assume $G_\xi(x)$ is the cumulative distribution function (cdf) of a beta distribution $B(x;\mu\phi,(1-\mu)\phi)$ with parameters $\xi = (\mu\phi,(1-\mu)\phi)$.

It follows that the calibrated density is $h_{\xi,F_1}(X) = (g_\xi \circ F_1)(X) f_1(X)$ where f_1 and g_ξ are the probability density functions of H_ξ and F_1. We denote with $\xi(\theta)$ the solution of Eq. (2).

The scoring function $S_{\alpha,\theta}(X,Y)$ can assume different forms. Following [5], we investigate three different consistent specifications. The first one is an elementary weighted average over elementary or extreme scores:

$$S_\theta(X,Y) = (Y-\theta)_+ - (Y-\theta)_+ - \mathbb{I}(x>\theta)(Y-X) \tag{3}$$

with $(t)_+ = max(t,0)$ and $\mathbb{I}(A)$ the indicator of the event A.

The second specification refers to a quantile consistent scoring function:

$$\begin{aligned} S^q_{\alpha,\theta}(X,Y) &= \{\mathbb{I}(Y<X)-\alpha\}\{\mathbb{I}(\theta<X)-\mathbb{I}(\theta<Y)\} \\ &= \begin{cases} 1-\alpha, & Y \le \theta < X \\ \alpha, & X \le \theta < Y \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

The third specification relies to an expectile consistent representation:

$$\begin{aligned} S^e_{\alpha,\theta}(X,Y) &= |\mathbb{I}(Y<X)-\alpha|\{(Y-\theta)_+ - (X-\theta)_+ - (Y-X)\mathbb{I}(\theta<X)\} \\ &= \begin{cases} (1-\alpha)|Y-\theta|, & Y \le \theta < X \\ \alpha|Y-\theta|, & X \le \theta < Y \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

We apply model in (2), with a beta calibration scheme and scoring functions in (3)-(5).

3 Numerical Illustration

The assume $Y_i \sim \mathcal{G}a(1,1)$ $i = 1,\dots,N$, where $\mathcal{G}a(a,b)$ denotes a gamma distribution with mean ab. The misspecified model is alternatively in the same distribution family $\mathcal{G}a(2,1)$ or in the lognormal distribution family $\Lambda(2,2)$. The score function $\mathbb{E}_{\mathbb{Q},\xi}(S_{\alpha,\theta}(X,Y))$ is evaluated on the data

$$\widehat{D(\theta,\xi)}_N = \frac{1}{N}\mathbb{E}_{\mathbb{Q},\xi}(S_{\alpha,\theta}(X,Y_i)) == \frac{1}{N}\sum_{i=1}^{N}\int S_{\alpha,\theta}(x,Y_i)h_{\xi,F}(x)dx \tag{6}$$

and the expectation approximated with M Monte Carlo samples from the predictive distribution, that is

$$\widehat{D(\theta,\xi)}_{M,N} = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N} S_{\alpha,\theta}(X_j,Y_i) \xrightarrow[M\to\infty]{as} \widehat{D(\theta,\xi)}_N \tag{7}$$

which converges to the empirical scoring function by the SLLN. Samples from the calibrated predictive distribution are obtained by inverse cdf methods, that is we first generate U from the standard uniform $\mathscr{U}(0,1)$ and then $X = F^{-1}(G_\xi^{-1}(U))$ where F^{-1} and G_ξ^{-1} are the inverse cdf of the misspecified model and of the beta distribution. The validity of the method follows from $P(X \leq x) = P(F^{-1}(G_\xi^{-1}(U)) < x) = P(G_\xi^{-1}(U)) < F(x)) = P(U < G_\xi(F(x))) = G_\xi(F(x)))$. In the numerical experiments we set $N = 100$ and $M = 1000$.

Fig. 1 reports the Murphy's diagrams for three scoring functions in (3)-(4)-(5) of the true model, the misspecified gamma models and the optimal calibration scheme. Results indicate that the calibrated forecasts is closer to the true model in all cases.

References

1. Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society, Series A*, **126**, 255–259.
2. Bassetti, F., Casarin, R., and Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, **113(522)**, 675–685.
3. Bates, J. and Granger, C. (1969). The combination of forecasts. *Operations Research Quarterly*, **20(4)**, 451–468.
4. Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, **177**, 213–232.
5. Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78(3)**, 505–562.
6. Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, **164**, 130–141.
7. Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747–1782.
8. Gneiting, T. G. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102(477)**, 359–378.
9. Granger, C. W. J. and Ramanathan, R. (1984). Improved Methods of Combining Forecasts. *Journal of Forecasting*, **3**, 197–204.
10. Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, **23**, 1–13.
11. Hoogerheide, L., Kleijn, R., Ravazzolo, R., van Dijk, H. K., and Verbeek, M. (2010). Forecast Accuracy and Economic Gains from Bayesian Model Averaging using Time Varying Weights. *Journal of Forecasting*, **29(1-2)**, 251–269.
12. McAlinn, K. and West, M. (2018). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, forthcoming.
13. Raftery, A., Karny, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, **52**, 52–66.
14. Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133**, 1155–1174.
15. Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92(437)**, 179–91.
16. Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72(1)**, 71–91.

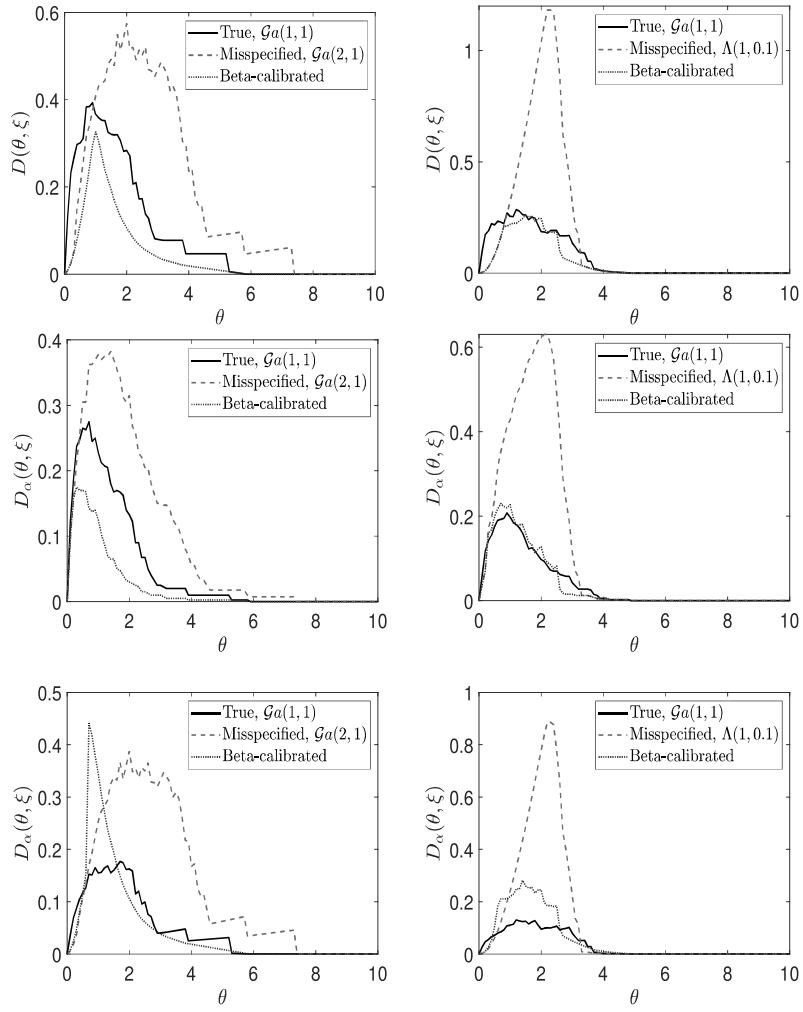

Fig. 1 Murphy's diagrams for score functions $S_\theta(X,Y)$ (first line), $S^q_{\alpha,\theta}(X,Y)$ (second line) and $S^e_{\alpha,\theta}(X,Y)$ (third line) with $\alpha = 0.25$. The misspecified models are in the Gamma (left) and lognormal (right) families of distributions. In each plot: the perfect (black solid), misspecified (red dotted) and calibrated (blue dashed) forecasters, where the calibrated forecaster is obtained by applying a beta calibration function to the misspecified model.

17. Roberts, H. V. (1965). Probabilistic prediction. *Journal of American Statistical Association*, **60**,50–62.
18. Terui, N. and van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, **18**, 421–438.