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 ($T \rightarrow \infty$). 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
0169-2070/Crown Copyright © 2015 Published by Elsevier B.V. on behalf of International Institute of Forecasters. All rights reserved.

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

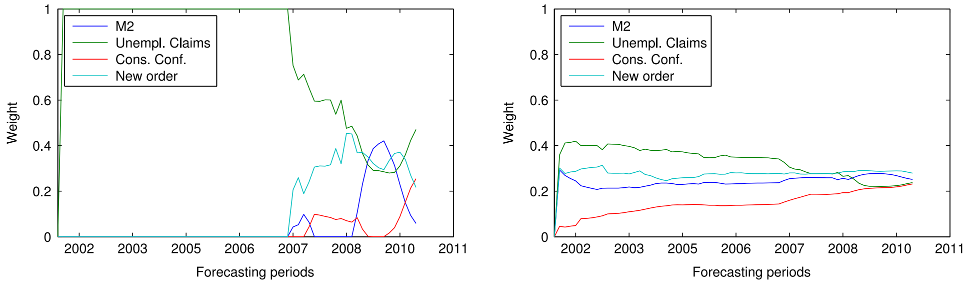
$$w_i^{PV} = \frac{1/|\bar{S}_i^l|}{\sum_{i=1}^{N} 1/|\bar{S}_i^l|} \qquad i = 1, \dots, N, \tag{4}$$

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.



(a) Optimal weights w_i^*.

(b) Alternative weights w_i^{LP}.

Fig. 1. Weights corresponding to the univariate models in the forecast combination. (For the color version of this figure, the reader is referred to the electronic version of this article.)

Table 1
Out-of-sample performances of the forecasts between May 2002 and April 2010.

Models	Scores		
	Log	Quad	Eps
H&P	−1.40	0.33	0.34
Equal weights	−0.86	0.46	0.28
Optimal weights w_i^*	−0.88	0.47	0.28
Alternative weights w_i^{PV}	−0.83	0.48	0.27
Univariate models:			
M2	−1.04	0.36	0.35
Unemployment claims	−1.10	0.44	0.30
Consumer Confidence	−1.12	0.33	0.37
New orders	−0.93	0.40	0.32

Notes: The numbers in the table are the Log, Quadratic and Epstein scoring rules, as used by Pauwels and Vasnev (2012). The scores are better for bigger Log and Quadratic numbers and for smaller Epstein numbers. The four variables used for the univariate models and the combination models correspond to the four variables selected by Hu and Phillips (2004a). *H&P* is a multivariate model with all four variables. *Equal weights* combines the probability forecasts of the univariate models equally. *Optimal weights* and *Alternative weights* refer to the models combining probability forecasts. Each univariate model features one of the listed variables as the main covariate. Only the FOMC meeting months are forecasted.

Table 2
Out-of-sample performances of the forecasts between May 2009 and April 2010.

Models	Scores		
	Log	Quad	Eps
Optimal weights w_i^*	−0.70	0.40	0.21
Alternative weights w_i^{PV}	−0.69	0.42	0.20

Notes: There are eight meetings during the period from May 2009 to April 2010. For further details, refer to the notes to Table 1.

forecasting model, the greater the weight given to its one-step-ahead forecast. Furthermore, the composition of the weights changes over time as the scores are averaged. Quadratic, Epstein and Brier scores can be used to construct alternative weights, see Pauwels and Vasnev (2012).

Fig. 1 shows the changes in the weights for the four models, with each featuring one covariate. Fig. 1(a) shows the weights w_i^*, and Fig. 1(b) displays the weights w_i^{PV}. For the optimal weight w_i^* of Hall and Mitchell (2007), we see that all of the weight is on unemployment claims for 41 of the 67 forecast FOMC meeting outcomes, with the three other covariates receiving zero weight. It is only once 41 periods have been forecast that the other models receive non-zero weights. In contrast, when using w_i^{PV}, the weights are shared across the four models, with unemployment claims receiving the largest weight (40%). Furthermore, in this particular empirical illustration, the forecast combination model using w_i^{PV} tends to outperform the one using w_i^* (see Table 1). After a lengthy training period, however, the weights w_i^* start to perform as well as w_i^{PV}, see Table 2.

Two important questions arise from this illustration. First, why do the optimal weights w_i^* select one model but neglect others for the first 41 one-step-ahead forecast periods? This would suggest that the one forecasting model should outperform forecast combinations for at least the first 41 periods. Second, does this result hold in general? The next section attempts to shed some light on these questions.

3. Analysis of optimal weights in simulations

Hall and Mitchell (2007) propose a set of weights for density forecast combination by maximizing the average logarithmic score of the combined density forecast. They call the weights "optimal" because they minimize the distance, measured by the estimated Kullback–Leibler Information Criterion (KLIC), between the combined forecast density and the unknown true density. A similar idea is used by Geweke and Amisano (2011). We follow Hall and Mitchell (2007) and use the optimal weights terminology, though optimality theory is not provided.

Suppose that there are N density forecasts, $g_{it}(\cdot)$, produced by models or analysts $i = 1, \dots, N$ of a real-valued variable y_t at time t, where $t = 1, \dots, T$ and T is the total number of forecasted periods.[4] The combined density forecast is defined as the finite mixture

$$p_t(\cdot) = \sum_{i=1}^{N} w_i g_{it}(\cdot), \tag{5}$$

where w_i are a set of non-negative weights that sum up to one.

[4] In the empirical illustration in Section 2, the density forecasts of y_t are discrete, which means that $g_{it}(y_t)$ is the forecasted probability of the observed outcome $P_{j,t}(y_t = j)$.

Definition 1 of Hall and Mitchell (2007) gives the weights vector $\boldsymbol{w}^* = (w_1^*, \ldots, w_N^*)$ as the solution of the optimization problem

$$\boldsymbol{w}^* = \arg \max_{(w_1,\ldots,w_N)} \frac{1}{T} \sum_{t=1}^{T} \ln p_t(y_t), \tag{6}$$

where $\frac{1}{T}\sum_{t=1}^{T} \ln p_t(y_t)$ is the average logarithmic score of the combined density forecast over the sample $t = 1, \ldots, T$.

Everitt (1996) lists the numerous difficulties that are associated with a general finite mixture problem, including slow convergence, a failure to reach the global optimum, and the absence of a solution. Hall and Mitchell (2007) warn that the optimization problem 'can be difficult'. We now investigate these difficulties in simulations.

3.1. Autocorrelated time series

We follow the experimental design of Smith and Wallis (2009, Section 3.1) closely; specifically, their case 2. We draw a sequence of $T + 1$ observations from a strictly stationary AR(2) process

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \epsilon_t \quad (t = 1, \ldots, T + 1),$$

where the $\{\epsilon_t\}$ are independent and identically distributed standard-normal variates and ϕ_1 and ϕ_2 are given parameters that are subject to the stationarity conditions $\phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1$, and $|\phi_2| < 1$. The variance of the process is given by

$$\sigma_z^2 = \mathrm{var}(z_t) = \frac{1 - \phi_2}{(1 + \phi_2)[(1 - \phi_2)^2 - \phi_1^2]},$$

and the first two autocorrelation coefficients are

$$\rho_1 = \mathrm{corr}(z_t, z_{t-1}) = \frac{\phi_1}{1 - \phi_2} \quad \text{and}$$

$$\rho_2 = \mathrm{corr}(z_t, z_{t-2}) = \phi_1 \rho_1 + \phi_2.$$

The aim is to forecast z_{T+1}, and two point forecasts are easily available:

$$y_1 = \rho_1 z_T \quad \text{and} \quad y_2 = \rho_2 z_{T-1}.$$

The forecast errors are

$$e_1 = z_{T+1} - y_1 \quad \text{and} \quad e_2 = z_{T+1} - y_2,$$

with variances

$$\sigma_1^2 = \mathrm{var}(e_1) = \sigma_z^2(1 - \rho_1^2) \quad \text{and}$$

$$\sigma_2^2 = \mathrm{var}(e_1) = \sigma_z^2(1 - \rho_2^2),$$

respectively. The density forecasts can be defined as

$$g_1(z) = \psi(z; y_1, \sigma_1^2) \quad \text{and} \quad g_2(z) = \psi(z; y_2, \sigma_2^2),$$

where $\psi(z; \mu, \sigma^2)$ is the normal density function with a mean of μ and a variance of σ^2, and the combined density forecast is

$$p(z) = w g_1(z) + (1 - w) g_2(z).$$

Because there are only two models, the subscript i for the weights is redundant. Thus, we can use w to denote the weight allocated to $g_1(z)$.

We are interested in the properties of the estimated weight w^* and the corresponding density forecasts for different values of ϕ_1 and ϕ_2 across different values of T. Forecast combinations using equal weights and weights w^{PV} are used for comparison. The forecasting horizon T varies from 3 to 50, and the weights w^* given by Eq. (6) and w^{PV} given by Eq. (4) are estimated using historic observations $(z_1, \ldots, z_T)$.

In order to obtain a better understanding of the uncertainty caused by the estimation of weights, we do *not* estimate the parameters ϕ_1 and ϕ_2. Instead, these are set to their true values, such that any uncertainty shown in the simulations is caused by weight estimation.

This experiment is repeated 10,000 times. For given values of ϕ_1 and ϕ_2, each replication produces the following forecasts for every point T:

1. density forecast $g_1(z)$ from model 1,
2. density forecast $g_2(z)$ from model 2,
3. combined density forecast $p(z)$ with equal weight $w = 1/2$,
4. combined density forecast $p(z)$ with $w = w^*$,
5. combined density forecast $p(z)$ with $w = w^{PV}$.

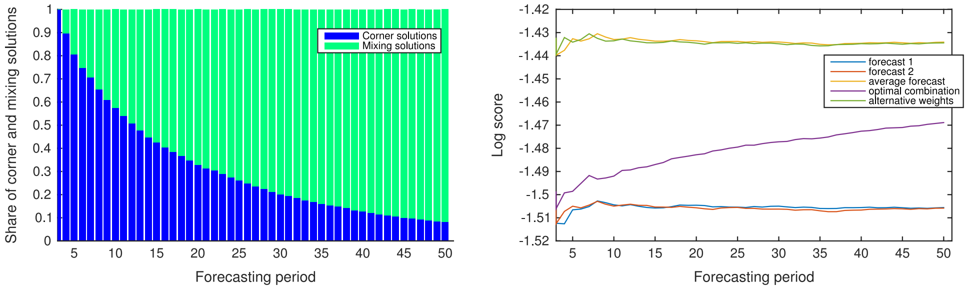
The forecasts are then evaluated at point z_{T+1} using log scores. The results are averaged across all simulations. In addition, we look at the percentage of corner solutions produced by the optimization problem.

Fig. 2 presents the results for $\phi_1 = \phi_2 = 0.4$. Fig. 2(b) shows that the performances of g_1, g_2, and forecasts with equal and w^{PV} weights are stable across different values of T. The performance of the combination with weight w^* continues to improve as more historical observations become available. This improvement can be attributed to the declining incidence of the corner solutions produced by the optimization problem, as shown in Fig. 2(a). Although the corner solutions in this situation are not theoretically optimal (because the data generating process (DGP) is an AR(2) model), they are observed frequently in the simulations. This property of the optimization problem corrects itself gradually as more historical observations become available.

Fig. 3 presents the results for $\phi_1 = 0.5, \phi_2 = -0.8$. Again, the behaviors of the forecasts are stable, with the exception of the combination with weight w^*, see Fig. 3(b). The performance of the combination with the weight w^* quickly improves as more historical observations become available. In this situation, forecast g_2 is considerably more important than forecast g_1, so it is not surprising that corner solutions represent half of the solutions even for $T = 50$ (see Fig. 3(a)). However, for $T < 15$, the corner solutions still have a negative impact on the performance of w^*.

The numerical results are presented in Tables 3 and 4. In Table 3, $\phi_1 = \phi_2$ for values ranging between −0.9 and 0.4. In Table 4, $\phi_1 = 0.5$ and ϕ_2 ranges between −0.9 and 0.4. Because only the performance of the combination with the weight w^* changes across T, we report its log score for $T = 5, 25, 50$, while we report the log scores of the other forecasts only for $T = 50$.

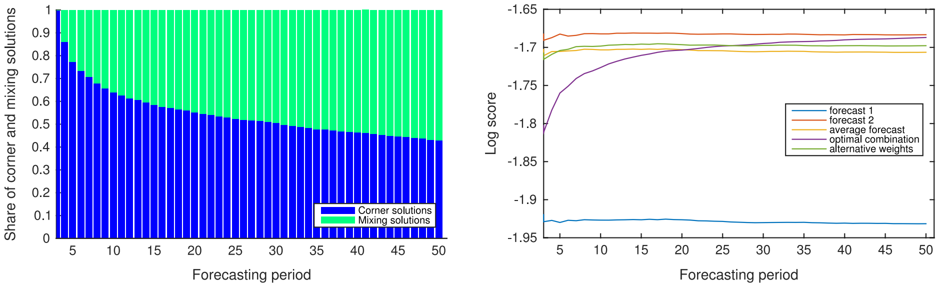
From Table 3, we can observe that when $\phi_1 = \phi_2$, the individual performances of the forecasts g_1 and g_2 are similar, but complement each other in the combinations with



(a) Share of corner solutions.

(b) Log scores of different forecasts.

Fig. 2. $\phi_1 = \phi_2 = 0.4$. (For the color version of this figure, the reader is referred to the electronic version of this article.)



(a) Share of corner solutions.

(b) Log scores of different forecasts.

Fig. 3. $\phi_1 = 0.5$, $\phi_2 = -0.8$. (For the color version of this figure, the reader is referred to the electronic version of this article.)

Table 3
Performances of difference forecasts, measured by the average log score when $\phi_1 = \phi_2 \in [-0.9, 0.4]$.

$\phi_1 = \phi_2$	Forecasts		Forecast combinations				
	g_1	g_2	$w = 1/2$	$w = w^{PV}$	$w = w^*$		
					$T = 5$	$T = 25$	$T = 50$
−0.9	−2.25093	−2.25019	−2.13054	−2.13362	−2.31262	−2.22474	−2.19715
−0.8	−1.93359	−1.93248	−1.83505	−1.83732	−1.96490	−1.91038	−1.88871
−0.7	−1.75641	−1.75569	−1.67809	−1.67980	−1.77979	−1.73746	−1.72142
−0.6	−1.64407	−1.64316	−1.58384	−1.58470	−1.64991	−1.62915	−1.61791
−0.5	−1.56394	−1.56513	−1.52092	−1.52163	−1.56851	−1.55542	−1.54721
−0.4	−1.50540	−1.50568	−1.47613	−1.47630	−1.50964	−1.50056	−1.49499
−0.3	−1.46795	−1.46844	−1.45055	−1.45084	−1.46600	−1.46647	−1.46266
−0.2	−1.44024	−1.44092	−1.43217	−1.43199	−1.44294	−1.44179	−1.43814
−0.1	−1.42287	−1.42289	−1.42060	−1.42044	−1.43039	−1.42263	−1.42247
0	−1.41930	−1.41930	−1.41930	−1.41938	−1.41847	−1.41922	−1.41938
0.1	−1.42324	−1.42332	−1.42049	−1.42080	−1.42038	−1.42286	−1.42312
0.2	−1.43703	−1.43720	−1.42446	−1.42448	−1.43607	−1.43343	−1.43352
0.3	−1.46521	−1.46546	−1.43209	−1.43205	−1.46061	−1.45571	−1.45134
0.4	−1.50590	−1.50678	−1.43484	−1.43512	−1.50636	−1.48240	−1.46963

Notes: The table reports forecasts g_1, g_2 and combinations with equal ($w = 1/2$) and w^{PV} weights for $T = 50$. The improvement in the combined forecast with the weight w^* is reported for $T = 5, 25, 50$. The log scores are negative by definition, with numbers closer to zero indicating better performances.

equal or w^{PV} weights. The combination with the weight w^* produces worse results than the individual forecasts for small values of T, but its performance improves as more historical information becomes available.

Table 4 shows that one forecast is superior (i.e., forecast g_2 is better for $\phi_2 < -0.5$, and forecast g_1 is better for $\phi_2 > -0.5$), whereas the combinations with the equal and w^{PV} weights have similar performances. The performance of the combination with the weight w^* is inferior for small values of T but improves quickly, reaching the level of the best forecast when $T = 50$. This result is as expected, because the sum used in the optimization problem in Eq. (6) converges to the expected log score, meaning that the solution minimizes the KLIC distance between the true density and the combined probability forecast.

3.2. Markov switching data generating process

For simplicity, consider an AR(1) model

$$y_t^{(1)} = \rho y_{t-1}^{(1)} + \nu_t, \quad \nu_t \sim i.i.d.\mathrm{N}(0, 1), \tag{7}$$

Table 4
Performances of difference forecasts, measured by the average log score when $\phi_1 = 0.5$ and $\phi_2 \in [-0.9, 0.4]$.

ϕ_2	Forecasts		Forecast combinations				
	g_1	g_2	$w = 1/2$	$w = w^{PV}$	$w = w^*$		
					$T = 5$	$T = 25$	$T = 50$
−0.9	−2.25082	−1.84008	−1.91819	−1.89494	−1.94143	−1.85616	−1.84598
−0.8	−1.93156	−1.68331	−1.70640	−1.69774	−1.75995	−1.69819	−1.68697
−0.7	−1.75548	−1.61830	−1.60921	−1.60683	−1.67562	−1.63464	−1.61939
−0.6	−1.64088	−1.58219	−1.55324	−1.55345	−1.61127	−1.58862	−1.57783
−0.5	−1.56293	−1.56213	−1.51902	−1.51935	−1.56962	−1.55242	−1.54471
−0.4	−1.50597	−1.54849	−1.49467	−1.49460	−1.53285	−1.51800	−1.51103
−0.3	−1.46578	−1.53960	−1.47736	−1.47664	−1.50365	−1.48604	−1.47852
−0.2	−1.43895	−1.53400	−1.46479	−1.46344	−1.48122	−1.46012	−1.45213
−0.1	−1.42351	−1.53092	−1.45555	−1.45378	−1.46738	−1.44438	−1.43630
0	−1.41844	−1.52996	−1.44871	−1.44683	−1.46310	−1.43943	−1.43138
0.1	−1.42343	−1.53103	−1.44357	−1.44192	−1.46765	−1.44368	−1.43551
0.2	−1.43881	−1.53424	−1.43946	−1.43837	−1.47978	−1.45478	−1.44587
0.3	−1.46559	−1.53999	−1.43548	−1.43520	−1.49822	−1.47021	−1.45904
0.4	−1.50459	−1.54901	−1.42933	−1.43026	−1.52061	−1.48586	−1.47046

Notes: The table reports forecasts g_1, g_2 and combinations with equal ($w = 1/2$) and w^{PV} weights for $T = 50$. The improvement of the combined forecast with the weight w^* is reported for $T = 5, 25, 50$. The log scores are negative by definition, with numbers closer to zero indicating better performances.

with $\rho = 0.3$, and an MA(1) model

$$y_t^{(2)} = \varepsilon_t, \qquad \varepsilon_t = \theta\varepsilon_{t-1} + \nu_t, \qquad \nu_t \sim i.i.d.\mathrm{N}(0, 1), \qquad (8)$$

with $\theta = 0.7$, assuming that the parameters are known and therefore there is no estimation noise. The true DGP combines the two models in Eqs. (7) and (8) by switching from one model to the other. The switch in DGP between the models in Eqs. (7) and (8) is uncertain. The probability that the prevailing model that determines the DGP will remain the same in the next period is 0.7, and the probability that there will be a switch to the alternative model is 0.3.

This Markov switching DGP has a stationary state in which the system follows model 1 for half of the time and model 2 for the other half. In this situation, it is theoretically optimal to combine the two models, because neither of the models captures the true DGP on its own. This can be seen from the simulation results in Fig. 4, where there are almost no corner solutions after 36 periods. When $T = 36$, the average w^* across simulations is 0.499, reflecting the fact that each of the models captures the DGP on its own roughly half of the time. Note, however, that the optimization problem in Eq. (6) still yields corner solutions in 20% of the simulations even after 24 forecasting periods (two years of monthly data).

3.3. Mixing DGP

Finally, if the true DGP in each period is a mix of the AR(1) model in Eq. (7) and the MA(1) model in Eq. (8), and hence the actual observations are generated as an ARMA(1,1) model

$$y_t = \alpha y_t^{(1)} + (1 - \alpha)y_t^{(2)},$$

then the solution of the optimization problem in Eq. (6), w^*, should converge to α. This is in fact what is observed in both Fig. 5(a), where $\alpha = 0.5$ and the average weight after 36 forecasting periods is 0.503, and Fig. 5(b), where $\alpha = 0.3$ and the average weight after 36 forecasting periods is 0.26. Note also that the convergence is slower for $\alpha = 0.3$, with approximately 20% of corner solutions occurring after 36 forecasting periods.

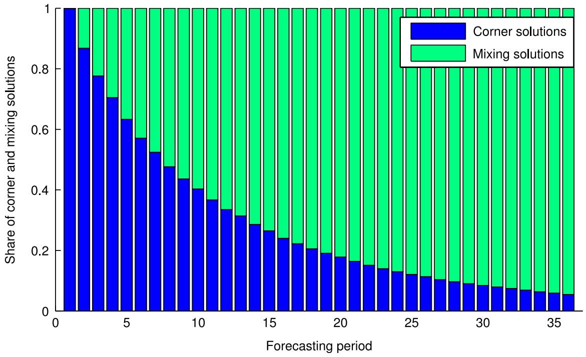
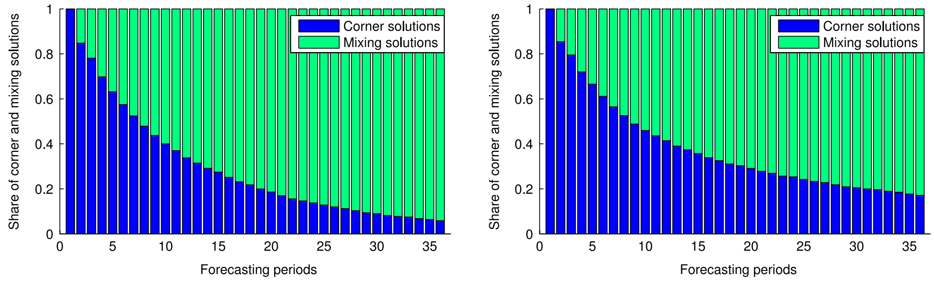


Fig. 4. The shares of corner and mixing solutions for two misspecified models across different forecasting periods. The DGP switches from one model to another with a probability of 0.3.

4. Concluding comments

The idea of using a training sample for parameter estimation before forecasting out-of-sample is acknowledged widely in the forecasting literature. The simulation and empirical results considered in this paper indicate the necessity of using a training sample for the optimal weights of Hall and Mitchell (2007) when combining forecasts. If no such training sample is used, one risks ending up with a corner solution. This is an artefact of the optimization problem given by Eq. (6) when the number of forecasting periods, T, is small. When T is sufficiently large, the asymptotic theory used by Hall and Mitchell (2007) and Geweke and Amisano (2011) to justify the optimal weights is valid, and the optimal weights have the expected properties. If one wishes the weights to behave as would be expected from theory, the authors' practical recommendation is to use at least 36 data points (three years of monthly data) when solving the optimization problem. Alternatively, one can use the weights proposed by Pauwels and Vasnev (2012), which do not need this extensive training period.



(a) $\alpha = 0.5$.

(b) $\alpha = 0.3$.

Fig. 5. The shares of corner and mixing solutions for two misspecified models across different forecasting periods. The DGP is mixing using parameter α, which is equal to the theoretically optimal weight.

Acknowledgments

The authors are grateful to the editor and an anonymous reviewer for their constructive comments. The authors would like to thank Rachida Ouysse, Valentyn Panchenko, Tommaso Proietti, Jaya Krishnakumar, Shubhabrata Das and Subramanian Chetan, as well as the participants of the Sydney Econometrics Reading Group (SERG) meeting at UNSW in November 2012 and the participants of the seminar at the Indian Institute of Management in Bangalore. The University of Sydney Business School's Research Grant 2012 is acknowledged.

References

Bache, I. W., Jore, A. S., Mitchell, J., & Vahey, S. P. (2011). Combining VAR and DSGE forecast densities. *Journal of Economic Dynamics and Control*, *35*, 1659–1670.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*, 187–203.

Corradi, V., & Swanson, N. R. (2006). Predictive density evaluation. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 197–284). North-Holland, Ch. 5.

Dueker, M. (1999). Measuring monetary policy inertia in target fed funds rate changes. *Federal Reserve Bank of St. Louis Review*, *81*(5), 3–9.

Everitt, B. S. (1996). An introduction to finite mixture distributions. *Statistical Methods in Medical Research*, *5*, 107–127.

Garratt, A., Mitchell, J., Vahey, S. P., & Wakerly, E. C. (2011). Real-time inflation forecast densities from ensemble Phillips curves. *The North American Journal of Economics and Finance*, *22*, 77–87.

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*, 114–148.

Geweke, J., & Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, *164*, 130–141.

Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, *23*, 1–13.

Hamilton, J. D., & Jordà, Ò (2002). A model of the federal funds rate target. *Journal of Political Economy*, *110*, 1135–1167.

Hu, L., & Phillips, P. C. B. (2004a). Dynamics of the federal funds target rate: A nonstationary discrete choice approach. *Journal of Applied Econometrics*, *19*, 851–867.

Hu, L., & Phillips, P. C. B. (2004b). Nonstationary discrete choice. *Journal of Econometrics*, *120*, 103–138.

Jore, A. S., Mitchell, J., & Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, *25*, 621–634.

Kapetanios, G., Mitchell, J., Price, S., & Fawcett, N. (2015). Generalised density forecast combinations. *Journal of Econometrics*, *188*, 150–165.

Kascha, C., & Ravazzolo, F. (2012). Combining inflation density forecasts. *Journal of Forecasting*, *29*, 231–250.

Kauppi, H. (2012). Predicting the direction of the Fed's target rate. *Journal of Forecasting*, *31*, 47–67.

Kim, H., Jackson, J., & Saba, R. (2009). Forecasting the FOMC's interest rate setting behavior: A further analysis. *Journal of Forecasting*, *28*, 145–165.

Monokroussos, G. (2011). Dynamic limited dependent variable modeling and U.S. monetary policy. *Journal of Money, Credit and Banking*, *43*, 519–534.

Ng, J., Forbes, C. S., Martin, G. M., & McCabe, B. P. M. (2013). Non-parametric estimation of forecast distributions in non-Gaussian, non-linear state space models. *International Journal of Forecasting*, *29*, 411–430.

Pauwels, L., & Vasnev, A. (2012). Forecast combination for discrete choice models: predicting FOMC monetary policy decisions. University of Sydney Business School working paper.

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*, 331–355.

Tay, A. S., & Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, *19*, 235–254.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). North-Holland, Ch. 4.

Laurent L. Pauwels received his Ph.D. in International Economics jointly from the Graduate Institute, Geneva and The University of Geneva. His main areas of research are in forecast combination, discrete choice modeling, structural breaks in time series and panels, estimation and inference in panels. His interests also include macroeconomic and exchange rate issues in China. During his doctoral studies, Laurent held positions briefly at the European Central Bank and the United Nations Economic Commission for Europe. Before joining the University of Sydney, he worked as an economist in the research department of the Hong Kong Monetary Authority.

Andrey L. Vasnev (Perm, 1976) graduated in Applied Mathematics from Moscow State University in 1998. In 2001 he completed his Master's degree in Economics in the New Economic School, Moscow. In 2006 he received Ph.D. degree in Economics from the Department of Econometrics and Operations Research at Tilburg University under the supervision of Jan R. Magnus. He worked as a credit risk analyst in ABN AMRO bank before joining the University of Sydney in 2008.