The University of Sydney Business School
The University of Sydney

BUSINESS ANALYTICS WORKING PAPER SERIES

Practical considerations for optimal weights in density forecast combination

Andrey L. Vasnev and Laurent L. Pauwels
University of Sydney

Abstract

The problem of finding appropriate weights to combine several density forecasts is an important issue currently debated in the forecast combination literature. Recently, a paper by Hall and Mitchell (IJF, 2007) proposes to combine density forecasts with optimal weights obtained from solving an optimization problem. This paper studies the properties of this optimization problem when the number of forecasting periods is relatively small and finds that it often produces corner solutions by allocating all the weight to one density forecast only. This paper's practical recommendation is to have an additional training sample period for the optimal weights. While reserving a portion of the data for parameter estimation and making pseudo-out-of-sample forecasts are common practices in the empirical literature, employing a separate training sample for the optimal weights is novel, and it is suggested because it decreases the chances of corner solutions. Alternative log-score or quadratic-score weighting schemes do not have this training sample requirement.

January 2013

BA Working Paper No: 01/2013
http://sydney.edu.au/business/business_analytics/research/working_papers

Practical considerations for optimal weights in density forecast combination

Andrey L. Vasnev and Laurent L. Pauwels
University of Sydney

This draft: January 2013

Abstract

The problem of finding appropriate weights to combine several density forecasts is an important issue currently debated in the forecast combination literature. Recently, a paper by Hall and Mitchell (IJF, 2007) proposes to combine density forecasts with optimal weights obtained from solving an optimization problem. This paper studies the properties of this optimization problem when the number of forecasting periods is relatively small and finds that it often produces corner solutions by allocating all the weight to one density forecast only. This paper's practical recommendation is to have an additional training sample period for the optimal weights. While reserving a portion of the data for parameter estimation and making pseudo-out-of-sample forecasts are common practices in the empirical literature, employing a separate training sample for the optimal weights is novel, and it is suggested because it decreases the chances of corner solutions. Alternative log-score or quadratic-score weighting schemes do not have this training sample requirement.

Key words: Forecast combination; Density forecast; Optimization; Optimal weight; Discrete choice models

1 Introduction

Since the early contributions on combining density forecasts, finding weights to combine density forecasts has been debated by the literature.[1] In a recent paper, Hall and Mitchell (2007) propose a practical way to select optimal weights in a linear combination of density forecasts, by maximizing the average logarithmic score of the combined density forecast. These optimal weights minimize the "distance" between the forecasted and true (but unknown) density, as measured by the Kullback–Leibler information criterion (KLIC). It shows how these optimal weights can be achieved but without detailing their theoretical properties. The motivation of that study relies on asymptotic theory, namely that the number of time periods grows to infinity ($T \to \infty$). Geweke and Amisano (2011) proposes a similar approach to Hall and Mitchell (2007) using Bayesian methods and provides the theoretical justification for using optimal weighting schemes in combining linear models.

The KLIC scores are used to evaluate forecast densities and have been used in the recent theoretical and empirical forecasting literature. Diks et al. (2010) develops a statistical test for comparing the predictive accuracy of competing copula specifications in multivariate density forecasts, based on out-of-sample KLIC scores. Diks et al. (2011) improves these testing techniques further with likelihood-based scoring rules. Jore et al. (2010) develops log-score recursive weights following Hall and Mitchell (2007), for vector autoregressive and autoregressive models of output growth, inflation and interest rates. Similarly, Garratt et al. (2011) applies these recursive weights for density forecasts of inflation in various industrialized countries. Wolden Bache et al. (2011) employs similar optimal weighting techniques to Hall and Mitchell (2007), for linear opinion pools to combine inflation forecast densities.

One would presume that when combining density forecasts with optimal weights, various density forecasts would receive positive weights in the combination rather than a single density forecast is chosen. This paper finds, however, that these "optimal weights" can behave unexpectedly when the number of forecasting periods is small by selecting one of the density forecasts rather than combining them with positive weight on each of the forecasts. In an empirical paper, Kascha and Ravazzolo (2012) find that combining densities is a better strategy than ex ante model selection. While it shows that combinations do not always outperform the best individual model, forecast combinations are more accurate and provide insurance against inappropriate model selection. This paper uses the empirical illustration of Pauwels and Vasnev (2012) to show that for the first 41 forecasting periods, one

[1] See Tay and Wallis (2000) and Corradi and Swanson (2006) for examples of early contributions. The reader is also invited to look at Timmermann (2006) for a thorough review of the forecast combination literature.

model is allocated all the weight while the other models have zero weight ("corner solutions"), and afterwards all models are allocated positive weights ("interior solutions"). While this could be an artefact of this particular empirical study, it nonetheless begs for formal investigation.

This paper investigates the performance of Hall and Mitchell (2007) optimal weights in combining density forecasts. It concentrates on the case when the number of forecasting periods is not infinite. Hence, it examines the theoretical properties of the optimal weights that minimize the *estimated* Kullback–Leibler Information Criterion. Tractable theoretical results can be obtained and geometric representations can be drawn up to two forecasting periods. When the forecast horizon is greater than or equal to three, theoretical properties are complex to derive and simple simulations actually provide better insights. It turns out that "corner solutions" do occur frequently, but ease out as the number of forecasting periods increase ($T \to \infty$), as expected in theory.

Although Hall and Mitchell (2007) considers continuous densities only, all theoretical findings found in this paper hold for both the continuous and the discrete cases. The empirical illustration, which motivates the questions raised in this paper, features discrete density forecast combinations. This illustration is presented in Section 2. Section 3 discusses the theoretical underpinnings. Section 4 provides simulation results to support the argument made in the paper. Section 5 concludes.

2 Empirical illustration: Predicting FOMC monetary policy decisions

The following empirical illustration discusses probability density forecast combinations with scoring rules as well as optimal weights proposed by Hall and Mitchell (2007). Early attempts to work with combinations of probability forecasts have been done in the context of aggregating probability distributions of expert opinions, as discussed in Genest and Zidek (1986) and Clemen and Winkler (1999). Pauwels and Vasnev (2012) uses 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 Jorda (2002), Monokroussos (2011), Hu and Phillips (2004a), Kim et al. (2009) and Kauppi (2012).

Following Dueker (1999), the model is

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

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

where $u_t \sim \mathrm{N}(0, \sigma^2)$ as in an ordered probit model, and both y_t^* and r_t^* are unobservable, $\boldsymbol{x}_{t-1}$ contains observable information relevant to the forecast including

initial claim for unemployment, annual growth of M2, consumer confidence, annual growth of manufacturers new orders.

The time period used in this example spans from January 1994 to April 2010, which represents 133 FOMC meetings.[2] Only the FOMC meeting months are forecasted. r_t^* is the optimal policy rate and it is assumed to exist. r_t is the federal funds target rate set by the FOMC in its last meeting. Fed decisions about the target interest rate are classified into three categories: "cut", "no change" or "hike." Hence,

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

is the observed decisions by 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 ($y_t = 1$). This divergence would need to be substantial to result in a change in the target rate as policy actions are often costly.[3]

In the discrete choice model with 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 as $P_j(\boldsymbol{x}_t; \boldsymbol{\theta})$. The parameters are estimated by maximizing the log-likelihood for the multiple choice model.

Model combination is done as follows. At every time period, t, each model $i \in \{1, \ldots, N\}$ 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 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

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

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. $\hat{\boldsymbol{\theta}}^{(i)}$ is the estimated parameter vector of $\boldsymbol{\theta}^{(i)}$ and ω_i is a scalar that weights model i. Note that the notation ω_i is used for simplicity as the weights can change over time and might be denoted as $\omega_{i,t}$.

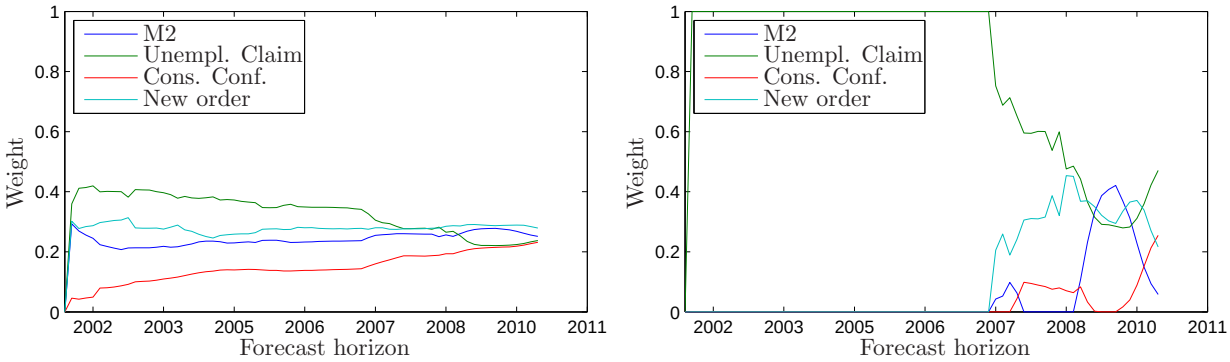
[2] Pauwels and Vasnev (2012) presents various robustness checks including forecasting up to December 2008, the last month the Fed used the basis point target before switching to the interval target.

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

The weights, ω_i, can be constructed, among other methods, by ranking the scores of each model's forecasting performance as proposed in Pauwels and Vasnev (2012) or by using optimal weights proposed by Hall and Mitchell (2007) and discussed in the next section. The log-score based weights, for example, are

$$\omega_i^l = \frac{1/|\bar{S}_i^l|}{\sum_{i=1}^{N} 1/|\bar{S}_i^l|} \quad i = 1, \ldots, N$$

where ω_i^l are the weight for forecast i based on the log-score $\bar{S}_i^l$ averaged over all one-step-ahead forecasts.[4] Hence, the better the score for a forecasting model, the higher the weight given to its one-step ahead forecast. Furthermore, the composition of the weights changes over time as the scores are averaged. See Pauwels and Vasnev (2012) for quadratic, Epstein and Brier score based weight.[5]



(a) Log score weights (b) Optimal weights

Figure 1: Weights corresponding to the univariate models in forecast combination.

Figure 1 shows the changes in the weights for the 4 models each featuring one covariate. Figure 1a displays the weights computed with a log-scoring rule and Figure 1b shows the optimal weights. In the optimal weight case, all the weight is on unemployment claims for 41 out of 67 forecasted FOMC meeting outcomes, while the three other covariates receive zero weight. It is only after 41 forecasted periods that the other models receive non-zero weight. In contrast, when using the log score scheme, the weights are shared across the 4 models, with unemployment

[4] If state j happens, then the log-score is given by $S^l = \log(\hat{P}_j)$, similarly to Ng et al. (2010). For multiple one-step ahead forecasts, the logarithmic scores are averaged over the number of forecasted periods for each model i over the period $[\tau_1 + 1, \tau_2]$: $\bar{S}_i^l = \frac{1}{\tau_2 - \tau_1} \sum_{t=\tau_1+1}^{\tau_2} S_{it}^l$, where S_{it}^l is the log-score obtained for model i at time t.

[5] Note that in a recent paper, Boero et al. (2011) provides a practical way to evaluate of some leading density forecast scoring rules such as Epstein, Brier and logarithmic rules, in the context of forecast surveys of UK inflation.

claim receiving the largest weight (40%). Furthermore, in this particular empirical illustration, the forecast combination model using log-score weighting tends to out-perform the one with optimal weighting scheme (see Table 1). After a lengthy training period, however, the optimal weights start to perform well as shown in Table 2.

Table 1: Out-of-Sample Forecasts

Predictions between May 2002 – April 2010			
Models|Scores	Log	Quad	Eps
H&P	-1.40	0.33	0.34
Equal weights	-0.86	0.46	0.28
Log weights	-0.83	0.48	0.27
Optimal weights	-0.88	0.47	0.28
	Univariate Models		
M2	-1.04	0.36	0.35
Unemployment claim	-1.10	0.44	0.30
Consumer Confidence 4	-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 in Pauwels and Vasnev (2012). The scores are higher for bigger Log and Quadratic numbers, and for smaller Epstein numbers. The four variables used for the univariate model and the combination models correspond to the 4 variables selected by Hu and Phillips (2004a). *H&P* is a univariate model with all 4 variables. *Equal weights* combines the probability forecast of the univariate models equally. *Log weights* and *Optimal weights* refer to the models combining probability forecasts. Each univariate model features one of the listed variables as a main covariate. Only the FOMC meeting months are forecasted.

Table 2: Out-of-Sample Forecasts

Predictions between May 2009 – April 2010			
Models|Scores	Log	Quad	Eps
Log weights	-0.69	0.42	0.20
Optimal weights	-0.70	0.40	0.21

Notes: There are 8 meetings during the period spanning from May 2009 and April 2010. For further details refer to notes of Table 1.

There are two important questions that arise from this illustration. First, why are optimal weights selecting one model while others are neglected for 41 one-step-ahead forecast periods? This would suggest that for at least the first 41 periods,

one forecasting model could outperform forecast combination. Second, does this result hold in general? In other words, is it possible to derive a general result for what is observed in the above empirical illustration? The next sections attempt to shed some light on these questions.

3 Theoretical analysis of optimal weights

Hall and Mitchell (2007) proposes a set of weights for density forecast combination by maximizing the average logarithmic score of the combined density forecast. It shows that the optimal weights minimize the estimated Kullback-Leibler Information Criterion (KLIC) distance between the true density and the combined probability forecast.[6]

The notation for the rest of the paper follows Hall and Mitchell (2007). Suppose that there are N density forecasts, $g_{it}(\cdot)$, produced by models or analysts $i = 1, ..., N$ of a real-valued variable y_t at time t, where $t = 1, ..., T$ and T is the total number of the forecasted periods.[7] The combined density forecast is defined as the finite mixture

$$p_t(\cdot) = \sum_{i=1}^{N} \omega_i g_{it}(\cdot), \tag{4}$$

where ω_i are a set of non-negative weights that sum up to one. Further the densities are evaluated at y_t, the actual realization, and $g_{it} = g_{it}(y_t)$ is used for notational convenience.

Definition 1 of Hall and Mitchell (2007) gives the optimal weights vector $\boldsymbol{\omega}^* = (\omega_1^*, \ldots, \omega_N^*)$ as the solution of the optimization problem

$$\boldsymbol{\omega}^* = \arg \max_{(\omega_1 \ldots \omega_N)} \frac{1}{T} \sum_{t=1}^{T} \ln p_t(y_t), \tag{5}$$

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 $t = 1, \ldots, T$.

Suppose there are two competing density forecasts which can be evaluated after observing the actual realization, g_{1t} and g_{2t}. The optimization problem (5) can be written as a one dimension problem

$$\omega^* = \arg \max_{0 \leq \omega \leq 1} \frac{1}{T} \sum_{t=1}^{T} \ln \left(\omega g_{1t} + (1 - \omega) g_{2t} \right), \tag{6}$$

[6] A similar idea is used in Geweke and Amisano (2011).

[7] 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 $P_{j,t}(y_t = j)$.

where ω is the weight of the first model. The objective function can be re-written in simpler terms as

$$\frac{1}{T}\sum_{t=1}^{T}\ln\left(1+\omega\delta_t\right)+\frac{1}{T}\sum_{t=1}^{T}\ln g_{2t},$$

where $\delta_t = \frac{g_{1t}-g_{2t}}{g_{2t}}$ represents the relative forecasting performance of forecast 1 over forecast 2. The term $\frac{1}{T}\sum_{t=1}^{T}\ln g_{2t}$ in the objective function can be ignored as it does not depend on ω and does not affect the optimization. Hence, the optimization problem boils down to

$$\omega^* = \arg\max_{0\leq\omega\leq 1}\frac{1}{T}\sum_{t=1}^{T}\ln\left(1+\omega\delta_t\right) \tag{7}$$

In order to visualize and understand the theoretical properties of (7), the next subsections closely study cases of one forecasting period ($T=1$), two forecasting periods ($T=2$) or more ($T\geq 3$).

3.1 $T=1$

When there is only one period to evaluate the performance, the optimization problem simplifies to

$$\omega^* = \arg\max_{0\leq\omega\leq 1}\ln\left(1+\omega\delta_1\right)$$

and the solution is either $\omega^*=1$ if $\delta_1>0$ or $\omega^*=0$ if $\delta_1<0$. Hence, the solution puts all the weight on one of the models depending on which model outperforms the other.

3.2 $T=2$

The optimization problem becomes more complex with two forecasting periods

$$\omega^* = \arg\max_{0\leq\omega\leq 1}\frac{1}{2}\left(\ln\left(1+\omega\delta_1\right)+\ln\left(1+\omega\delta_2\right)\right).$$

In the simplest case, if forecast 1 is better in both periods, $\delta_1,\delta_2>0$, then $\omega^*=1$ and if forecast 2 is better in both periods, $\delta_1,\delta_2<0$, then $\omega^*=0$. However, competing models could perform well at different times. Hence, it is possible to get $\delta_1>0$ for $T=1$ and $\delta_2<0$ for $T=2$, or the reverse. In this case, it is not clear whether the weight is on one model or shared between both models. In order to understand the situation the quadratic form

$$f(\omega)=(1+\omega\delta_1)(1+\omega\delta_2)=1+(\delta_1+\delta_2)\omega+\delta_1\delta_2\omega^2$$

can be studied as the logarithmic function is monotonically increasing. When $\delta_1 > 0$, $\delta_2 < 0$, $f(\omega)$ has roots $-1/\delta_1 < 0$ and $-1/\delta_2 > 0$ of different sign. $f(\omega)$ is an inverted u-shape function with a maximum at

$$\widetilde{\omega} = -\frac{(\delta_1 + \delta_2)}{2\delta_1\delta_2}.$$

The location of $\widetilde{\omega}$ inside or outside of the $[0, 1]$-interval, yields three possible values of ω^*:

(a) $\omega^* = 0$, if $(\delta_1 + \delta_2) \leq 0$ (all the weight is on forecast 2),

(b) $0 < \omega^* < 1$, if $(\delta_1 + \delta_2) > 0$ and $-\frac{1}{2\delta_1\delta_2}(\delta_1 + \delta_2) < 1$ (both forecasts have non zero weights).

(c) $\omega^* = 1$, if $(\delta_1 + \delta_2) > 0$ and $-\frac{1}{2\delta_1\delta_2}(\delta_1 + \delta_2) \geq 1$ (all the weight is on forecast 1),

In case (a) forecast 2 outperforms forecast 1 in the second period more than forecast 1 outperforms forecast 2 in the first period, then all the weight goes to forecast 2. On the other hand, in case (c) if forecast 2 outperforms forecast 1 in the second period less than forecast 1 outperforms forecast 2 in the first period, then it is possible that forecast 1 retains all the weight. The interior solution $(0 < \omega^* < 1)$ is only found when $\tilde{\omega}$ is less than 1. The borderline scenario between cases (b) and (c) is given by the line $-\frac{1}{2\delta_1\delta_2}(\delta_1 + \delta_2) = 1$ (or $\delta_2 = -\frac{\delta_1}{2+1/\delta_1}$), which provides the limit set of points for which $\omega^* = 1$.

Figure 2 illustrates those three possibilities. It is clear that when $T = 2$, the chances of obtaining a corner solution (with ω^* equals to 1 or 0) are much greater than having an interior solution with positive weights for each forecast (with $0 < \omega^* < 1$).

The chances of obtaining a distribution of weights across models depend on how far each model's relative forecasting performance is from each other. Consider the following situation. Let $-a \leq \delta_1, \delta_2 \leq a$, i.e. the relative forecasting performance of model 1 over model 2 in each period is bounded by a, such that if $a = 0.1$ for example, the performance of model 1 and 2 are within 10% of each other. Hence, the percentage of solutions where $0 < \omega^* < 1$ is given by

$$\mu = \frac{1}{2a^2}\left(\frac{a^2}{2} - \int_0^a \frac{\delta_1}{1 + 2\delta_1} d\delta_1\right)$$

solving the integral yields

$$\int_0^a \frac{\delta_1}{1 + 2\delta_1} d\delta_1 = \frac{a}{2} - \frac{1}{4}\ln(1 + 2a)$$

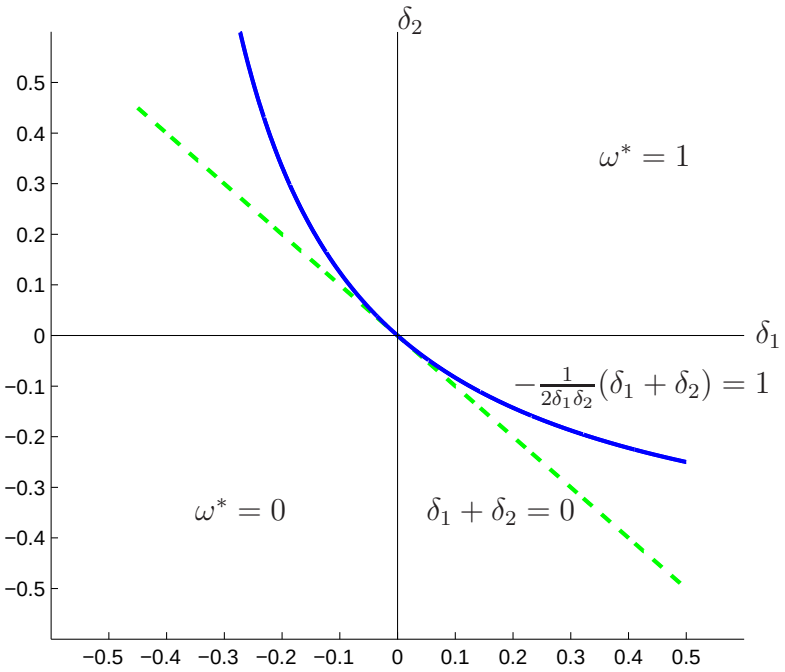


Figure 2: Areas of corner and interior solutions. Case (a): all the weight goes to forecast 2, $\omega^* = 0$, in the area below the green dashed line. Case (b): interior solution, $0 < \omega^* < 1$, in the area between the green dashed line and the blue solid curve. Case (c): all weight goes to forecast 1, $\omega^* = 1$, in the area above the blue solid curve given by $-\frac{1}{2\delta_1\delta_2}(\delta_1 + \delta_2) = 1$.

and

$$\mu = \frac{1}{2a^2}\left(\frac{a^2}{2} - \frac{a}{2} + \frac{1}{4}\ln(1+2a)\right).$$

The numerical results are given in Table 3. The proportion of interior solutions is rather small, but increasing when the bound a increases. In summary, with the

Table 3: Analytical proportion of mixing solutions as a function of a.

a	μ
0.1	2.9%
0.2	5.1%
0.3	6.9%

two forecast periods, the optimization problem (7) returns corner solutions most of the times. This situation is rather puzzling as one would expect more frequent

interior solutions when models are closely competing, as in Geweke and Amisano (2011).

3.3 $T \geq 3$

When there are 3 or more periods to forecast, the algebra of the analytical solution is less tractable. However, here are some theoretical considerations to foster some understanding of what is happening in the case $T \geq 3$.

Since the logarithm function is monotonically increasing, one can focus on the objective function of the optimization problem (7)

$$\psi(\omega) = \sum_{t=1}^{T} \ln(1 + \omega\delta_t) \tag{8}$$

or substitute $\psi(\omega)$ for the polynomial $f(\omega)$

$$f(\omega) = \prod_{t=1}^{T} (1 + \omega\delta_t). \tag{9}$$

In other words, on $[0, 1]$-interval and for reasonable[8] values of δ_t optimizing $\psi(\omega)$ is equivalent to optimizing $f(\omega)$.

The polynomial function $f(\omega)$ is somewhat easier to analyze. Without loss of generality assume that $\delta_1 < \delta_2 < \ldots < \delta_\tau < 0 < \delta_{\tau+1} < \ldots < \delta_{T+1} < \delta_T$. By plotting the function in Figure 3, it is clear that the largest positive δ_T and the smallest negative δ_1 have the most influence on ω^*, while other δ_t retain some influence on ω^*. Figure 4 magnifies Figure 3 over the $[0, 1]$-interval. With the assumption that $-\delta_1 < 1$, i.e. forecast 2 does not outperform forecast 1 more than 100% in relative terms, which is realistic for competing prediction models. As labeled in Figure 4 on [0,1]-interval $f(\omega)$ is decreasing and $\omega^* = 0$ in case (a), $f(\omega)$ is increasing and $\omega^* = 1$ in case (c) and $f(\omega)$ is well-behaved and has one maximum in case (b). The cases can be easily distinguished by the first derivatives $f'(0)$ and $f'(1)$. Below is a description of the meaning of three cases in terms of forecasting performance.

Analysis of $f'(\omega)$

The first order derivative of $f(\omega)$ with respect to ω is given by

$$f'(\omega) = \sum_{t=1}^{T} \delta_t \prod_{j \neq t} (1 + \omega\delta_j)$$

[8] $\delta_1 < 0$ and $-1/\delta_1 > 1$

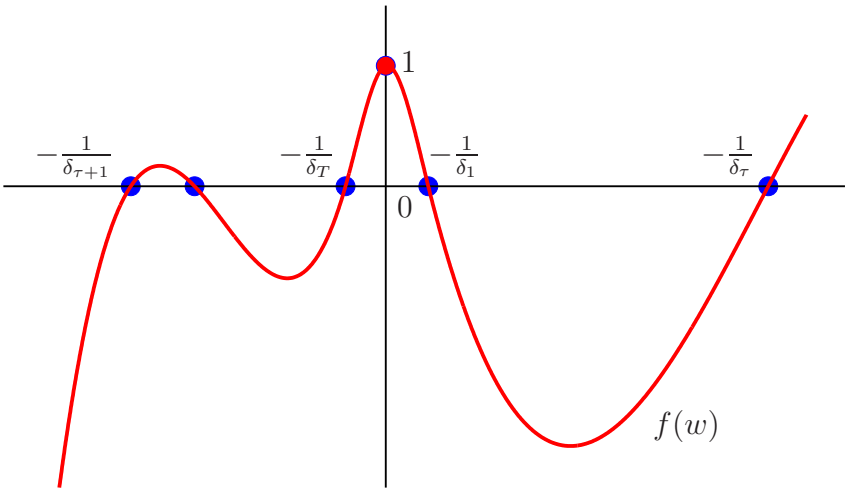


Figure 3: Stylized representation of the polynomial function $f(\omega)$ and its roots

and $f'(0) = \sum_{t=1}^{T} \delta_t$ and $f'(1) = \sum_{t=1}^{T} \delta_t \prod_{j \neq t}(1+\delta_j)$. Cases (a) - (c) presented in Figure 4 correspond to

(a) $f'(0) < 0$ and $f'(1) < 0$ with $\omega^* = 0$ if $\sum_{t=1}^{T} \delta_t < 0$: forecast 2 cumulatively outperforms forecast 1,

(b) $f'(0) > 0$ and $f'(1) < 0$ with $0 < \omega^* < 1$: this case does not have a tractable interpretation,

(c) $f'(0) > 0$ and $f'(1) > 0$ with $\omega^* = 1$,: this case does not have a tractable interpretation except for the special case when all $\delta_t > 0$ and forecast 1 is better every period.

Although substituting $\psi(\omega)$ for the polynomial $f(\omega)$ is helpful for the stylized representations in Figures 3 and 4, the three cases are generally complex to interpret. Perhaps the derivatives of the original objective function $\psi(\omega)$ can shed more light on the matter.

Analysis of $\psi'(\omega)$

The first order derivative of $\psi(\omega)$ with respect to ω is given by

$$\psi'(\omega) = \sum_{t=1}^{T} \frac{\delta_t}{1+\omega\delta_t}$$

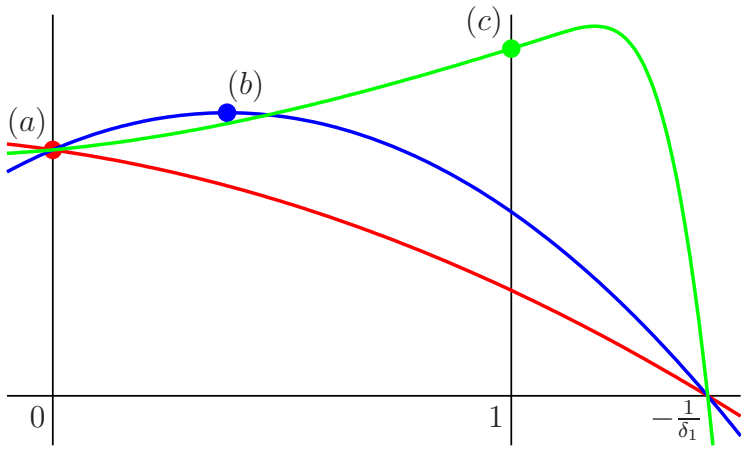


Figure 4: Stylized representation of the polynomial function $f(\omega)$ on $[0, 1]$-interval

and $\psi'(0) = \sum_{t=1}^{T} \delta_t$ and $\psi'(1) = \sum_{t=1}^{T} \frac{\delta_t}{1+\delta_t}$. $\psi'(0) = f'(0)$ and is still easy to interpret, but $\psi'(1)$ is somewhat simpler than $f'(1)$ and can be written as

$$\psi'(1) = \sum_{t=1}^{T} \frac{\delta_t}{1+\delta_t} = T - \sum_{t=1}^{\tau} \frac{1}{1+\delta_t} - \sum_{t=\tau+1}^{T} \frac{1}{1+\delta_t}.$$

Since $\frac{1}{1+\delta_t} > 1$ for $t \leq \tau$ and $\frac{1}{1+\delta_t} < 1$ for $t > \tau$, periods when forecast 2 is better are compared against periods when forecast 1 is better. The comparison is done cumulatively over all periods with an additional $1/(1+x)$ transformation due to the logarithm form. Cases (a) - (c) in Figure 4 correspond to

(a) $\psi'(0) < 0$ and $\psi'(1) < 0$ with $\omega^* = 0$, if $\sum_{t=1}^{T} \delta_t < 0$, i.e. forecast 2 cumulatively outperforms forecast 1,

(b) $\psi'(0) > 0$ and $\psi'(1) < 0$ with $0 < \omega^* < 1$, if $\sum_{t=1}^{T} \delta_t > 0$, i.e. forecast 1 cumulatively outperforms forecast 2, but $1 < \frac{1}{T}\sum_{t=1}^{\tau} \frac{1}{1+\delta_t} + \frac{1}{T}\sum_{t=\tau+1}^{T} \frac{1}{1+\delta_t}$, which implies that the average performance of forecast 1 does not outweigh the average performance of forecast 2,

(c) $\psi'(0) > 0$ and $\psi'(1) > 0$ with $\omega^* = 1$, if $1 > \frac{1}{T}\sum_{t=1}^{\tau} \frac{1}{1+\delta_t} + \frac{1}{T}\sum_{t=\tau+1}^{T} \frac{1}{1+\delta_t}$, i.e. the good average performance of forecast 1 outweighs the average good performance of forecast 2.

Although the analysis of $\psi'(\omega)$ better elucidates the three cases encountered, the following special case provides further insights into the matter.

Special case

Suppose that forecast 1 is only better than forecast 2 in the last period T, $\delta_T > 0$, while $\delta_t < 0$ for $t = 1, \ldots, T-1$. In this case only if $\delta_T > -\sum_{t=1}^{T-1} \delta_t$, forecast 1 is used in the combination and its weight is $\omega^* \neq 0$. Otherwise, if $\delta_T < -\sum_{t=1}^{T-1} \delta_t$, then $\omega^* = 0$ and forecast 1 is not used in the combination. Furthermore, in the border case, between case (a) and case (b) in Figure 4, the performance of forecast 1 in the last period balances out its previous poor performance history

$$\psi'(0) = \sum_{t=1}^{T-1} \delta_t + \delta_T = 0$$

and so $\delta_T = -\sum_{t=1}^{T-1} \delta_t$, which implies that $\omega^* = 0$. Moreover,

$$\psi'(1) = T - \sum_{t=1}^{T-1} \frac{1}{1+\delta_t} - \frac{1}{1+\delta_T}$$

and the other border case, between case (b) and case (c) in Figure 4, is given by $\psi'(1) = 0$, i.e. $\delta_T = \frac{1}{T - \sum_{t=1}^{T-1} \frac{1}{1+\delta_t}} - 1$. For simplicity assume that $\delta_t = \delta$ for $t = 1, \ldots, T-1$, then the interior solutions $0 < \omega^* < 1$ are in the region

$$(T-1)\delta < \delta_T < \frac{(T-1)\delta}{1-T\delta}.$$

However, this interval does not always exist and can be quite narrow.

In other words, in order to exit the corner solution $\omega^* = 0$, the performance of forecast 1 in the last period needs to be outstanding, with the risk of ending up in the other corner where $\omega^* = 1$. For example, if $\delta < 1$ and $1 - T\delta < 0$, then

$$\psi'(1) = T - \frac{(T-1)}{1-\delta} - \frac{1}{1+\delta_T} = \frac{1-T\delta}{1-\delta} - \frac{1}{1+\delta_T} < 0$$

and $\omega^* < 1$. However, if T is large enough and δ is fixed, then the interior solution $0 < \omega^* < 1$ is only possible if $-(T-1)\delta < \delta_T$. In practical terms, for $\delta = -10\%$ and $T = 11$, δ_T must be greater than 100% for forecast 1 to outperform its recent history, which is unrealistic from two closely competing models.

In summary, the theoretical analysis points to the dominance of corner solutions over interior ones when there are closely competing models and few forecasting periods. In order to see clearly what happens to the optimal weights when $T \geq 3$ and to consider more than two competing models, one needs to recourse to Monte Carlo simulations covered in the next section.

4 Simulations

4.1 Competing models

Two competing models

The experiment is set up on the premises of the theoretical considerations discussed in the previous section. The optimization problem is given in equation (7). There are two closely competing models, g_{1t}, g_{2t}, in terms of density forecasting performance. The relative model performance in each period, δ_t, is generated as follows

$$\delta_t = \rho\,\delta_{t-1} + \varepsilon_t \quad \text{with} \quad \varepsilon_t \sim \mathrm{N}\left(0, \left(\frac{0.3}{2\sqrt{(1-\rho^2)}}\right)^2\right)$$

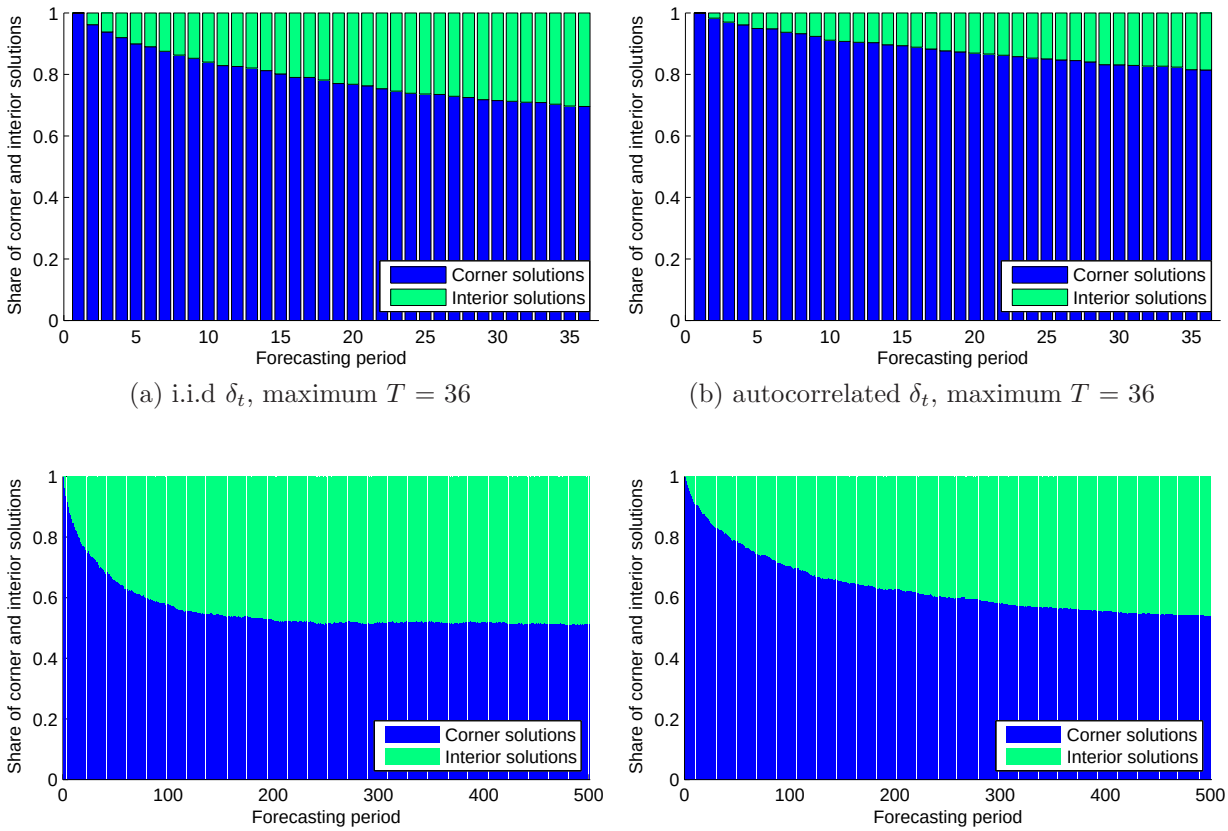
with ρ as the correlation coefficient. Hence, the correlation between the forecasting performance of the two models is such that if $\delta_{t-1} > 0$, the chances of it to remain positive in period t increases with ρ. The variance is chosen such that $|\delta_t| \leq 0.3$ with high probability, i.e. the models are competing and always produce similar types of density forecast. In simulations ρ is set to 0 when there is no correlation in the forecasting performance of each model (*i.i.d.* δ_t), and to 0.5 when there is correlation across time (autocorrelated δ_T).

The advantage in simulating δ_t directly rather than generating separate data process for the two models is that it permits to focus solely on the relative performance of the models. The forecasting sample is generated at once for $t = 1, \ldots, T$ and this process is repeated 10,000 times.[9]

Two sets of figures are reported. In Figures 5a and 5b, the maximum number of forecasting periods is set to 36, which would be equivalent to 3 years of monthly data, and in Figures 5c and 5d it is 500 periods in order to assess how the optimal weights are behaving when T goes to infinity. All following figures including Figures 5a - 5d can be interpreted in the following way. The vertical axis measures the relative frequency of corner solutions, when one model is given 100% of the weight while the other receives 0%, against interior solutions, when the share of the weight for one model is less than 100% and the other model receive non-zero weight. The horizontal axis depicts the number of forecasting periods, T.

One can observe from Figure 5a that interior solutions accounts for approximately 30% of the time even after 36 time periods. In fact, it takes more than 200 time periods to find an equal chance of corner and interior solutions (Figure 5c). As expected, this trend is slower when there is correlation ($\rho = 0.5$) as seen from Figures 5b and 5d. As shown in Figure 5d, it takes almost the full 500 periods to

[9]Experiments for which the forecasting sample is generated independently one period at a time yield the same results as those presented here. The results are available upon request.



(a) i.i.d δ_t, maximum $T = 36$

(b) autocorrelated δ_t, maximum $T = 36$

(c) i.i.d δ_t, maximum $T = 500$

(d) autocorrelated δ_t, maximum $T = 500$

Figure 5: Simulation results for two competing models

have the same results as in the *i.i.d* case.

These findings underline the importance of having enough observations to evaluate forecast performance and to find the optimal weights in (5). Remember that Hall and Mitchell (2007) relies on asymptotic results and assumes that $T \to \infty$ to choose weights that minimize the Kullback-Leibler distance. Geweke and Amisano (2011), for example, run an empirical exercise where the time series contain thousands of observations.

Ten competing models

When there are more than two competing models, the optimization problem can be written as

$$\boldsymbol{\omega}^* = \arg\max_{\boldsymbol{\omega}} \frac{1}{T} \sum_{t=1}^{T} \ln \left(\sum_{i=1}^{N} \omega_i g_{it} \right), \tag{10}$$

where N is the number of models and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)'$ is a vector of weights assigned to each model.

In the simulations for ten models, it is assumed that the models produce similar forecasts. This is captured by simulating $g_{it} = \phi(\xi_{it})$, where ϕ is the standard normal density and ξ_{it} are *i.i.d.* (across both i and t) random variables simulated from the normal distribution with the mean 0 and the standard deviation 1/2.

The results are presented in Figure 6. For short term horizons, the optimal

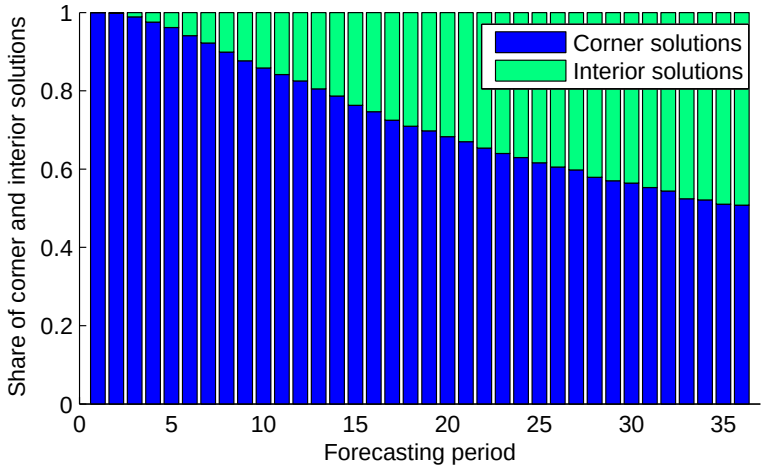


Figure 6: The share of corner and interior solutions for ten competing models across different forecasting periods.

solution for (10) is a corner solution. Only after 36 forecasting periods (3 years of monthly data) there is an approximately equal share of corner and interior solutions.

4.2 Forecast combination for different data generating processes

In order to know the theoretical value of the optimal weight, the data generating process (DGP) and models need to be specified. This is explored in the next three simulation cases for different types of DGPs.

Alternating DGP

For simplicity, consider an AR(1) model

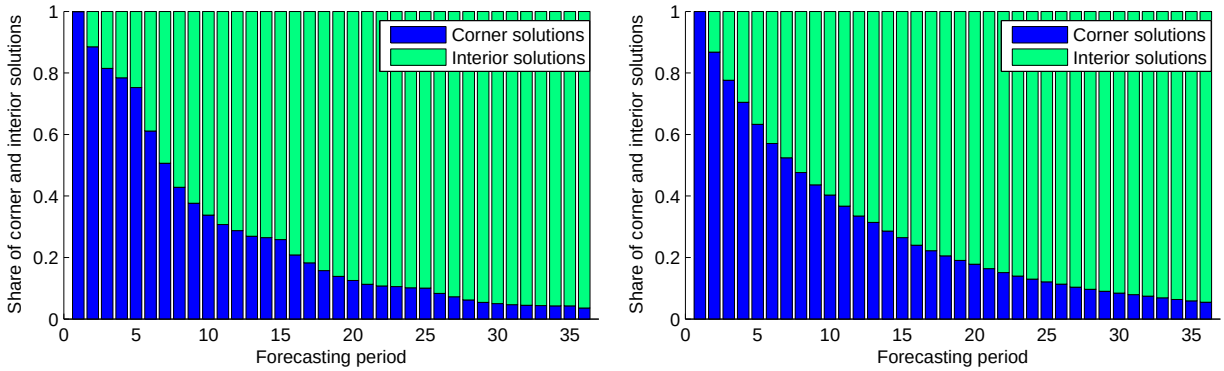
$$y_t^{(1)} = \rho y_{t-1}^{(1)} + \nu_t, \quad \nu_t \sim i.i.d.\ \mathrm{N}(0,1) \tag{11}$$

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

$$y_t^{(2)} = \varepsilon_t, \quad \varepsilon_t = \theta \varepsilon_{t-1} + \nu_t, \quad \nu_t \sim i.i.d.\ \mathrm{N}(0,1) \tag{12}$$

with $\theta = 0.7$, assuming that the parameters are known so there is no estimation noise. The true DGP to be forecasted by (11) and (12), combines both models (11) and (12) by switching from one model to the other every 5 periods.

In this situation, it is theoretically optimal to combine both models as none of the models captures the true DGP on their own. This is visible in the simulation results in Figure 7a where, after 36 periods, almost all solutions are interior. When $T = 36$ the average optimal weight is 0.56 which reflects that each of the models captures the DGP roughly half of the time on their own. However, the convergence to the optimal solution is slow and after 24 forecasting periods (2 years of monthly data) the optimization problem (6) yields corner solutions in 10% of the simulations.



(a) The DGP switches every 5 periods.

(b) The DGP switches with probability 0.3.

Figure 7: The share of corner and interior solutions for two misspecified models across different forecasting periods. The DGP switches between the two models.

Markov switching DGP

In this simulation, the switch in DGP between model (11) and (12) is uncertain. The probability that the prevailing model determining the DGP remains the same in the next period is 0.7, while the probability that there is a switch to the alternative model is 0.3.

This Markov switching DGP has a stationary state where the system follows model 1 half of the time and model 2 the other half. The situation is similar to the predetermined switch between the models. As before, it is theoretically optimal to combine two models. This is corroborated in Figure 7b where after 36 periods most of the solutions are interior. When $T = 36$ the average optimal weight is 0.499 which indicates that each model captures the true DGP on their own only half of the time. The convergence to the optimal solution is also slow and after

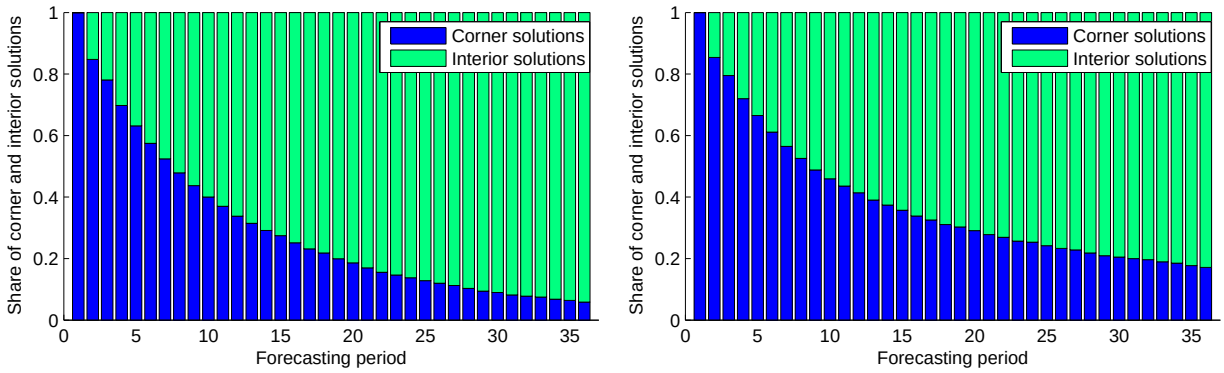
24 forecasting periods (2 years of monthly data) the corner solutions occur about 20% of the times.

Mixing DGP

Finally, if the true DGP is a mix of AR(1) model (11) and MA(1) model (12) in every period, and hence the actual observation are generated as an ARMA(1,1) model

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

then the theoretically optimal weight from solving (6) should converge to α. This is indeed what is observed in Figure 8a where $\alpha = 0.5$ and the average weight after 36 forecasting periods is 0.503 and in Figure 8b 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 roughly 20% of corner solutions after 36 forecasting periods.



(a) $\alpha = 0.5$ (b) $\alpha = 0.3$

Figure 8: The share of corner and interior solutions for two misspecified models across different forecasting periods. The DGP is mixing with parameter α which is equal to the theoretically optimal weight.

5 Concluding comments

The idea of having a training sample for parameter estimation before forecasting out-of-sample is widely acknowledged in the forecasting literature. All the theoretical, simulation and empirical results considered in this paper point to the necessity of an additional training sample for the optimal weights 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 in (5) when the number of the forecasting periods, T, is small. When T is large enough, the asymptotic theory used for justifying the optimal weights in Hall and Mitchell (2007) and in Geweke and Amisano (2011) is valid and the optimal weights have expected properties. If one wishes the optimal weights to behave as expected in theory, the authors' practical recommendation is the use of at least 36 data points (3 years of monthly data) when solving the optimization problem. Alternatively one can use log or quadratic weights proposed in Pauwels and Vasnev (2012) that do not need that extensive training period.

Acknowledgements

The authors are grateful to Rachida Ouysse, Valentyn Panchenko, Tommaso Proietti and the participants of the SERG meeting in November 2012 at UNSW. The authors thank Jaya Krishnakumar, Subhabrata Das and Subramanian Chetan, as well as 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

Boero, G., J. Smith, and K. F. Wallis (2011). Scoring rules and survey density forecasts. *International Journal of Forecasting 27*, 379–393.

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

Corradi, V. and N. R. Swanson (2006). Predictive density evaluation. In C. J. W. G. G. Elliot and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Chapter 5, pp. 198–284. North-Holland.

Diks, C., V. Panchenko, and D. van Dijk (2010, September). Out-of-sample comparison of copula specifications in multivariate density forecasts. *Journal of Economic Dynamics and Control 34*(9), 1596–1609.

Diks, C., V. Panchenko, and D. van Dijk (2011, August). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics 163*(2), 215–230.

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

Garratt, A., J. Mitchell, S. P. Vahey, and E. C. Wakerly (2011, January). Real-time inflation forecast densities from ensemble phillips curves. *The North American Journal of Economics and Finance 22*(1), 77–87.

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

Geweke, J. and G. Amisano (2011, September). Optimal prediction pools. *Journal of Econometrics 164*(1), 130–141.

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

Hamilton, J. D. and O. Jorda (2002). A model of the federal funds rate target. *Journal of Political Economy 110*(5), 1135–1167.

Hu, L. and P. C. B. Phillips (2004a). Dynamics of the federal funds target rate: A nonstationary discrete choice approach. *Journal of Applied Econometrics 19*(7), 851–867.

Hu, L. and P. C. B. Phillips (2004b). Nonstationary discrete choice. *Journal of Econometrics 120*(1), 103–138.

Jore, A. S., J. Mitchell, and S. P. Vahey (2010). Combining forecast densities from vars with uncertain instabilities. *Journal of Applied Econometrics 25*(4), 621–634.

Kascha, C. and F. Ravazzolo (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*(1), 47–67.

Kim, H., J. Jackson, and R. Saba (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*(2-3), 519–534.

Ng, J., C. S. Forbes, G. M. Martin, and B. P. McCabe (2010, October). Non-parametric estimation of forecast distributions in non-gaussian state space models. Monash University, Mimeo.

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

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

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*, pp. 135–196. Elsevier, Horth-Holland.

Wolden Bache, I., A. Sofie Jore, J. Mitchell, and S. P. Vahey (2011, October). Combining var and dsge forecast densities. *Journal of Economic Dynamics and Control 35*(10), 1659–1670.