Center for Quantitative Economic Research
WORKING PAPER SERIES

Optimal Prediction Pools

John Geweke and Gianni Amisano

CQER Working Paper 09-05
October 2009

FEDERAL RESERVE BANK *of* ATLANTA

Center for Quantitative Economic Research | WORKING PAPER SERIES

Optimal Prediction Pools

John Geweke and Gianni Amisano

CQER Working Paper 09-05
October 2009

Abstract: A prediction model is any statement of a probability distribution for an outcome not yet observed. This study considers the properties of weighted linear combinations of n prediction models, or linear pools, evaluated using the conventional log predictive scoring rule. The log score is a concave function of the weights, and, in general, an optimal linear combination will include several models with positive weights despite the fact that exactly one model has limiting posterior probability one. The paper derives several interesting formal results: For example, a prediction model with positive weight in a pool may have zero weight if some other models are deleted from that pool. The results are illustrated using S&P 500 returns with prediction models from the ARCH, stochastic volatility, and Markov mixture families. In this example models that are clearly inferior by the usual scoring criteria have positive weights in optimal linear pools, and these pools substantially outperform their best components.

JEL classification: C11, C53

Key words: forecasting, GARCH, log scoring, Markov mixture, model combination, S&P 500 returns, stochastic volatility

This paper was originally prepared for the Forecasting in Rio Conference, Graduate School of Economics, Getulio Vargas Foundation, Rio de Janeiro, July 2008. The authors acknowledge helpful comments from James Chapman, Frank Diebold, Joel Horowitz, James Mitchell, Luke Tierney, Mattias Villani, Kenneth Wallis, Robert Winkler, Arnold Zellner, and participants in presentations at the Bank of Canada, Erasmus University, the NBER-NSF Time Series Conference, Princeton University, Rimini Centre for Economic Analysis, Sveriges Riksbank, Rice University, and the University of Queensland. The authors gratefully acknowledge financial support from NSF grant SBR-0720547. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions regarding content to John Geweke, Departments of Statistics and Economics, W210 Pappajohn Business Bldg., University of Iowa, Iowa City, IA 52242-1000, john-geweke@uiowa.edu and Gianni Amisano, University of Brescia and European Central Bank, amisano@eco.unibs.it.

1 Introduction and motivation

The formal solutions of most decision problems in economics, in the private and public sectors as well as academic contexts, require probability distributions for magnitudes that are as yet unknown. Point forecasts are rarely sufficient. For econometric investigators whose work may be used by clients in different situations the mandate to produce predictive distributions is compelling. Increasing awareness of this context, combined with advances in modeling and computing, is leading to a sustained emphasis on these distributions in econometric research (Diebold et al. (1998); Christoffersen (1998); Corradi and Swanson (2006a, 2006b); Gneiting et al. (2007)). In many situations there are several models with predictive distributions available, leading naturally to questions of model choice or combination. While there is a large econometric literature on choice or combination of point forecasts, dating at least to Bates and Granger (1969) and extending through many more contributions reviewed recently by Timmermann (2006), the treatment of predictive density combination in the econometrics literature is much more limited. Granger et al. (1989) and Clements (2006) attacked the related problems of event and quantile forecast combination, respectively. Wallis (2005) was perhaps the first econometrician to take up combinations of predictive densities explicitly. Mitchell and Hall (2007) is the closest precursor of the approach taken here.

We consider the situation in which alternative models provide predictive distributions for a vector time series $\mathbf{y}_t$ given its history $\mathbf{Y}_{t-1} = \{\mathbf{y}_h, \ldots, \mathbf{y}_{t-1}\}$; h is a starting date for the time series, $h \leq 1$. A prediction model A (for "assumptions") is a construction that produces a probability density for $\mathbf{y}_t$ with respect to an appropriate measure ν from the history $\mathbf{Y}_{t-1}$ denoted $p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A)$. There are many kinds of prediction models. Leading examples begin with parametric conditional densities $p(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_A, A)$. Then, in a formal Bayesian approach

$$p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A) = p(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, A) = \int p(\mathbf{y}_t | \mathbf{Y}_{t-1}, \boldsymbol{\theta}_A, A) p(\boldsymbol{\theta}_A \mid \mathbf{Y}_{t-1}, A) d\boldsymbol{\theta}_A, \quad (1)$$

where $p(\boldsymbol{\theta}_A \mid \mathbf{Y}_{t-1}, A)$ is the posterior density

$$p(\boldsymbol{\theta}_A \mid \mathbf{Y}_{t-1}, A) \propto p(\boldsymbol{\theta}_A \mid A) \prod_{s=1}^{t-1} p(\mathbf{y}_s \mid \mathbf{Y}_{s-1}, \boldsymbol{\theta}_A, A)$$

and $p(\boldsymbol{\theta}_A \mid A)$ is the prior density for $\boldsymbol{\theta}_A$. A non-Bayesian approach might construct the parameter estimates $\widehat{\boldsymbol{\theta}}_A^{t-1} = f_{t-1}(\mathbf{Y}_{t-1})$ and then

$$p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A) = p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \widehat{\boldsymbol{\theta}}_A^{t-1}, A\right). \quad (2)$$

The specific construction of $p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A)$ does not concern us: in the extreme, it could be entirely judgmental. What is critical is that it rely only on information available at time $t-1$ and that it provide a mathematically complete predictive

density for $\mathbf{y}_t$. The primitives are these predictive densities and the realizations of the time series $\mathbf{y}_t$, which we denote $\mathbf{y}_t^o$ (o for "observed") in situations where the distinction between the random vector and its realization is important. This set of primitives is the one typically used in the few studies that have addressed these questions (e.g. Diebold et al. (1998, p. 879)). As Gneiting et al. (2007, p. 244) notes, the assessment of a predictive distribution on the basis of $p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A)$ and $\mathbf{y}_t^o$ only is consistent with the prequential principle of Dawid (1984).

1.1 Log scoring

Our assessment of models and combinations of models will rely on the log predictive score function. For a sample $\mathbf{Y}_T = \mathbf{Y}_T^o$ the log predictive score function of a single prediction model A is

$$LS(\mathbf{Y}_T^o, A) = \sum_{t=1}^{T} \log p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A\right). \tag{3}$$

In a full Bayesian approach $p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A) = p(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, A)$ and (3) becomes

$$LS(\mathbf{Y}_T^o, A) = \sum_{t=1}^{T} \log p\left(\mathbf{y}_t^o \mid \mathbf{Y}_{t-1}^o, A\right) = \log p(\mathbf{Y}_T^o \mid A) = \log \int p(\mathbf{Y}_T^o, \boldsymbol{\theta}_A \mid A)\, d\boldsymbol{\theta}_A$$

(Geweke (2001) and (2005, Section 2.6.2)). In a parametric non-Bayesian approach (2) the log predictive score is

$$LS(\mathbf{Y}_T^o, A) = \sum_{t=1}^{T} \log p\left(\mathbf{y}_t^o \mid \mathbf{Y}_{t-1}^o, \widehat{\boldsymbol{\theta}}_A^{t-1}, A\right)$$

which is smaller than the full-sample log-likelihood function evaluated at the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_A^T$.

For some of the analytical results in this study we assume that there is a data generating process D that gives rise to the ergodic vector time series $\{\mathbf{y}_t\}$. That is, there is a true model D, but it is not necessarily one of the models under consideration. For most D and A

$$E_D[LS(\mathbf{Y}_T, A)] = \int \left[\sum_{t=1}^{T} \log p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A)\right] p(\mathbf{Y}_T|D)\, d\nu(\mathbf{Y}_T)$$

exists and is finite. Given the ergodicity of $\{\mathbf{y}_t\}$,

$$T^{-1} LS(\mathbf{Y}_T, A) \xrightarrow{a.s.} \lim_{T\to\infty} T^{-1} E_D[LS(\mathbf{Y}_T, A)] = LS(A; D). \tag{4}$$

Whenever we invoke a true model D, we shall assume that (4) is true for D and any model A under consideration.

The log predictive score function is a measure of the out-of-sample prediction track record of the model. Other such scoring rules are, of course, possible, mean square prediction error being perhaps the most familiar. One could imagine using a scoring rule to evaluate the predictive densities provided by a modeler. Suppose that the modeler then produced predictive densities in such a way as to maximize the expected value of the scoring rule, the expectations being taken with respect to the modeler's subjective probability distribution. The scoring rule is said to be proper if, in such a situation, the modeler is led to report a predictive density that is coherent and consistent with his subjective probabilities. (The term "proper" was coined by Winkler and Murphy (1968), but the general idea dates back at least to Brier (1950) and Good (1952).) If the scoring rule depends on $\mathbf{Y}_T^o$ and $p\left(\mathbf{y}_t; \mathbf{Y}_{t-1}, A\right)$ only through $p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A\right)$ then it is said to be local (Bernardo (1979)).

The only proper local scoring rule takes the form

$$g\left(\mathbf{Y}_t^o\right) + c \sum_{t=1}^{T} \log p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A\right)$$

with $c > 0$, a linear transformation of (3). This was shown by di Finetti and Savage (1963) and Shuford et al. (1966) for the case in which the support of $\{\mathbf{y}_t\}$ is a finite set of at least three discrete points; for further discussion see Winkler (1969, p. 1075). It was shown for the case of continuously distributed $\{\mathbf{y}_t\}$ by Bernardo (1979); for further discussion see Gneiting and Raftery (2007, p. 366).

This study will consider alternative prediction models $A_1, \ldots, A_n$. Propriety of the scoring rule is important in this context because it guarantees that if one of these models were to coincide with the true data generating process D, then that model would attain the maximum score as $T \to \infty$.

There is a long-standing literature on scoring rules for discrete outcomes and in particular for Bernoulli random variables (DeGroot and Fienberg (1982), Clemen et al. (1995)). However, as noted in the recent review article by Gneiting et al. (2007, p. 364) and Bremmes (2004) the literature on scoring rules for probabilistic forecasts of continuous variables is sparse.

1.2 Linear pooling

This study explores using the log scoring rule (3) to evaluate combinations of probability densities $p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}^o, A_j\right)$ $(j = 1, \ldots, n)$. There are, of course, many ways in which these densities could be combined, or aggregated; see Genest et al. (1984) for a review and axiomatic approach. McConway (1981) showed that, under mild regularity conditions, if the process of combination is to commute with any possible marginalization of the distributions involved, then the combination must be linear. Moreover, such combinations are trivial to compute, both absolutely and in compar-

ison with alternatives. Thus we study predictive densities of the form

$$\sum_{i=1}^{n} w_i p\left(\mathbf{y}_t; \mathbf{Y}_{t-1}^o, A_i\right); \ \sum_{i=1}^{n} w_i = 1; \ \ w_i \geq 0 \quad (i = 1, \ldots, n). \tag{5}$$

The restrictions on the weights w_i are necessary and sufficient to assure that (5) is a density function for all values of the weights and all arguments of any density functions. We evaluate these densities using the log predictive score function

$$\sum_{t=1}^{T} \log \left[\sum_{i=1}^{n} w_i p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_i\right) \right]. \tag{6}$$

Combinations of subjective probability distributions are known as opinion pools, a term due to Stone (1961), and linear combinations are known as linear opinion pools (Bacharach (1974)). We use the term prediction pools to describe the setting specific to this study. While all models are based on opinions, only formal statistical models are capable of producing the complete predictive densities that, together with the data, constitute our primitives. Choice of weights in any combinations like (5) is widely regarded as a difficult and important question. This study uses past performance of the pool to select the weights; in the language of Jacobs (1995) the past constitutes the training sample for the present. Sections 3 and 5 show that this is easy to do. This study compares linear prediction pools using the log scoring rule. An optimal prediction pool is one with weights chosen so as to maximize (6).

Hall and Mitchell (2007) proposed combining predictive probability densities by finding the nonnegative weights w_i that maximize (6). The motivation of that study is asymptotic: as $T \to \infty$, the weights so chosen are those that minimize the Kullback-Leibler directed distance from an assumed data generating process D to the model (5). Hall and Mitchell (2007) show that direct maximization of (6) is more reliable than some other methods, involving probability integral transforms, that have been proposed in the literature. The focus of our work is complementary and more analytical, and we also provide a larger-scale implementation of optimal pooling than does Hall and Mitchell (2007).

The characteristics of optimal prediction pools turn out to be strikingly different from those that are constructed by means of Bayesian model averaging (which is always possible in principle and often in practice) as well as those that result from conventional frequentist testing (which is often problematic since the models are typically non-nested). Given a data generating process D that produces ergodic $\{\mathbf{y}_t\}$ a limiting optimal prediction pool exists, and unless one of the models A_j coincides with D, several of the weights in this pool typically are positive. In contrast, the posterior probability of the model A_j with the smallest Kullback-Leibler directed distance from D will tend to one and all others to zero. Any frequentist procedure based on testing will have a similar property, but with a distance measure specific to the test.

The contrast is rooted in the fact that Bayesian model averaging and frequentist tests are predicated on the belief that $A_j = D$ for some j, whereas optimal prediction pools make no such assumption. If $A_j \neq D$ $(j = 1, \ldots, n)$ then it is possible that one model would dominate the optimal pool, but this result seems to us unusual and this supposition is supported in the examples studied here. Our findings show that optimal pools can, and do, perform substantially better than any of their constituent models as assessed by a log predictive scoring rule. We show that there must exist a model, not included in the pool, with a log predictive score at least as good as, and in general better than, that of the optimally scored prediction pool.

The paper develops the basic ideas for a pool of two models (Section 2) and then applies them to prediction model pools for daily S&P 500 returns, 1972 through 2005 (Section 3). It then turns to the general case of pools of n models and studies how changes in the composition of the pool change the optimal weights (Section 4). Section 5 constructs an optimal pool of six alternative prediction models for the S&P 500 returns. Section 6 studies the implications of optimal prediction pools for the existence of prediction models as yet undiscovered that will compare favorably with those in the pool as assessed by a log predictive scoring rule. The final section concludes.

2 Pools of two models

Consider the case of two competing prediction models $A_1 \neq A_2$. From (4)

$$T^{-1}\left[LS\left(\mathbf{Y}_T, A_1\right) - LS\left(\mathbf{Y}_T, A_2\right)\right] \overset{a.s.}{\to} LS\left(A_1; D\right) - LS\left(A_2; D\right).$$

If A_1 corresponds to the data generating process D, then in general $LS\left(A_1; D\right) - LS\left(A_2; D\right) = LS\left(D; D\right) - LS\left(A_2; D\right) \geq 0$ and the limiting value coincides with the Kullback-Leibler distance from D to A_2. If A_2 also nests A_1 then $LS\left(A_1; D\right) - LS\left(A_2; D\right) = 0$, but in most cases of interest $LS\left(A_1; D\right) \neq LS\left(A_2; D\right)$ and so if $A_1 = D$ then $LS\left(A_1; D\right) - LS\left(A_2; D\right) > 0$. These special cases are interesting and informative, but in application most econometricians would agree with the dictum of Box (1980) that all models are false. Indeed the more illuminating special case might be $LS\left(A_1; D\right) - LS\left(A_2; D\right) = 0$ when neither model A_j is nested in the other: then both A_1 and A_2 must be false.

In general $LS\left(A_1; D\right) - LS\left(A_2; D\right) \neq 0$. For most prediction models constructed from parametric models of the time series $\{\mathbf{y}_t\}$ a closely related implication is that one of the two models will almost surely be rejected in favor of the other as $T \to \infty$. For example in the Bayesian approach (1) the Bayes factor in favor of one model over the other will converge to zero, and in the non-Bayesian construction (2) the likelihood ratio test or another test appropriate to the estimates $\widehat{\boldsymbol{\theta}}^t_{A_j}$ will reject one model in favor of the other. We mention these well-known results here only to emphasize the contrast with those in the remainder of this section.

Given the two prediction models A_1 and A_2, the prediction pool $A = \{A_1, A_2\}$ consists of all prediction models

$$p\left(\mathbf{y}_t; \mathbf{Y}_{t-1}, A\right) = wp\left(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_1\right) + (1-w)\, p\left(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_2\right), \quad w \in [0,1]. \tag{7}$$

The predictive log score function corresponding to given $w \in [0,1]$ is

$$f_T(w) = \sum_{t=1}^{T} \log\left[wp\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_1\right) + (1-w)\, p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_2\right)\right]. \tag{8}$$

The optimal prediction pool corresponds to $w_T^* = \arg\max_w f_T(w)$ in (8).[1] The determination of such a pool was, of course, impossible for purposes of forming the elements $wp\left(\mathbf{y}_t; \mathbf{Y}_{t-1}^o, A_1\right) + (1-w)\, p\left(\mathbf{y}_t; \mathbf{Y}_{t-1}^o, A_2\right)$ $(t = 1, \ldots, T)$ because it is based on the entire sample. But it is just as clear that weights w could be determined recursively at each date t based on information through $t-1$. We shall see subsequently that the required computations are practical, and in the examples in the next section there is almost no difference between the optimal pool considered here and those created recursively when the two procedures are evaluated using a log scoring rule.

The first two derivatives of f_T are

$$\begin{aligned} f_T'(w) &= \sum_{t=1}^{T} \frac{p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_1\right) - p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_2\right)}{wp\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_1\right) + (1-w)\, p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_2\right)}, \\ f_T''(w) &= -\sum_{t=1}^{T} \left[\frac{p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_1\right) - p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_2\right)}{wp\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_1\right) + (1-w)\, p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_2\right)}\right]^2 < 0. \end{aligned} \tag{9}$$

For all $w \in [0,1]$, $T^{-1} f_T(w) \xrightarrow{a.s.} f(w)$. If

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} E_D\left[p\left(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_1\right) - p\left(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_2\right)\right] \neq 0 \tag{10}$$

then $f(w)$ is concave. The condition (10) does not necessarily hold, but it seems to us that the only realistic case in which it does not occurs when one of the models nests the other and the restrictions that create the nesting are correct for the pseudo-true parameter vector. We have in mind, here, prediction models A_1 and A_2 that are typically non-nested and, in fact, differ substantially in functional form for their predictive densities. Henceforth we shall assume that (10) is true. Given

[1] The setup in (8) is formally similar to the nesting proposed by Quandt (1974) in order to test the null hypothesis $A_1 = D$ against the alternative $A_2 = D$. (See also Gourieroux and Monfort (1989, Section 22.2.7).) That is not the objective here. Moreover, Quant's test involves simultaneously maximizing the function in the parameters of both models and w, and is therefore equivalent to the attempt to estimated by maximum likelihood the mixture models discussed in Section 6; Quandt (1974) clearly recognizes the pitfalls associated with this procedure.

this assumption $w_T^* = \arg\max_w f_T(w)$ converges almost surely to the unique value $w^* = \arg\max_w f(w)$. Thus for a given data generating process D there is a unique, limiting optimal prediction pool. As shown in Hall and Mitchell (2007) this prediction pool minimizes the Kullback-Leibler directed distance from D to the prediction model (5).

It will prove useful to distinguish between several kinds of prediction pools, based on the properties of f_T. If $w_T^* \in (0,1)$ then A_1 and A_2 are each *competitive* in the pool $\{A_1, A_2\}$. If $w_T^* = 1$ then A_1 is *dominant* in the pool $\{A_1, A_2\}$ and A_2 is *excluded* in that pool;[2] equivalently $f_T'(1) \geq 0$, which amounts to

$$T^{-1}\sum_{t=1}^{T} p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_2\right) / p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_1\right) \leq 1.$$

By mild extension A_1 and A_2 are each competitive in the population pool $\{A_1, A_2\}$ if $w^* \in (0,1)$, and if $w^* = 1$ then A_1 is dominant in the population pool and A_2 is excluded in that pool.

Some special cases are interesting, not because they are likely to occur, but because they help to illuminate the relationship of prediction pools to concepts familiar from model comparison. First consider the hypothetical case $A_1 = D$.

Proposition 1 *If $A_1 = D$ then A_1 is dominant in the population pool $\{A_1, A_2\}$ and $f'(1) = 0$.*

Proof. If $A_1 = D$,

$$f'(1) = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E_D\left[1 - \frac{p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_2)}{p(\mathbf{y}_t; \mathbf{Y}_{t-1}, D)}\right] = 0.$$

From (9) and the strict concavity of f it follows that A_1 is dominant in the population pool. ■

A second illuminating hypothetical case is $LS(A_1; D) = LS(A_2; D)$. Given (10) then $A_1 \neq D$ and $A_2 \neq D$ in view of Proposition 1. The implication of this result for practical work is that if two non-nested models have roughly the same log score then neither is "true." Section 6 returns to this implication at greater length.

Turning to the more realistic case $LS(A_1; D) \neq LS(A_2; D)$, $w^* \in (0,1)$ implies also that $A_1 \neq D$ and $A_2 \neq D$. In fact one never observes f, of course, but the familiar log scale of $f_T(w)$ provides some indication of the strength of the evidence that neither $A_1 = D$ nor $A_2 = D$. There is a literature on testing that formalizes this idea in the context of (7); see Gourieroux and Monfort (1989, Chapter 22), and Quandt (1974). Our motivation is not to demonstrate that any prediction model is

[2]Dominance is a necessary condition for forecast encompassing (Chong and Hendry (1986)) asymptotically. But it is clearly weaker than forecast encompassing.

false; we know at the outset that this is the case. What is more important is that (7) evaluated at w_T^* provides a lower bound on the improvement in the log score predictive density that could be attained by models not in the pool, including models not yet discovered. We return to this point in Section 6.

If $w^* \in (0,1)$ then for a sufficiently large sample size the optimal pool will have a log predictive score superior to that of either A_1 or A_2 alone, and as sample size increases $w_T^* \xrightarrow{a.s.} w^*$. This is in marked contrast to conventional Bayesian model combination or non-Bayesian tests. Both will exclude one model or the other asymptotically, although the procedures are formally distinct. For Bayesian model combination the contrast is due to the fact that the conventional setup conditions on one of either $D = A_1$ or $D = A_2$ being true. As we have seen, in this case the posterior probability of A_1 and w_T^* have the same limit. By formally admitting the contingency that $A_1 \neq D$ and $A_2 \neq D$ we change the conventional assumptions, leading to an entirely different result: even models that are arbitrarily inferior, as measured by Bayes factors, can substantially improve predictions from the superior model as indicated by a log scoring rule. For non-Bayesian testing the explanation is the same: since a true test rejects one model and accepts the other, it also conditions on one of either $D = A_1$ or $D = A_2$ being true. We turn next to some examples.

3 Examples of two-model pools

We illustrate some properties of two-model pools using daily percent log returns of the Standard and Poors (S&P) 500 index and six alternative models for these returns. All of the models used rolling samples of 1250 trading days, about five years. The first sample consisted of returns from January 3, 1972 ($h = -1249$, in the notation of the previous section) through December 14, 1976 ($t = 0$), and the first predictive density evaluation was for the return on December 15, 1976 ($t = 1$). The last predictive density evaluation was for the return on December 16, 2005 ($T = 7324$).

Three of the models are estimated by maximum likelihood and predictive densities are formed by substituting the estimates for the unknown parameters: a Gaussian i.i.d. model ("Gaussian," hereafter); a Gaussian generalized autoregressive conditional heteroscedasticity model with parameters $p = q = 1$, or GARCH (1,1) ("GARCH"); and a Gaussian exponential GARCH model with $p = q = 1$ ("EGARCH"). Three of the models formed full Bayesian predictive densities using MCMC algorithms: a GARCH(1,1) model with i.i.d. Student t shocks ("t-GARCH"); the stochastic volatility model of Jacquier et al. (1994) ("SV"); and the hierarchical Markov normal mixture model with serial correlation and $m_1 = m_2 = 5$ latent states described in Geweke and Amisano (2007) ("HMNM").

Table 1 provides the log predictive score for each model. That for t-GARCH exceeds that of the nearest competitor, HMNM, by 19. Results for each are based on full Bayesian inference but the log predictive scores are not the same as log marginal likelihoods because the early part of the data set is omitted and rolling rather than

full samples are used. Nevertheless the difference between these two models strongly suggests that a formal Bayesian model comparison would yield overwhelming posterior odds in favor of t-GARCH. Of course the evidence against the other models in favor of t-GARCH is even stronger: 143 against SV, 232 against EGARCH, 257 against GARCH, and 1253 against Gaussian.

Pools of two models, one of which is t-GARCH, reveal that t-GARCH is not dominant in all of these pools. Figure 1 shows the function $f_T(w)$ for pools of two models, one of which is t-GARCH with w denoting the weight on the t-GARCH predictive density. The vertical scale is the same in each panel. All functions $f_T(w)$ are, of course, concave. In the GARCH and t-GARCH pool $f_T(w)$ has an internal maximum at $w = 0.944$ with $f_T(0.944) = -9317.12$, whereas $f_T(1) = -9315.50$. This distinction is too subtle to be evident in the upper left panel in which it appears that $f'_T(w) \approxeq 0$. For the EGARCH and t-GARCH pool, and for the HMNM and t-GARCH pool, the maximum is clearly internal. For the SV and t-GARCH pool $f_T(w)$ is monotone increasing, with $f'_T(1) = 1.96$. In the Gaussian and t-GARCH pool, not shown in Figure 1 , t-GARCH is again dominant with $f'_T(1) = 54.4$. Thus while all two-model comparisons strongly favor t-GARCH, it is dominant only in the pool with Gaussian and the pool with SV.

Figure 2 portrays $f_T(w)$ for two-model pools consisting of HMNM and one other predictive density, with w denoting the weight on HMNM. The scale of the vertical axis is the same as in Figure 1 in all panels except the upper left, which shows $f_T(w)$ in the two-model pool consisting of Gaussian and HMNM. The latter model nests the former, and it is dominant in this pool with $f'_T(1) = 108.3$. In pools consisting of HMNM on the one hand and GARCH, EGARCH or SV, on the other, the models are mutually competitive. Thus SV is excluded in a two-model pool with t-GARCH, but not in a two-model pool with HMNM. This is not a logical consequence of the fact that t-GARCH has a higher log predictive score than HMNM. Indeed, the optimal two-model pool for EGARCH and HMNM has a higher log predictive score than any two-model pool that includes t-GARCH, as is evident by comparing the lower left panel of Figure 2 with all the panels of Figure 1.

Table 2 summarizes some key characteristics of all the two-model pools that can be created for these predictive densities. The entries above the main diagonal indicate the log score of the optimal linear pool of the two prediction models. The entries below the main diagonal indicate the weight w_T^* on the model in the row entry in the optimal pool. In each cell there is a pair of entries. The upper entry reflects pool optimization exactly as described in the previous section. In particular, the optimal prediction model weight is determined just once, on the basis of the predictive densities for all T data points. This scheme could not be used in practice because only past data are available for optimization. The lower entry in each pair reflects pool optimization using the predictive densities $p\left(y_s^o; \mathbf{Y}_{s-1}^o, A_j\right)$ $(s = 1, \ldots, t-1)$ to form the optimal pooled predictive density for y_t. The log scores (above the main diagonal in Table 1) are the sums of the log scores for pools formed in this way. The

weights (below the main diagonal in Table 1) are averages of the weights w_t^* taken across all T predictive densities. (For $t = 1$, w_1^* was arbitrarily set at 0.5.)

For example, in the t-GARCH and HMNM pool, the log score using the optimal weight based on all T observations is -9284.72. If, instead, the optimal weight is recalculated in each period using only past predictive likelihoods, then the log score is -9287.28. The weight on the HMNM model is 0.289 in the former case, and the average weight on this model is 0.307 in the latter case. Note that in every case the log score is lower when it is determined using only past predictive likelihoods, than when it is determined using the entire sample. But the values are, at most, about 3 points lower. The weights themselves show some marked differences – pools involving EGARCH seem to exhibit the largest contrasts. The fact that the two methods can produce substantial differences in weights, but the log scores are always nearly the same, is consistent with the small values of $|f_T''(w)|$ in substantial neighborhoods of the optimal value of w evident in Figures 1 and 2 .

Figure 3 shows the evolution of the weight w_t^* in some two-model pools when pools are optimized using only past realizations of predictive densities. Not surprisingly w_t^* fluctuates violently at the start of the sample. Although the predictive densities are based on rolling five-year samples, w_t^* should converge almost surely to a limit under the conditions specified in Section 2. The HMNM and t-GARCH pool, upper left panel, might be interpreted as displaying this convergence, but the case for the pools involving EGARCH is not so strong.

Whether or not Section 2 provides a good asymptotic paradigm for the behavior of w_t^* is beside the point, however. The important fact is that a number of pools of two models outperform the model that performs best on its own (t-GARCH), performance being assessed by the log scoring rule in each case. The best of these two-model pools (HMNM and EGARCH) does not even involve t-GARCH, and it outperforms t-GARCH by 37 points. These findings illustrate the fresh perspective brought to model combination by linear pools of prediction models. Extending pools to more than two models provides additional interesting insights.

4 Pools of multiple models

In a prediction pool with n models the log predictive score function is

$$f_T(\mathbf{w}) = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{n} w_i p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_i) \right]$$

where $\mathbf{w} = (w_1, \ldots, w_n)'$, $w_i \geq 0$ $(i = 1, \ldots, n)$ and $\sum_{i=1}^{n} w_i = 1$. Given our assumptions about the data generating process D,

$$T^{-1} f_T(\mathbf{w}) \overset{a.s.}{\to} \lim_{T\to\infty} T^{-1} \int \log \left[\sum_{i=1}^{n} w_i p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_i) \right] p(\mathbf{Y}_T|D)\, d\nu(\mathbf{Y}_T) = f(\mathbf{w}).$$

Denote $p_{ti} = p\left(\mathbf{y}_t^o; \mathbf{Y}_{t-1}^o, A_i\right)$ $(t = 1, \ldots, T; i = 1, \ldots, n)$. Substituting $w_1 = 1 - \sum_{i=2}^n w_i$,

$$\partial f_T(\mathbf{w}) / \partial w_i = \sum_{t=1}^T \frac{p_{ti} - p_{t1}}{\sum_{j=1}^n w_j p_{tj}} \quad (i = 2, \ldots, n); \tag{11}$$

and

$$\partial^2 f_T(\mathbf{w}) / \partial w_i \partial w_j = -\sum_{t=1}^T \frac{(p_{ti} - p_{t1})(p_{tj} - p_{t1})}{\left[\sum_{k=1}^n w_k p_{tk}\right]^2} \quad (i, j = 2, \ldots, n).$$

The $n \times n$ Hessian matrix $\partial^2 f_T / \partial \mathbf{w} \partial \mathbf{w}'$ is non-positive definite for all $\mathbf{w}$ and, pathological cases aside, negative definite. Thus $f(\mathbf{w})$ is strictly concave on the unit simplex. Given the evaluations p_{ti} over the sample from the alternative prediction models, finding $\mathbf{w}_T^* = \arg\max_{\mathbf{w}} f_T(\mathbf{w})$ is a straightforward convex programming problem. The limit $f(\mathbf{w})$ is also concave in $\mathbf{w}$ and $\mathbf{w}^* = \arg\max_{\mathbf{w}} f(\mathbf{w}) = \lim_{T\to\infty} \mathbf{w}_T$.

Proposition 1 generalizes immediately to pools of multiple models.

Proposition 2 *If $A_1 = D$ then A_1 is dominant in the population pool $\{A_1, \ldots, A_m\}$ and*

$$\partial f(\mathbf{w}) / \partial w_j \mid_{\mathbf{w}=\widetilde{\mathbf{w}}} = 0 \quad (j = 1, \ldots, n)$$

where $\widetilde{\mathbf{w}} = (1, 0, \ldots, 0)'$.

Proof. *From (11),*

$$\frac{\partial f(\mathbf{w})}{\partial w_j} \mid_{\mathbf{w}=\widetilde{\mathbf{w}}} = \lim_{T\to\infty} T^{-1} \sum_{t=1}^T E_D\left[\frac{p(\mathbf{y}_t; \mathbf{Y}_{t-1}, A_i)}{p(\mathbf{y}_t; \mathbf{Y}_{t-1}, D)} - 1\right] = 0 \quad (j = 2, \ldots, m)$$

and consequently $\partial f(\mathbf{w}) / \partial w_1 \mid_{\mathbf{w}=\widetilde{\mathbf{w}}} = 0$ as well. From the concavity of $f(\mathbf{w})$, $\mathbf{w}^* = \widetilde{\mathbf{w}}$.

■

Extending the definitions of Section 2, models $A_1, \ldots, A_m$ $(m < n)$ are *jointly excluded* in the pool $\{A_1, \ldots, A_n\}$ if $\sum_{i=1}^m w_{Ti}^* = 0$; they are *jointly competitive* in the pool if $0 < \sum_{i=1}^m w_{Ti}^* < 1$; and they *jointly dominate* the pool if $\sum_{i=1}^m w_{Ti}^* = 1$. Obviously any pool has a smallest dominant subset. A pool trivially dominates itself. There are useful relations between exclusion, competitiveness and dominance that are useful in interpreting and constructing optimal prediction pools.

Proposition 3 *If $\{A_1, \ldots, A_m\}$ dominates the pool $\{A_1, \ldots, A_n\}$ then $\{A_1, \ldots, A_m\}$ dominates $\{A_1, \ldots, A_m, A_{j_1}, \ldots, A_{j_k}\}$ for all $\{j_1, \ldots, j_k\} \subseteq \{m+1, \ldots, n\}$.*

Proof. By assumption $\{A_{m+1}, \ldots, A_n\}$ is excluded in the pool $\{A_1, \ldots, A_n\}$. The pool $\{A_1, \ldots, A_m, A_{j_1}, \ldots, A_{j_k}\}$ imposes the constraints $w_i = 0$ for all $i > m$, $i \neq \{j_1, \ldots, j_k\}$. Since $\{A_{m+1}, \ldots, A_n\}$ was excluded in $\{A_1, \ldots, A_n\}$ these constraints are not binding. Therefore $\{A_{j_1}, \ldots, A_{j_k}\}$ is excluded in the pool $\{A_1, \ldots, A_m, A_{j_1}, \ldots, A_{j_k}\}$.

■

Thus a dominant subset of a pool is dominant in all subsets of the pool in which it is included.

Proposition 4 *If* $\{A_1, \ldots, A_m\}$ *dominates all pools* $\{A_1, \ldots, A_m, A_j\}$ $(j = m+1, \ldots, n)$ *then* $\{A_1, \ldots, A_m\}$ *dominates the pool* $\{A_1, \ldots, A_n\}$.

Proof. The result is a consequence of the concavity of the objective functions. The assumption implies that there exist weight $w_2^*, \ldots, w_m^*$ such that $\partial f_T\left(w_2^*, \ldots, w_m^*, w_j\right)/\partial w_j < 0$ when evaluated at $w_j = 0$ $(j = m+1, \ldots, n)$. Taken jointly these $n-m$ conditions are necessary and sufficient for $w_{m+1} = \ldots w_n = 0$ in the optimal pool created from the models $\{A_1, \ldots, A_n\}$. ■

The converse of Proposition 4 is a special case of Proposition 3. Taken together these propositions provide an efficient means to show that a small group of models is dominant in a large pool.

Proposition 5 *The set of models* $\{A_1, \ldots, A_m\}$ *is excluded in the pool* $\{A_1, \ldots, A_n\}$ *if and only if* A_j *is excluded in each of the pools* $\{A_j, A_{m+1}, \ldots, A_n\}$ $(j = 1, \ldots, m)$.

Proof. This is an immediate consequence of the first-order conditions for exclusion, just as in the proof of Proposition 4. ■

Proposition 6 *If the model* A_1 *is excluded in all pools* (A_1, A_i) $(i = 2, \ldots, n)$ *then* A_1 *is excluded in the pool* $(A_1, \ldots, A_n)$.

Proof. From (9) and the concavity of f_T the assumption implies

$$T^{-1}\sum_{t=1}^{T} p_{t1}/p_{ti} \leq 1 \quad (i = 2, \ldots, n)\,. \tag{12}$$

Let $\widetilde{w}_i$ $(i = 2, \ldots, n)$ be the optimal weights in the pool $(A_2, \ldots, A_n)$. From (11)

$$T^{-1}\sum_{t=1}^{T} \frac{p_{ti}}{\sum_{j=2}^{n} \widetilde{w}_j p_{tj}} = \lambda \text{ if } \widetilde{w}_i > 0 \;\; (i = 2, \ldots, n) \tag{13}$$

for some positive but unspecified constant λ. From (12) and Jensen's inequality

$$T^{-1}\sum_{t=1}^{T} \frac{p_{t1}}{\sum_{j=2}^{n} \widetilde{w}_j p_{tj}} < T^{-1}\sum_{t=1}^{T}\sum_{i=2}^{n} \widetilde{w}_i \frac{p_{t1}}{p_{ti}} < 1. \tag{14}$$

Suppose $\widetilde{w}_i > 0$. From (13)

$$T^{-1}\sum_{t=1}^{T} \frac{p_{ti}}{\sum_{j=2}^{n} \widetilde{w}_j p_{tj}} = T^{-1}\sum_{T=1}^{T}\sum_{\ell=2}^{n} \widetilde{w}_\ell \frac{p_{t\ell}}{\sum_{j=2}^{n} \widetilde{w}_j p_{tj}} = 1 \;\; (i = 2, \ldots, n)\,. \tag{15}$$

From (14) and (15),

$$T^{-1}\sum_{t=1}^{T} \frac{p_{ti} - p_{t1}}{\sum_{j=2}^{n} \widetilde{w}_j p_{tj}} \geq 0 \;\; (i = 2, \ldots, n)\,.$$

Since $w_1 = 1 - \sum_{i=2}^{n} w_i$, it follows from (11) that $\partial f_T(\mathbf{w})/\partial w_1 \leq 0$ at the point $\mathbf{w} = (0, \widetilde{w}_2, \ldots, \widetilde{w}_n)'$. Because f_T is concave this is necessary and sufficient for A_1 to be excluded in the pool $(A_1, \ldots, A_n)$. ■

Proposition 6 shows that one can establish the exclusion of A_1 in the pool $\{A_1, \ldots, A_n\}$, or for that matter any subset of the pool $\{A_1, \ldots, A_n\}$ that includes A_1, by showing that A_1 is excluded in the two-model pools $\{A_1, A_i\}$ for all A_i that make up the larger pool.

The converse of Proposition 6 is false. That is, a model can be excluded in a pool with three or more models, and yet it is competitive in some (or even all) pairwise pools. Consider $T = 2$ and the following values of p_{ti}:

	A_1	A_2	A_3
$t = 1$	0.4	0.1	1.0
$t = 2$	0.4	1.0	0.1

The model A_1 is competitive in the pools $\{A_1, A_2\}$ and $\{A_1, A_3\}$ because in (9) $f_T'(0) > 0$ and $f_T'(1) < 0$ in each pool. In the optimal pool $\{A_2, A_3\}$ the models A_2 and A_3 have equal weight with $\sum_{t=1}^{2} \sum_{j=2}^{3} \widetilde{w}_j p_{tj} = 0.55$. The first-order conditions in (11) are $\partial f_T(\mathbf{w})/\partial w_2 = \partial f_T(\mathbf{w})/\partial w_3 = 0.3/0.55 > 0$ and therefore the constraint $w_1 \geq 0$ is binding in the optimal pool $\{A_1, A_2, A_3\}$. The contours of the log predictive score function are shown in Figure 4(a).

Notice also in this example that $LS(\mathbf{Y}_T^o, A_1) = -1.833 > -2.302 = LS(\mathbf{Y}_T^o, A_2) = LS(\mathbf{Y}_T^o, A_3)$, and thus the model with the highest log score can be excluded from the optimal pool. The same result holds in the population: the Kullback-Leibler distance from D to A_1 may be less than the distance from D to A_j $(j = 2, \ldots, m)$ and yet A_1 may be excluded in the population pool $\{A_1, \ldots, A_m\}$ so long as $m > 2$. If $m = 2$ then the model with the higher log score is always included in the optimal pool.

No significantly stronger version of Proposition 6 appears to be true. Consider the conjecture that if model A_1 is excluded in one of the pools $\{A_1, A_i\}$ $(i = 2, \ldots, n)$, then A_1 is excluded in the pool $\{A_1, \ldots, A_n\}$. The contrapositive of this claim is that if A_1 is competitive in $\{A_1, \ldots, A_n\}$ then it is competitive in $\{A_1, A_i\}$ $(i = 2, \ldots, n)$, and by extension A_1 wold be competitive in any subset of $\{A_1, \ldots, A_n\}$ that includes A_1. That this not true may be seen from the following example with $T = 4$:

	A_1	A_2	A_3
$t = 1$	0.8	0.9	1.3
$t = 2$	1.2	1.1	0.7
$t = 3$	0.9	1.0	1.1
$t = 4$	1.1	1.0	0.9

The optimal pool $\{A_1, A_2, A_3\}$ weights the models equally, as may be verified from (11). But A_1 is excluded in the pool $\{A_1, A_2\}$: assigning w to A_1, (9) shows

$$f_T'(0) = \frac{-0.1}{0.9} + \frac{0.1}{1.1} + \frac{-0.1}{1} + \frac{0.1}{1} < 0.$$

The contours of the log predictive score function are shown in Figure 4(b).

5 Multiple-model pools: An example

Using the same S&P 500 returns data set described in Section 3 it is easy to find the optimal linear pool of all six prediction models described in that section. (The optimization required 0.22 seconds using conventional Matlab software, illustrating the trivial computations required for log score optimal pooling once the predictive density evaluations are available.) The first line of Table 3 indicates the composition of the optimal pool and the associated log score. The EGARCH, t-GARCH and HMNM models are jointly dominant in this pool while Gaussian, GARCH and SVOL are excluded. In the optimal pool the highest weight is given to t-GARCH, the next highest to EGARCH, and the smallest positive weight to HMNM.

Weights do not indicate a predictive model's contribution to log score, however. The next three lines of Table 3 show the impact of excluding one of the models dominant in the optimal pool. The results show that HMNM makes the largest contribution to the optimal score, 31.25 points; EGARCH the next largest, 19.47 points; and t-GARCH the smallest, 15.51 points. This ranking strictly reverses the ranking by weight in the optimal pool. When EGARCH is removed GARCH enters the dominant pool with a small weight, whereas the same models are excluded in the optimal pool when either t-GARCH or HMNM is removed.

These characteristics of the pool are evident in Figure 5, which shows log predictive score contours for the dominant three-model pool on the unit simplex. Weights for EGARCH and t-GARCH are shown explicitly on the horizontal and vertical axes, with residual weight on HMNM. Thus the origin corresponds to HMNM, the lower right vertex of the simplex to EGARCH, and the upper left vertex to t-GARCH. Values of the log score for the pool at those points can be read from Table 1. The small circles indicate optimal pools formed from two of the three models: EGARCH and HMNM on the horizontal axis, t-GARCH and HMNM on the vertical axis, and EGARCH and t-GARCH on the diagonal. Values of the log score for the pool at those points can be read from the last three entries in the last column of Table 3. The optimal pool is indicated by the asterisk. Moving away from this point, the log-score function is much steeper moving toward the diagonal than toward either axis. This reflects the large contribution of HMNM to log-score relative to the other two models just noted.

The optimal pool could not be used in actual prediction 1976-2005 because the weights draw on all of the returns from that period. As in Section 3, optimal weights can be computed each day to form a prediction pool for the next day. These weights are portrayed in Figure 6. There is substantial movement in the weights, with a noted tendency for the weight on EGARCH to be increasing at the expense of t-GARCH even late in the period. Nevertheless the log score function for the prediction model pool constructed in this way is -9267.82, just 3 points lower than the pool optimized

over the entire sample. Moreover this value substantially exceeds the log score for any model over the same period, or for any optimal pool of two models (see Table 3).

This insensitivity of the pool log score to substantial changes in the weights reflects the shallowness of the objective function near its mode: a pool with equal weights for the three dominant models has a log score of -9265.62, almost as high as that of the optimal pool. This leaves essentially no possible return (as measured by the log score) to more elaborate methods of combining models like bagging (Breiman (1996)) or boosting (Friedman et al. (2000)). Whether these circumstances are typical can be established directly by applying the same kind of analysis undertaken in this section for the relevant data and models, a question left to future research.

6 Pooling and model improvement

The linear pool $\{A_1, A_2\}$ is superficially similar to the mixture of the same models. In fact the two are not the same, but there is an interesting relationship between their log predictive scores. Denote the mixture of A_1 and A_2

$$p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_{A_1}, \boldsymbol{\theta}_{A_2}, w, A_{1\cdot 2}\right) = wp\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_{A_1}\right) + (1-w)\, p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_{A_2}\right). \quad (16)$$

Equivalently there is an i.i.d. latent binomial random variable $\widetilde{w}_t$, independent of $\mathbf{Y}_{t-1}$, $P\left(\widetilde{w}_t = 1\right) = w$, with $y_t \sim p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_{A_1}\right)$ if $\widetilde{w}_t = 1$ and $y_t \sim p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_{A_2}\right)$ if $\widetilde{w}_t = 0$.

If the prediction model A_j is fully Bayesian (1) or utilizes maximum likelihood estimates in (2) then under weak regularity conditions

$$\begin{aligned} T^{-1} LS\left(\mathbf{Y}_T, A_j\right) &\overset{a.s.}{\to} \lim_{T\to\infty} T^{-1} \int \log p\left(\mathbf{Y}_T \mid \boldsymbol{\theta}^*_{A_j}, A_j\right) p\left(\mathbf{Y}_T \mid D\right) d\nu\left(\mathbf{Y}_T\right) \\ &= LS\left(A_j; D\right) \quad (j = 1, 2) \end{aligned}$$

where

$$\boldsymbol{\theta}^*_{A_j} = \arg\max_{\boldsymbol{\theta}_{A_j}} \lim_{T\to\infty} T^{-1} \int \log p\left(\mathbf{Y}_T \mid \boldsymbol{\theta}_{A_j}, A_j\right) p\left(\mathbf{Y}_T \mid D\right) d\nu\left(\mathbf{Y}_T\right) \quad (j = 1, 2), \quad (17)$$

sometimes called the pseudo-true values of $\boldsymbol{\theta}_{A_1}$ and $\boldsymbol{\theta}_{A_2}$. However $\boldsymbol{\theta}^*_{A_1}$ and $\boldsymbol{\theta}^*_{A_2}$ are not, in general, the pseudo-true values of $\boldsymbol{\theta}_{A_1}$ and $\boldsymbol{\theta}_{A_2}$ in the mixture model $A_{1\cdot 2}$, and w^* is not the pseudo-true value of w. These values are instead

$$\begin{aligned} \left\{\boldsymbol{\theta}^{**}_{A_1}, \boldsymbol{\theta}^{**}_{A_2}, w^{**}\right\} &= \arg\max_{\boldsymbol{\theta}_{A_1}, \boldsymbol{\theta}_{A_2}, w} \lim_{T\to\infty} T^{-1} \int \sum_{t=1}^{T} \log\left[wp\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_{A_1}\right)\right. \\ &\quad \left. + (1-w)\, p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}_{A_2}\right)\right] p\left(\mathbf{Y}_T \mid D\right) d\nu\left(\mathbf{Y}_T\right). \quad (18) \end{aligned}$$

Let $w^* = \arg\max_w f(w)$. Note that

$$\begin{aligned}
& \lim_{T\to\infty} T^{-1} \int \sum_{t=1}^{T} \log\left[w^{**} p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}^{**}_{A_1}\right)\right. \\
& \left. + (1-w^{**})\, p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}^{**}_{A_1}\right)\right] p\left(\mathbf{Y}_T \mid D\right) d\nu\left(\mathbf{Y}_T\right) \\
\geq\ & \lim_{T\to\infty} T^{-1} \int \sum_{t=1}^{T} \log\left[w^{*} p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}^{*}_{A_1}\right)\right. \\
& \left. + (1-w^{*})\, p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}, \boldsymbol{\theta}^{*}_{A_1}\right)\right] p\left(\mathbf{Y}_T \mid D\right) d\nu\left(\mathbf{Y}_T\right) \\
=\ & w^* LS\left(A_j; D\right) + \left(1-w^*\right) LS\left(A_j; D\right).
\end{aligned}$$

Therefore the best log predictive score that can be obtained from a linear pool of the models A_1 and A_2 is a lower bound on the log predictive score of a mixture model constructed from A_1 and A_2. This result clearly generalizes to pools and mixtures of n models.

To illustrate these relationships, suppose the data generating process D is $y_t \sim N(1,1)$ if $y_{t-1} > 0$, $y_t \sim N(-1,1)$ if $y_{t-1} < 0$. In model A_1, $y_t \overset{iid}{\sim} N(\mu, \sigma^2)$ with $\mu \geq 1$ and in model A_2, $y_t \overset{iid}{\sim} N(\mu, \sigma^2)$ with $\mu \leq -1$. Corresponding to (17) the pseudo-true value of μ is 1 in A_1 and -1 in A_2; the psuedo-true value of σ^2 is 3 in both models. The expected log score, approximated by direct simulation, is -1.974 in both models. This value is indicated by the dashed (green) horizontal line in Figure 7. The function $f(w)$, also approximated by direct simulation, is indicated by the concave solid (red) curve in the same figure. The maximum, at $w = 1/2$, is $f(w) = -1.866$. Thus $f_T(w)$ would indicate that neither model could coincide with D, even for small T.

The mixture model (16) will interpret the data as independent and identically distributed, and the pseudo-true values corresponding to (18) will be $\mu = 1$ for one component, $\mu = -1$ for the other, and $\sigma^2 = 1$ in both. The expected log score, approximated by direct simulation, is -1.756, indicated by the dotted (blue) horizontal line in Figure 7. In the model $A = D$, $y_t \mid (y_{t-1}, A)$ has mean -1 or 1, and variance 1. Its expected log score is $-(1/2)[\log(2\pi) - 1] = -1.419$, indicated by the solid (black) horizontal line in the figure.

The example illustrates that $\max f(w)$ can fall well short of the mixture model expected log score, and that the latter can, in turn, be much less than the data generating process expected log score. It is never possible to show that $A = D$: only to adduce evidence that $A \neq D$.

7 Conclusion

In any decision-making setting requiring prediction there will be competing models. If one is willing to condition on one of the models available being true, then econometric

theory is comparatively tidy. In both Bayesian and non-Bayesian approaches, it is typically the case that one of a fixed number of models will come to dominate as sample size increases without bound.

At least in social science applications there is no reason to believe that any of the models under consideration is true and in many instances there is ample evidence that none could be true. This study develops an approach to model combination designed for such settings. It shows that linear prediction pools generally yield superior predictions as assessed by a conventional log score function. (This finding does not depend on the existence of a true model.) An important characteristic of these pools is that prediction model weights do not necessarily tend to zero or one asymptotically, as is the case for posterior probabilities. (This result invokes the existence of a true model.) The example studied here involves six models and a large sample. One of these models has posterior probability very nearly one. Yet three of the six models in the pool have positive weights, all substantial.

Optimal log scoring of prediction pools has three practical advantages. First, it is easy to do: compared with the cost of specifying the constituent models and conducting formal inference for each, it is practically costless. Second, the behavior of the log score as a function of model weights can show clearly that none of the models under consideration is true, or even close to true as measured by Kullback-Leibler distance. Third, linear prediction pools provide an easy way to improve predictions as assessed by the log score function. The example studied in this paper illustrates how acknowledging that all the available models are false can result in improved predictions, even as the search for better models goes on.

The last result is especially important. Our examples showed how models that are clearly inferior to others in the pool nevertheless substantially improve prediction by being part of the pool rather than being discarded. The analytical results in Section 4 and the examples in Section 5 establish that the most valuable model in a pool need not be the one most strongly favored by the evidence interpreted under the assumption that one of several models is true. It seems to us that this is a lesson that should be heeded generally in decision-making of all kinds.

References

Bacharach J (1974). Bayesian dialogues. Unpublished manuscript, Christ Church College, Oxford University.

Bates JM, Granger CWJ (1969). The combination of forecasts. Operational Research Quarterly 20: 451-468.

Bernardo JM (1979). Expected information as expected utility. The Annals of Statistics 7: 686-690.

Box GEP (1980). Sampling and Bayes inference in scientific modeling and robustness. Journal of the Royal Statistical Society Series A 143: 383-430.

Breiman L (1996). Bagging predictors. Machine Learning 26: 123-140.

Bremmes JB (2004). Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. Monthly Weather Review 132: 338-347.

Brier GW (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review 78: 1-3.

Chong YY, Hendry DF (1986). Econometric evaluation of linear macro-economic models. Review of Economic Studies 53: 671-690.

Christoffersen PF (1998). Evaluating interval forecasts. International Economic Review 39: 841-862.

Clemen RT, Murphy AH, Winkler RL (1995). Screening probability forecasts: contrasts between choosing and combining. International Journal of Forecasting 11: 133-146.

Clements MP (2006). Evaluating the Survey of professional Forecasters probability distributions of expected inflation based on derived event probability forecasts. Empirical Economics 31: 49-64.

Corradi V, Swanson NR (2006a). Predictive density evaluation. Elliott G, Granger CWJ, Timmermann A (eds.), Handbook of Economic Forecasting. Amsterdam: North-Holland. Chapter 5, pp. 197-284.

Corradi V, Swanson NR (2006b). Predictive density and conditional confidence interval accuracy tests. Journal of Econometrics 135: 187-228.

Dawid AP (1984). Statistical theory: The prequential approach. Journal of the Royal Statistical Society Series A 147: 278-292.

DeGroot MH, Fienberg SE (1982). Assessing probability assessors: Calibration and refinement. Gupta SS, Berger JO (eds.), Statistical Decision Theory and Related Topics III, Vol. 1. New York: Academic Press. pp. 291-314.

di Finetti B, Savage LJ (1963). The elicitation or personal probabilities. Unpublished manuscript.

Diebold FX, Gunter TA, Tay AS (1998). Evaluating density forecasts with applications to financial risk management. International Economic Review 39: 863-883.

Friedman J, Hastie T, Tibshirani R (2000). Additive logistic regression: A statistical view of boosting. Annals of Statistics 28: 337-374.

Genest C, Weerahandi S, Zidek JV (1984). Aggregating opinions through logarithmic pooling. Theory and Decision 17: 61-70.

Geweke J (2001). Bayesian econometrics and forecasting. Journal of Econometrics 100: 11-15.

Geweke J (2005). Contemporary Bayesian Econometrics and Statistics. Hoboken: Wiley.

Geweke J, Amisano G (2007). Hierarchical Markov normal mixture models with applications to financial asset returns. European Central Bank working paper 831.

Gneiting T, Raftery AE (2007). Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association 102: 359-378.

Gneiting T, Balabdaoul F, Raftery AE (2007). Probability forecasts, calibration and sharpness. Journal of the Royal Statistical Society Series B 69: 243-268.

Good IJ (1952). Rational decisions. Journal of the Royal Statistical Society Series B 14: 107-114.

Gourieroux C, Monfort A (1989). Statistics and Econometric Models, Vol 2. Cambridge: Cambridge University Press.

Granger CWJ, White H, Kamstra M (1989). Interval forecasting: an analysis based upon ARCH-quantile estimators. Journal of Econometrics 40: 87-96.

Hall SG, Mitchell J (2007). Combining density forecasts. International Journal of Forecasting 23: 1-13.

Jacobs RA (1995). Methods for combining experts' probability assessments. Neural Computation 7: 867-88.

Jacquier E, Polson NG, Rossi PE (1994). Bayesian analysis of stochastic volatility models. Journal of Business and Economic Statistics 12: 371-389.

McConway KJ (1981). Marginalization and linear opinion pools. Journal of the American Statistical Association 76: 410-414.

Quandt RE (1974). A comparison of methods for testing nonnested hypotheses. Review of Economics and Statistics 56: 92-99.

Shuford EH, Albert A, Massengill HE (1966). Admissible probability measurement procedures. Psychometrika 31: 125-145.

Stone M (1961). The opinion pool. Annals of Mathematical Statistics 32: 1339-1342.

Timmermann A (2006). Forecast combination. Elliott G, Granger CWJ, Timmermann A (eds.), Handbook of Economic Forecasting. Amsterdam: North-Holland. Chapter 4, pp. 135-196.

Wallis KF (2005). Combining density and interval forecasts: A modest proposal. Oxford Bulletin of Economics and Statistics 67: 983-994.

Winker RL (1969). Scoring rules and the evaluation of probability assessors. Journal of the American Statistical Association 64: 1073-1078.

Winkler RL, Murphy AM (1968). "Good" probability assessors. Journal of Applied Meteorology 7: 751-758.

Table 1: Log predictive scores of the alternative models

Gaussian	GARCH	EGARCH	t-GARCH	SV	HMNM
−10570.80	−9574.41	−9549.41	−9317.50	−9460.93	−9336.60

Table 2: Optimal pools of two predictive models

	Gaussian	GARCH	EGARCH	t-GARCH	SV	HMNM
Gaussian		−9539.72 −9541.42	−9505.57 −9507.73	−9317.50 −9318.65	−9460.45 −9461.99	−9336.60 −9337.48
GARCH	0.957 0.943		−9514.26 −9516.47	−9317.12 −9317.48	−9417.88 −9419.84	−9310.59 −9313.55
EGARCH	0.943 0.920	0.628 0.386		−9296.08 −9298.29	−9380.07 −9383.15	−9280.34 −9282.68
t-GARCH	1.000 0.984	0.944 0.931	0.677 0.861		−9317.50 −9318.15	−9284.72 −9287.28
SV	0.986 0.971	0.494 0.384	0.421 0.453	0.000 0.007		−9323.88 −9325.50
HMNM	1.000 0.996	0.628 0.611	0.529 0.670	0.289 0.307	0.713 0.787	

Entries above the diagonal are log scores of optimal pools. Entries below the diagonal provide the weight of the model in that row in the optimal pool. The top entry in each pair reflects optimization using the entire sample and the bottom entry reflects continuous updating of weights using only the data available on each date. Bottom entries below the diagonal indicate the average weight over the sample.

Table 3: Optimal pools of 6 and 5 models

Gaussian	GARCH	EGARCH	t-GARCH	SV	HMNM	log score
0.000	0.000	0.319	0.417	0.000	0.264	-9264.83
0.000	0.060	X	0.653	0.000	0.286	-9284.30
0.000	0.000	0.471	X	0.000	0.529	-9280.34
0.000	0.000	0.323	0.677	0.000	X	-9296.08

The first six columns provide the weights for the optimal pools and the last column indicates the log score of the optimal pool. "X" indicates that a model was not included in the pool.

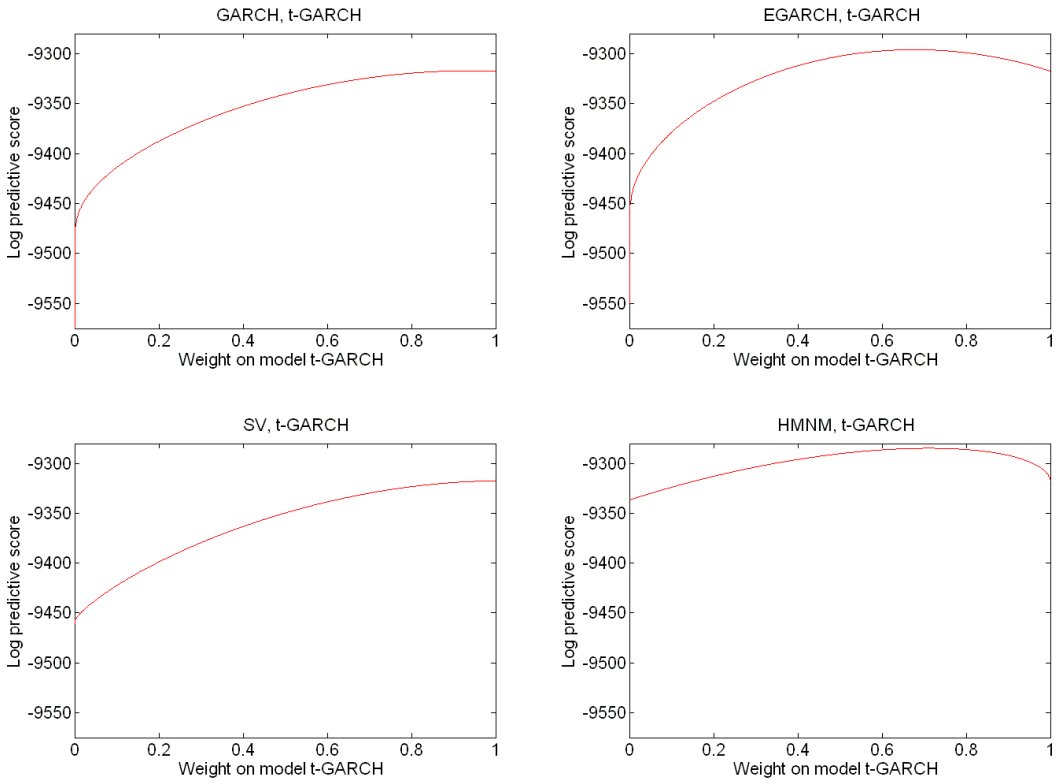

Figure 1: Log predictive score as a function of model weight in some two-model pools of S&P 500 predictive densities 1976-2005. Prediction models are described in the text.

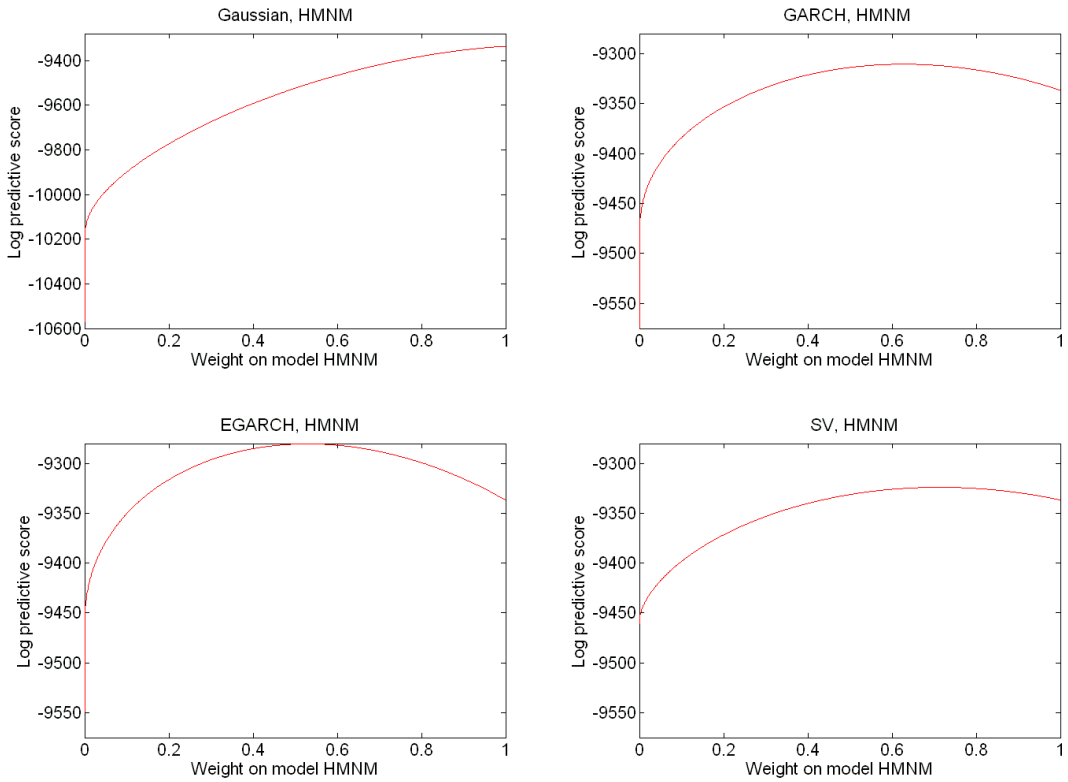

Figure 2: Log predictive score as a function of model weight in some two-model pools of S&P 500 predictive densities 1976-2005. Prediction models are described in the text.

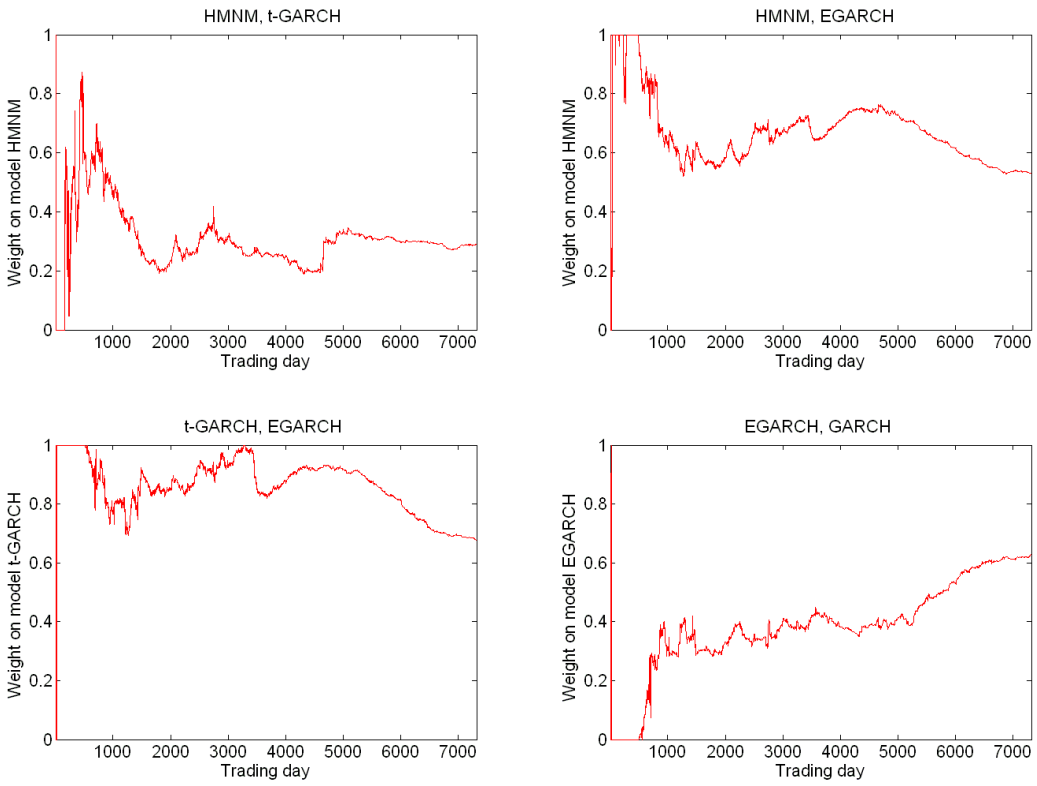

Figure 3: Evolution of model weights in some some two-model pools of S&P 500 predictive densities 1976-2005. Prediction models are described in the text.

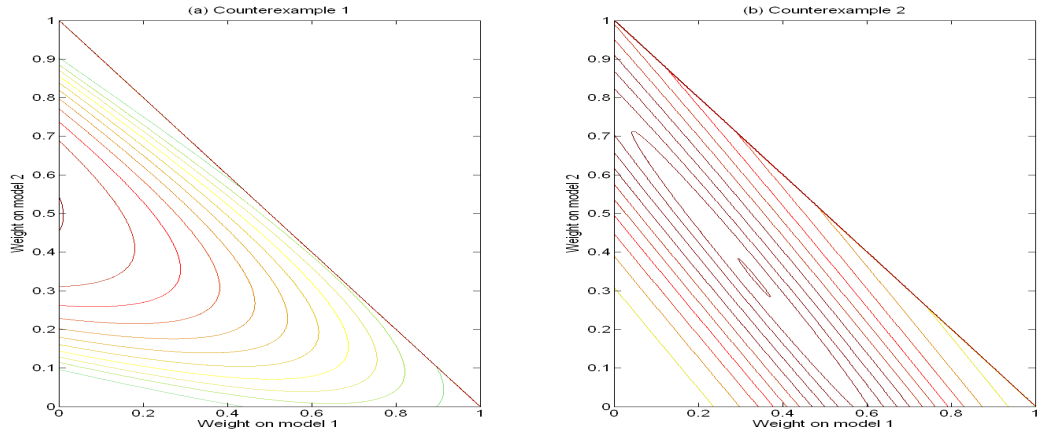

Figure 4: Panel (a) is a counterexample to the converse of Proposition 6. Panel (b) is a counterexample to a conjectured strengthening of Proposition 6.

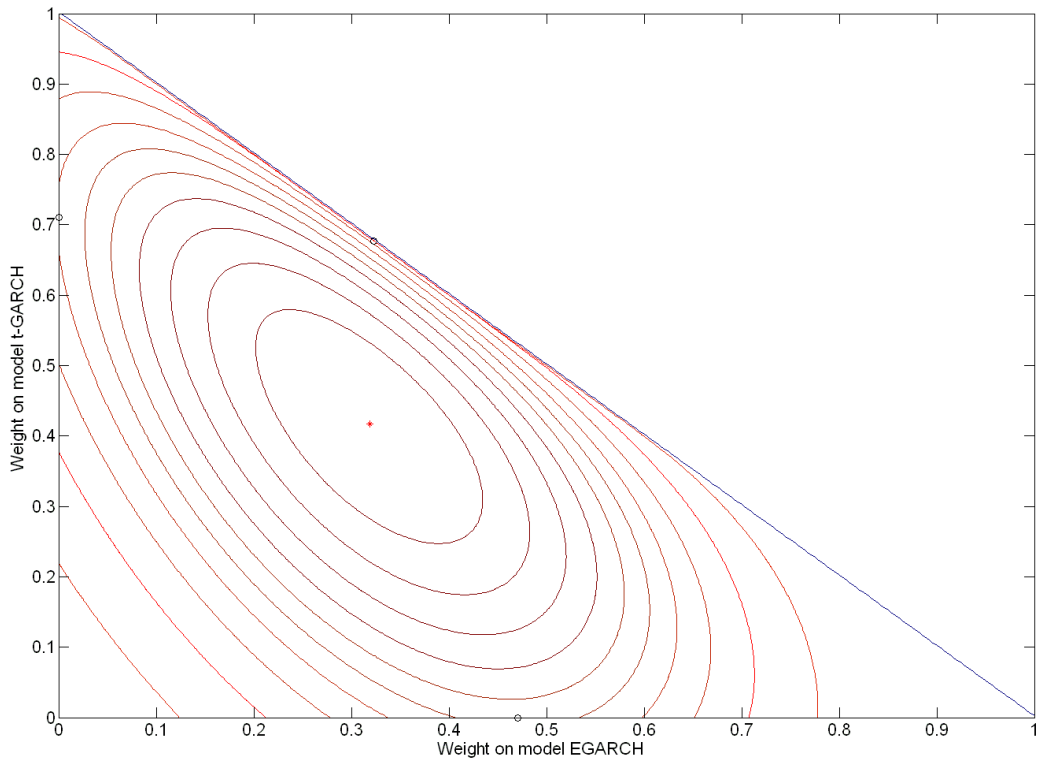

Figure 5: Contours of the log-score function (6) for the three models dominant in the six-model prediciton pool for S%P 500 returns 1972-2005. Residual weight accrues to the HMNM model. The three small circles indicate optimal two-model pools.

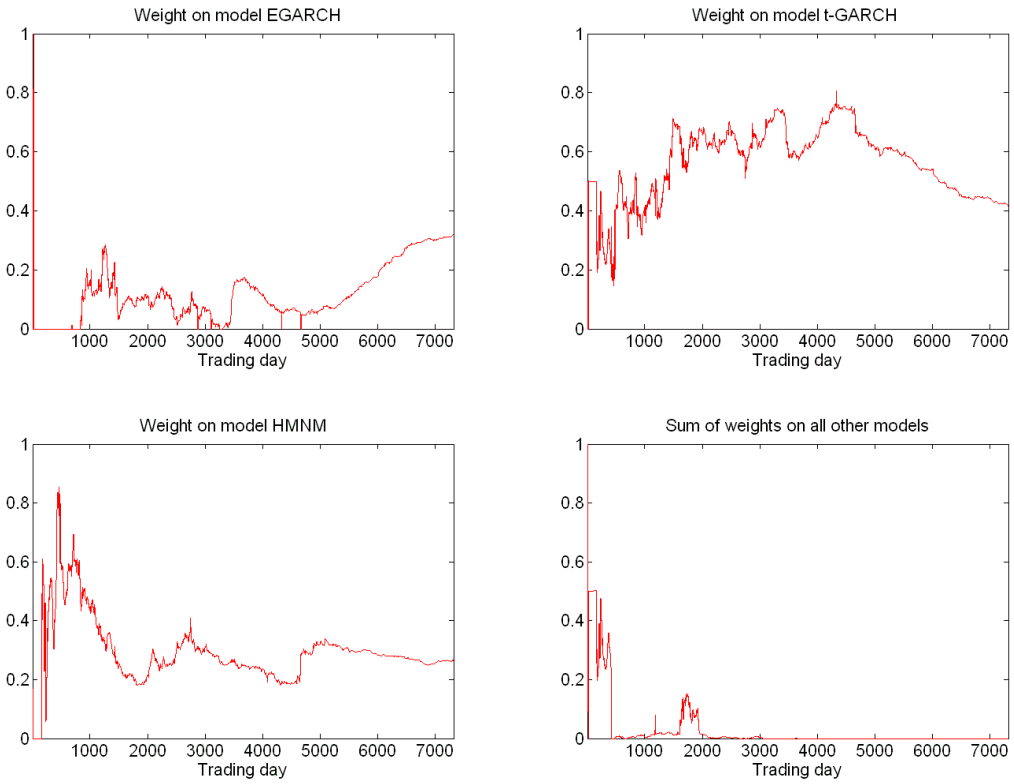

Figure 6: Evolution of model weights in the six-model pool of S&P 500 predictive densities 1976-2005. Prediction models are described in the text.

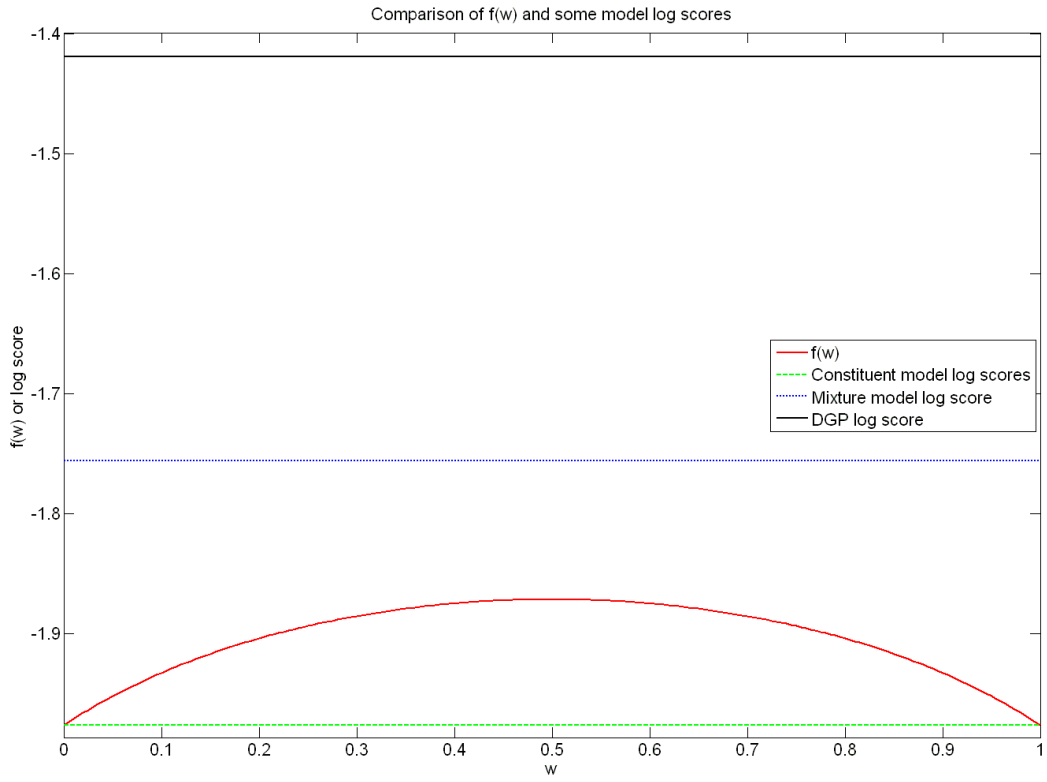

Figure 7: Expected log scores for individual models, a linear model pool, a mixture model, and the data generating process.