International Journal of Forecasting 32 (2016) 391–397

Contents lists available at ScienceDirect

International Journal of Forecasting

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

A note on the estimation of optimal weights for density forecast combinations

Laurent L. Pauwels, Andrey L. Vasnev *

University of Sydney, New South Wales 2006, Australia

ARTICLE INFO

Keywords:
Forecast combination
Density forecast
Optimization
Optimal weight
Discrete choice models

ABSTRACT

The problem of finding appropriate weights for combining several density forecasts is an important issue that is currently being debated in the forecast combination literature. A recent paper by Hall and Mitchell (2007) proposes that density forecasts be combined using the weights obtained from solving an optimization problem. This paper documents the properties of this optimization problem through a series of simulation experiments. When the number of forecasting periods is relatively small, the optimization problem often produces solutions that are dominated by a number of simple alternatives.

1. Introduction

The question of finding weights for combining density forecasts is non-trivial, and is currently being debated in the forecast combination literature. The latest work in this area is by Kapetanios, Mitchell, Price, and Fawcett (2015), and examples of early contributions are provided by Tay and Wallis (2000)and Corradi and Swanson (2006). The reader is also invited to peruse the review by Timmermann (2006) for a thorough review of the forecast combination literature.

In a recent paper, Hall and Mitchell (2007) propose a practical way of obtaining weights in a linear combination of density forecasts. The weights are found by maximizing the average logarithmic score of the combined density forecast. Hall and Mitchell (2007) call these weights "optimal" because they minimize the "distance" between the forecast and true (but unknown) densities, as measured by the Kullback–Leibler Information Criterion (KLIC). Although Hall and Mitchell (2007) show how these weights can be used, the paper does not detail the theoretical properties of the estimators of the weights. The motivation for the study relies on asymptotic theory, namely that the number of time periods grows to infinity ($T \to \infty$). Geweke and Amisano (2011) propose an approach that is similar to that of Hall and Mitchell (2007) using Bayesian methods, and provide a theoretical justification for the use of optimal linear combinations.

Several studies have followed in the footsteps of Hall and Mitchell (2007) in developing weighting techniques for density forecasts. For example, Jore, Mitchell, and Vahey (2010) develop log-score recursive weights for autoregressive models of output growth, inflation and interest rates. Similarly, Garratt, Mitchell, Vahey, and Wakerly (2011) apply these recursive weights to density forecasts of inflation in various industrialized countries. Bache, Jore, Mitchell, and Vahey (2011) employ weighting techniques similar to those of Hall and Mitchell (2007) for combining inflation forecast densities in linear opinion pools.

One would assume that the combination of various density forecasts implies that several density forecasts would be assigned positive weights in the combination. However, this paper finds that the "optimal weights" of Hall and Mitchell (2007) can behave unexpectedly when the number of forecasting periods is small. The weights can be such

* Correspondence to: Office 4160, Business School (H70), The University of Sydney, NSW 2006, Australia. Tel.: +61 2 9036 9435; fax: +61 2 9351 6409.

E-mail addresses: laurent.pauwels@sydney.edu.au (L.L. Pauwels), andrey.vasnev@sydney.edu.au (A.L. Vasnev).

http://dx.doi.org/10.1016/j.ijforecast.2015.09.002

that only one density is selected ("corner solution"), rather than combining the densities ("mixing solution"). Empirical work often provides evidence that combining densities is a better strategy than selecting one model. Kascha and Ravazzolo (2012) show that, although combinations do not always outperform individual models, they are beneficial because they are more accurate overall, and provide insurance against inappropriate model selection. Pauwels and Vasnev (2012) find that, when predicting the Fed's decisions to change the interest rate, the optimal weights of Hall and Mitchell (2007) select only one model for 41 forecasting periods. After 41 periods, each of the models is allocated a positive weight. While this result could be an artefact of the specific empirical study, it nonetheless begs for a formal investigation.

This paper examines the properties of Hall and Mitchell (2007) optimal weights when the number of forecasting periods is not infinite. Simple simulations provide clear insights; it turns out that "corner solutions" do occur frequently, but disappear as the number of forecasting periods increases ($T \to \infty$). The paper is organized as follows. An empirical illustration that motivates the questions raised in this paper is presented in Section 2. Section 3 provides simulation results to support the argument made in the paper. Section 4 concludes.

2. Empirical illustration: Predicting FOMC monetary policy decisions

The following empirical illustration discusses probability density forecast combinations, including the combination using the optimal weights proposed by Hall and Mitchell (2007). Early attempts to work with combinations of probability forecasts have been made in the context of aggregating probability distributions of expert opinions, as was discussed by Genest and Zidek (1986) and Clemen and Winkler (1999).

Pauwels and Vasnev (2012) use a conditional ordered probit model to estimate the dynamics of the federal funds target rate changes, following in the steps of Dueker (1999), Hamilton and Jordà (2002), Monokroussos (2011), Hu and Phillips (2004a), Kim, Jackson, and Saba (2009) and Kauppi (2012). Dueker (1999) uses the model

$$r_t^* = \boldsymbol{x}_{t-1}'\boldsymbol{\beta} - u_t \tag{1}$$

$$y_t^* = r_t^* - r_{t-1}, \tag{2}$$

where $u_t \sim \mathrm{N}(0, \sigma^2)$, both y_t^* and r_t^* are unobservable, and $\boldsymbol{x}_{t-1}$ contains observable information that is relevant to the forecast, including initial claims for unemployment insurance, annual growth of M2, consumer confidence, and annual growth of manufacturers' new orders.

In Eq. (2), r_t^* is the optimal policy rate, which is assumed to exist. r_t is the federal funds target rate set by the Federal Open Market Committee (FOMC) at its last meeting. Only the FOMC meeting months are forecasted. The time period used in this example is from January 1994 to April 2010, which represents 133 FOMC meetings.[1]

The Fed's decisions about the target interest rate are classified into three categories: "cut", "no change" and "hike". Hence,

$$y_t = \begin{cases} -1 & \text{if } y_t^* < \mu_1 \\ 0 & \text{if } \mu_1 \le y_t^* \le \mu_2 \\ 1 & \text{if } y_t^* > \mu_2, \end{cases} \tag{3}$$

is the observed decision of the Fed. For example, if the difference between the optimal policy rate (r_t^*) and the actual federal funds target rate (r_{t-1}) is greater than the threshold μ_2, then the model would predict a rate hike ($y_t = 1$).[2]

In the discrete choice model with the error distribution Φ, the probability distribution of y_t, $\Pr(y_t = j)$, depends on $(\boldsymbol{x}_t; \boldsymbol{\theta})$ with the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mu_1, \mu_2, \sigma^2)'$. For simplicity, it is denoted $P_{j,t}(\boldsymbol{x}_t; \boldsymbol{\theta})$. The parameters are estimated by maximizing the log-likelihood for the multiple-choice model.

Model combination is performed as follows. At each time t, each model $i \in \{1, \dots, N\}$ produces a probability forecast $P_{j,t}^{(i)}(\boldsymbol{x}_t^{(i)}; \boldsymbol{\theta}^{(i)})$ for each state $j = -1, 0, 1$. The vector of covariates $\boldsymbol{x}_t^{(i)}$ and the parameter vector $\boldsymbol{\theta}^{(i)}$ can be different for each model. Hence, the combined one-step-ahead probability forecast, $\hat{\boldsymbol{P}}_t^{(c)}$, simply follows from

$$\hat{\boldsymbol{P}}_t^{(c)} = \sum_{i=1}^{N} w_i \hat{\boldsymbol{P}}_t^{(i)}(\boldsymbol{x}_t^{(i)}; \hat{\boldsymbol{\theta}}^{(i)}),$$

where $\hat{\boldsymbol{P}}_t^{(i)} = \left(\hat{P}_{-1,t}^{(i)}, \hat{P}_{0,t}^{(i)}, \hat{P}_{1,t}^{(i)}\right)'$ is a 3×1 vector of estimated probabilities, $\hat{\boldsymbol{\theta}}^{(i)}$ is the estimated parameter vector of $\boldsymbol{\theta}^{(i)}$, and w_i is a scalar that weights model i. The weights w_i are non-negative and sum to one. Note that the notation w_i is used for simplicity, as the weights can change over time.

Among other methods, the weights w_i can be constructed using the weights proposed by Hall and Mitchell (2007). We denote those weights as w_i^* and call them optimal, following the terminology of Hall and Mitchell (2007), but introduce them formally in the next section. Alternatively, the weights can be constructed by ranking the scores of the models' forecasting performances, as was proposed by Pauwels and Vasnev (2012). If the log score is used to evaluate the performance, then the weights are

$$w_i^{PV} = \frac{1/|\bar{S}_i^l|}{\sum_{i=1}^{N} 1/|\bar{S}_i^l|} \quad i = 1, \dots, N, \tag{4}$$

where the log scores $\bar{S}_i^l$ are averaged over all one-step-ahead forecasts.[3]Hence, the better the score for a

[1] Pauwels and Vasnev (2012) present various robustness checks, including forecasting up to December 2008, which was the last month in which the Fed used the basis point target, before switching to the interval target.

[2] When the vector $\boldsymbol{x}_t$ contains integrated processes, the thresholds can be scaled by the sample size, as was shown by Hu and Phillips (2004a,b) and applied by Pauwels and Vasnev (2012).

[3] If state j happens, then the log-score is given by $S^l = \log(\hat{P}_j)$, similarly to the study by Ng, Forbes, Martin, and McCabe (2013). For multiple one-step-ahead forecasts, the logarithmic scores are averaged over the number of forecasted periods for each model i.