Improving Asset Price Prediction when All Models are False

Garland Durham[*] and John Geweke[†]

February 2011

Abstract

This study considers three alternative sources of information about volatility potentially useful in predicting daily asset returns: past daily returns, past intraday returns, and a volatility index based on observed option prices. For each source of information the study begins with several alternative models, and then works from the premise that all of these models are false to construct a single improved predictive distribution for daily S&P 500 index returns. The criterion for improvement is the log predictive score, equivalent to the average probability ascribed *ex ante* to observed returns. The first implication of the premise is that conventional models within each class can be improved. The paper accomplishes this by introducing flexibility in the conditional distribution of returns, in volatility dynamics, and in the relationship between observed and latent volatility. The second implication of the premise is that model pooling will provide prediction superior to the best of the improved models. The paper accomplishes this by constructing *ex ante* optimal pools, which place a premium on diversification in much the same way as optimal portfolios. All procedures are strictly out-of-sample, recapitulating one-step-ahead predictive distributions that could have been constructed for daily returns beginning January 2, 1992, and ending March 31, 2010. The prediction probabilities of the optimal pool exceed those of the conventional models by as much as 7.75%. The optimal pools place substantial weight on models using each of the three sources of information about volatility.

KEYWORDS: EGARCH, intradaily returns, model combination, optimal pool, S&P 500, stochastic volatility, VIX.

JEL classification: G17 primary, C52 secondary

[*] Division of Finance, Leeds School of Business, University of Colorado USA; Garland.Durham@colorado.edu

[†] Centre for the Study of Choice, University of Technology Sydney, Ultimo NSW 2007, Australia; and Division of Finance, Leeds School of Business, University of Colorado; John.Geweke@uts.edu.au. Geweke acknowledges financial support for this work through Australia Research Council grant DP110104732.

1 Introduction

Prediction of financial asset prices is important in a variety of private and public sector policy contexts. Examples include the pricing of options by private traders; the measurement of risk in mortgage pools by banks and Federal agencies; and assessment of systemic risk by regulatory agencies and macroeconomic policy makers. In all of these decision-making activities formal prediction models for asset prices are essential tools, and the academic literature has responded with a wide variety of candidates. Yet, those with responsibilities for such decisions recognize that all of these models are incomplete descriptions of reality. How should a decision-maker proceed, knowing that all the models at her disposal are false?

The academic literature provides little practical guidance on this point. The orthodox rational expectations framework is not designed for this purpose. It avoids the issue by assuming that reality is fully described by a parametric model that is known to economic agents and policy makers. When this approach is extended to the situation where economic agents and policy makers must learn about reality, it is typically in the context of a correctly specified parametric model with unknown parameters. The mainstream econometric literature is also unhelpful for the decision-maker confronted with an array of alternative model-based predictive distributions for asset prices. Non-Bayesian econometrics emphasizes specification testing. But when all the available models are false, passing a battery of tests is an indicator of insufficient sample size and test power rather than evidence that a particular model is true and others are false. With sample sizes sufficiently large and tests sufficiently powerful, all models will be rejected, leaving the prediction question unresolved. Bayesian econometrics provides an elegant operational theory of model combination, but because it is founded on the explicit condition that one of the models under consideration is a literal description of reality it shares the limitations of the rational expectations literature.

This paper looks at several different classes of models that generate predictive distributions for asset prices by making use of alternative sources of information (daily returns only, high-frequency intraday returns, and option-implied volatility), and uses the method of optimal prediction pooling developed in Geweke and Amisano (2010) to construct predictive densities that outperform any of the individual models. The situation is typical in that each class of models provides a window into the underlying reality, but we do not believe any of them to be literally true. The optimal pooling idea makes explicit allowance for the possibility that all of the models under consideration are false and reflects the observed behavior of decision-makers, who are likely to consult several different models when making policy, even models that have been rejected by formal statistical tests. While such behavior seems paradoxical, it is supported by the finding that pools are typically able to outperform even the best of the individual models they encompass, sometimes by a large margin, while placing significant weight on models that are easily rejected by conventional tests.

The study proceeds in two steps. First, we construct the collection of individual models. We look at a total of 42 models categorized into three groups based on the source of information used to forecast volatility. The models allow for flexibility in the shape of the predictive distribution, the way this shape changes over time, and the relationship between observed and latent volatility, building on methods introduced in Durham (2007). The second step involves building on this extended collection of asset price prediction models, constructively using the premise that none of the models are true to generate improved predictive distributions using the method of optimal prediction pooling. The application uses daily S&P 500 index returns from the first trading day of 1990 through the end of March 2010.

The models are described in Section 2. The first group uses the history of daily returns only and comprises six models: two stochastic volatility (SV) models with leverage and four exponential generalized autoregressive heteroscedasticity (EGARCH) models. In the second group, consisting of 18 models, the indicator of daily volatility is the sum of squared 5-minute S&P 500 index returns from previous trading days. The third group, also consisting of 18 models, uses the Chicago Board Options Exchange Market Volatility Index (VIX), which is a model-free indicator of daily volatility constructed from options prices.

In each group of models, we begin with a simple base model and elaborate on it in several directions. In all groups the daily return shock is either Gaussian or a mixture of two normal distributions. In all models except the stochastic volatility models the volatility component is either a single factor (the conventional treatment) or the sum of two independent factors with different autocorrelation properties. Permitting more than one volatility factor introduces flexibility into the autocorrelation of the latent volatility state, allowing the models to generate, for example, long memory-like behavior. In the high-frequency and options groups there is a third dimension of flexibility related to the form of the mapping from the volatility state to spot volatility (the scaling factor for daily returns).

Turning to the second step, Section 3 reviews the essentials of optimal prediction pools. Each model generates a predictive density rule (PDR), which is a mapping from the information set at time $t-1$ to a predictive density for the return at time t. While there are many ways that one could construct a prediction rule from a given model, we use a maximum likelihood plug-in rule. That is, given a model, for each time t in the sample we calculate the maximum likelihood estimator (MLE) over historical data available at time $t-1$ (expanding samples) and use the predictive density obtained from the model with that value for the parameter vector. This must be repeated for each model and each time t in the sample (42 models and nearly 5000 observations in our application), necessitating computationally efficient means of estimating the models. For the SV models, we integrate across uncertainty in the volatility states. For the other models, the volatility state is a deterministic function of the model parameters and history of observables. The theory does not rely upon the precise manner in which the PDR is formed. Rather than MLE, one could use, for example, a generalized method of moments (GMM) plug-in rule or take a fully Bayesian approach where the PDR is formed by integrating across both parameter and state uncertainty.

Any linear combination of prediction rules, subject to the constraints that the weights are nonnegative and sum to one, is itself a prediction rule referred to as a prediction pool. Geweke and Amisano (2010) describe the construction of an optimal prediction rule, which uses weights obtained by optimizing over historical performance. Performance is assessed using a log score criterion. Although other scoring criteria are possible, the log score has attractive optimality conditions and is analogous to model assessments based on log likelihood or marginal likelihood in the conventional frequentist or Bayesian settings, respectively. The critical point is that the analysis is entirely out-of-sample. That is, the individual model PDRs as well as the pool weights depend only on data available in real-time. Since assessment is always based on out-of-sample performance, there is no need to attempt the usual adjustments penalizing model complexity.

The formation of optimal prediction pools has interesting parallels to optimal portfolio construction. Given a collection of assets, suppose the objective is to construct the portfolio which optimizes some criterion function, perhaps Sharpe ratio.

The optimal portfolio will typically include a mix of assets rather than placing all weight on the single asset with the highest Sharpe ratio. Even though a particular asset may not perform well on its own, it can improve the performance of the portfolio (diversification). The optimal portfolio will typically have better performance than any of the individual assets alone.

These same points also hold true for prediction pools. Given a collection of models, the goal is to optimize some criterion function (e.g., log score). The optimal pool typically includes a mix of models rather than placing all weight on a single model. Even though a particular model may do poorly on its own, it can improve the performance of the pool. The pool will generally have better performance than any of the models it comprises.

As with optimal portfolio construction, optimal pooling tends to reward diversification, with several models (potentially all) having positive weights. It does so for similar reasons: just as an asset can enter a portfolio despite a comparatively low return when it is negatively correlated with the market return, a prediction model that produces an inferior predictive likelihood on average can enter an optimal prediction pool with positive weight if it occasionally but regularly outperforms the other prediction models. In either case the weights have well-defined almost sure limits given a particular choice of criterion function and stationary data-generating processes.

These features of optimal prediction pools are fundamentally different from Bayesian model averaging. In a stationary environment the posterior probability of the model with the highest average predictive likelihood, and therefore its weight in model averaging, tends to one as sample size increases. This feature arises precisely because Bayesian model averaging conditions on the truth of one of the n models. Given this condition, the true model is distinguished by having the highest expected predictive likelihood under the data generating process, and as sample size increases the evidence bearing on which model enjoys this distinction becomes overwhelming. In the more realistic case where none of the models is the data-generating process, all weight devolves onto the model closest to it in Kullback-Leibler distance. Non-Bayesian model selection methods have similar properties, identifying the model closest to the data generating process in Kullback-Leibler distance asymptotically.

Thus whereas a conventional econometric approach to model combination leads to a horse race with a single ultimate winner, an optimal pool typically consists of several models, each contributing a strength that balances some weakness in the other models entering the optimal pool. Under the log scoring rule the optimal pool proves superior to the model winning the horse race in the conventional approach.

Empirical results are reported in Section 4. For the simplest models in each group there are very substantial differences in log scores across the three classes (Section 4.1), with the basic daily model outperforming the basic options model which in turn outperforms the basic high-frequency model. The various model extensions described in Section 2 eliminate the bulk of the differences in log scores among the best models in each group, with the best daily model followed by the best high-frequency model followed by the best options model.

Yet these differences are still large. Conventional model combination procedures, motivated by Bayesian model averaging and described in Section 4.2, amount to "winner takes all": on most trading days predictions are based almost entirely on one group of models, but there are sharp fluctuations between groups. Daily models are most often chosen, followed by options models; high-frequency models never dominate. The performance of this model averaging procedure is poor, both in comparison with the best of our extended models and with some simple benchmark predictive density combinations.

Optimal prediction pools, constructed in Section 4.3, behave very differently. Following initial fluctuations, weights in the optimal pools stabilize several years into the sample. The sum of weights on the daily models is somewhat less than one-half. The sum of weights on the high-frequency models is about the same as for the options models. A related measure of model value indicates that in the latter years of the sample the values of the daily and high-frequency groups are nearly equal and about three times as great as the value of the options models. The optimal pools substantially outperform all of the individual models in log score, and they also outperform the simple benchmark predictive density combinations.

This study concentrates on the specific problem of extending and combining models that use alternative sources of information about volatility for the purpose of improving the one-step-ahead prediction of an index of asset prices. For sake of transparency we do not introduce notation or techniques more general than required to address this particular task. Yet the methodology in the study can be extended to a much wider set of similar problems. Some of these extensions are quite modest while others require addressing additional technical issues. Section 5 summarizes the findings of this study and then briefly discusses a much larger set of prediction problems amenable to similar treatment and the work involved.

2 Models and estimation techniques

We look at several classes of models, corresponding to different sources of information about volatility: daily returns, high-frequency, and options. For all of the models, returns are of the form

$$y_t = \mu_Y + \sigma_t \varepsilon_t, \tag{1}$$

where y_t is the daily log return, σ_t is the volatility scaling factor and ε_t is a mixture of normals standardized to have mean zero and variance 1. The model classes differ in the information used to estimate σ_t. In each class, we examine a hierarchy of models with varying amounts of flexibility in several relevant dimensions. All models are estimated by maximum likelihood. Predictive densities are then formed by replacing unknown parameters with their point estimates. In all cases, predictive densities for the time t return are constructed using only information available at time $t-1$.

Our objective is not just to obtain forecasts that match the observed data in, say, first or second moments. Rather, the object of interest is the full predictive density, with assessment using a likelihood-based metric closely related to Kullback-Leibler distance. Thus it is important for the models to have sufficient flexibility to generate realistic distributions, motivating the use of the mixture models. Mixtures of normals have nice properties. Given enough components they are able to fit any distribution arbitrarily closely (McLachlan and Peel, 2000). For distributions encountered in applications similar to the one in this paper, good fits are typically obtained with a small number of components.

These mixture models are closely related to the jump models commonly used in this literature. But, we do not take a stand on the nature of the intradaily price movements: what part is diffusive, what part is due to jumps, and what the characteristics of those jumps are. We are only interested in the shape of the daily return distributions. The mixture distributions are useful for this purpose. See Durham (2007) for additional detail.

We examined mixtures of up to three components. The three-component models perform well in the later part of the sample but have difficulty in the early part, where the quantity of available data is more limited. In full Bayesian estimation, the problems in the early part of the sample could be alleviated by using an appropriate prior. With the maximum likelihood approach used in this paper, an analagous effect could be achieved by adding curvature to the likelihood surface in an ad hoc manner. However, for the application in this paper we restrict attention to models with a maximum of two mixture components.

Some of the models include multiple volatility factors, providing flexibility in the autocorrelation characteristics of the latent volatility state. In models with two factors, for example, one captures a persistent long-term trend in the level of volatility, while the other captures short-term fluctuations around it. Such models are capable of generating long memory-like behavior (Bollerslev and Mikkelsen, 1996).

The class of daily models consists of two stochastic volatility (SV) and four exponential generalized autoregressive heteroscedasticity (EGARCH) models. The SV models are of the form

$$\begin{aligned} y_t &= \mu_Y + \sigma_Y \exp\left(v_{t-1}/2\right) \varepsilon_t \qquad (2) \\ v_t &= \phi v_{t-1} + \sigma_V \eta_t, \end{aligned}$$

where y_t is the log return and v_t is the unobserved volatility state. The volatility innovations are of form $\eta_t = \rho\varepsilon_t + \left(1-\rho^2\right)^{1/2} u_t$, where $u_t \sim N(0,1)$ is uncorrelated with ε_t. Thus $\mathrm{E}(\eta_t) = 0$, $\mathrm{var}(\eta_t) = 1$ and $\mathrm{corr}(\eta_t, \varepsilon_t) = \rho$, but because ε_t is non-Gaussian so is η_t. Negative values for ρ capture a leverage effect, whereby negative returns are associated with increased volatility on subsequent days. The nature of the relationship between ε_t and η_t implies that extreme price changes will tend to generate large changes in volatility as well. Estimation is done using the simulated maximum likelihood algorithm and EIS sampler of Richard and Liesenfeld (2006). Predictive densities are formed by integrating across uncertainty in the volatility state. We look at two partcular cases of the SV model: `sv_1` uses a Gaussian distribution for ε_t, and `sv_2` uses a mixture of two normal distributions.

The EGARCH models are of form

$$\begin{aligned} y_t &= \mu_Y + \sigma_Y \exp\left(\sum_{i=1}^{k} v_{it}/2\right) \varepsilon_t \qquad (3) \\ v_{i,t+1} &= \alpha_i v_{it} + \beta_i \left(|\varepsilon_t| - (2/\pi)^{1/2}\right) + \gamma_i \varepsilon_t \qquad (i = 1, \ldots, k). \end{aligned}$$

The model `egarch_kj` includes k volatility factors v_{it} and the normal mixture has j components ($k = 1, 2$; $j = 1, 2$).

The high-frequency models use a volatility signal extracted from five-minute intraday S&P 500 returns. Following Andersen et al. (2001), Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002), daily realized volatility was calculated by summing over squared intraday returns for each day t,

$$RV_t^{(\Delta)} = \sum_{j=1}^{1/\Delta} \left(y_{t-1+j\Delta} - y_{t-1+(j-1)\Delta} \right)^2, \tag{4}$$

where Δ is the sampling interval for the intraday data. In the application Δ corresponds to five-minute intervals. In (4) $t-1$ denotes the opening of the market on day t and t denotes the close (so intraday volatility does not include the return from market close on one day to market open on the following day).

In principle, high-frequency returns are capable of providing very precise information about the latent volatility state. In practice, there is measurement error related to, for example, market microstructure effects and non-synchronous trading, which the use of five-minute returns is intended to help alleviate (longer sampling intervals decrease the measurement error but at the cost of greater discretization error). Perhaps more critically, we are using the realized volatility observed on day t as a basis for forecasting day $t+1$ returns. Consistent with the literature, we also ignore the overnight return. So there is little reason to expect the realized volatility to be either an efficient or unbiased estimator for the variance of the next day's return.

We address these issues in two steps. First, we apply a filter to extract estimates of the latent volatility state from the realized volatility observations. We tried several different filters with up to three factors for the volatility state: exponential weighting; heterogeneous autoregressive model (Corsi, 2009); and Kalman filter. We also tried using the raw unfiltered data directly as an estimate of the volatility state. Among the various filtering schemes, there was not much difference; all improved predictions substantially. In all cases, the multifactor models performed much better in forecasting realized volatility, but the single-factor models were slightly better at forecasting returns (the objective of this paper). The results reported in the application use one-

or two-factor Kalman filters,

$$\begin{aligned} \log RV_t &= \mu_{RV} + \sum_{i=1}^{k} v_{it} + \omega_t \\ v_{i,t+1} &= \phi_i v_{it} + \nu_{it} \qquad (i = 1, \ldots, k). \end{aligned}$$

We decided to proceed with the Kalman filter rather than any of the alternative filters that perform about as well, because it is well-motivated as a basis for extracting a signal from noisy observations.

The second step is a mapping from the volatility state extracted in the previous step to σ_t, the scaling factor for daily returns:

$$\psi : \log \widehat{RV}_t \longrightarrow \log \sigma_t.$$

Since this mapping is of unknown form, we estimate it using flexible parametric methods. Polynomial expansions of sufficiently high degree are capable of approximating any smooth function to arbitrary accuracy on compact sets, and so are useful for this purpose. We looked at Legendre polynomials up to order three (the volatility states were first scaled and translated to mean zero and unit variance), but found no improvements beyond order two. The parameters of the mapping are estimated simultaneously with the parameters of (1), conditioning on the point estimates for the volatility state.

The model `hifreq_kjp` uses k independent latent volatility factors, j normal components in the mixture for ε_t, and a polynomial of order p in the mapping. We consider the cases ($k = 0, 1, 2$; $j = 1, 2$; $p = 0, 1, 2$) for a total of 18 high-frequency models. For the case $k = 0$, no filtering is done (that is, the observed RV_t is mapped directly into σ_t). The case $p = 0$ refers to a linear polynomial where the constant is estimated and the slope coefficient is one.

The options models have the same structure as the high-frequency models except that they substitute a measure of option-implied volatility IV_t in place of the high-frequency measure RV_t. We use the VIX index, a model-free measure of volatility implied by options prices (Britten-Jones and Neuberger, 2000). There is some measurement error involved when using the VIX index as a signal about the volatility state due to, for example, truncation and discreteness effects (Jiang and Tian, 2007). The measure is also biased due to the existence of risk-premia. Thus, similar considerations to those discussed in the context of the high-frequency models apply here as well.

The model `vix_kjp` uses k independent latent volatility factors, j normal components in the mixture for ε_t, and a polynomial of order p in the mapping. We consider the cases ($k = 0, 1, 2$; $j = 1, 2$; $p = 0, 1, 2$) for a total of 18 options models.

3 Pooling

Each of the 42 models just described provides a predictive density rule. A predictive density rule is an operator mapping information sets into predictive densities. For the asset return y_t on day t conditional on information available at the close of trading on day $t-1$, the predictive density takes the form $p_t\left(y_t \mid Y^o_{t-1}, X^o_{t-1}, \boldsymbol{\theta}_{A_i}, A_i\right)$, where the superscript "o" denotes the observed value (data) as distinguished from the *ex ante* random variable or argument of the density function and A_i indicates the particular model. The symbols Y_{t-1} and X_{t-1} indicate the sets of historical daily asset returns and covariates, respectively, available at the end of day $t-1$. In the high-frequency models X_{t-1} consists of the five-minute intraday returns on days $s < t$; for the options models it consists of the VIX index on days $s < t$; and for the daily models $X_{t-1} = \emptyset$. This section uses this generic notation throughout.

The decision-making context requires a single predictive density $p_t\left(y_t; Y^o_{t-1}, X^o_{t-1}\right)$ at the end of trading day $t-1$. Broadly speaking these contexts include any situation in which normative behavior presumes a subjective distribution for relevant unknown magnitudes, including conventional expected utility maximization. Special cases are conventional theories of asset derivative pricing and optimal macroeconomic policy. A decision-maker could choose among the alternative predictive densities $p_t\left(y_t; Y^o_{t-1}, X^o_{t-1}, A_i\right)$ or combine them.

3.1 Assessing the performance of predictive densities

Choosing among the possibilities requires a criterion. The decision-maker can use the observed values of past returns and covariates available at time $t-1$ to assess the performance of any stipulated predictive density rule, just as an investor can use the history of returns in creating an optimal portfolio. This set of primitives — the history $\left(Y^o_{t-1}, X^o_{t-1}\right)$ and a predictive density rule — 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 this basis is consistent with the prequential principle of Dawid (1984). These assessment procedures are widely known as scoring rules.

This study uses the log scoring rule

$$LS\left(Y_{t-1}^{o}; X_{t-1}^{o}, A_i\right) = \sum_{s=1}^{t-1} \log p_s\left(y_s^o; Y_{s-1}^{o}, X_{s-1}^{o}, A_i\right). \tag{5}$$

to assess the prediction performance of a model A_i over the sample period up to time $t-1$. This rule is easy to interpret, grounded in the literature, and has a significant axiomatic justification. With regard to interpretation, there is a simple relationship between (5) and the performance of alternative prediction rules. For alternative rules A_1 and A_2,

$$\Delta\left(A_1, A_2\right) = \exp\left\{\left[LS\left(Y_{t-1}^{o}; X_{t-1}^{o}, A_1\right) - LS\left(Y_{t-1}^{o}; X_{t-1}^{o}, A_2\right)\right] / \left(t-1\right)\right\} \tag{6}$$

is the geometric average of the ratio of probability densities assigned to the observed returns $y_1^o, \ldots, y_{t-1}^o$. This justifies the colloquial interpretation, "observed returns were $100 \cdot \left[\Delta\left(A_1, A_2\right) - 1\right]$ percent more probable under predictive density A_1 than they were under A_2."

With reference to the econometrics literature, for the specific case of Bayesian predictive densities $LS\left(Y_{t-1}^{o}; X_{t-1}^{o}, A_i\right)$ is the log predictive likelihood. In the even more specific case in which the sample begins at time $t = 1$ and sample size is T, $LS\left(Y_T^o; X_T^o, A_i\right)$ is the log marginal likelihood, which in turn is the foundation of the Bayesian approach to the model combination issue addressed in this study. (On predictive and marginal likelihoods see Geweke, 2005, Section 2.6.) The predictive densities employed in the work described here are not Bayesian, but Section 4.2 uses this relationship in drawing contrasts between model pooling and conventional model averaging procedures.

There is a superficial resemblance between (5) and the log likelihood function

$$\log L(\boldsymbol{\theta}_i; Y_T^o, X_T^o, A_i) = \sum_{s=1}^{T} \log p_s\left(y_s^o \mid Y_{s-1}^{o}, X_{s-1}^{o}, \boldsymbol{\theta}_i, A_i\right)$$

where T is sample size and the parameter vector $\boldsymbol{\theta}_i$ is the argument of the likelihood function in model A_i. The resemblance is incomplete and potentially misleading. It is incomplete because the candidate values for $\boldsymbol{\theta}_i$ are estimates that are functions of the entire sample, whereas only data from periods s and earlier enter p_s in (5). As a consequence over-fitting issues are critical in maximum likelihood estimation, leading to corrections like the familiar AIC and SBIC criteria. These issues do not arise in this study because all measures of performance are strictly out-of-sample.

With reference to the finance literature, the rule (5) parallels a time separable utility function in which the quantity of the single good consumed in period s is $p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o\right)$ and instantaneous utility is logarithmic. In the prototypical situation, consumption is return on wealth and the motivating problem is optimal portfolio allocation. Higher $p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o\right)$ is better than lower just as more consumption is preferred to less. Like all analogies, this one is incomplete. A major distinction that works to practical advantage in this study is that while the class of regular instantaneous utility functions is quite broad, the logarithmic form has the following axiomatic foundation.

Suppose a decision-maker asks her technical staff to provide predictive densities and announces a scoring rule she will use to assess these densities. Suppose further that each staff member reports the predictive density that maximizes the expected value of the announced scoring rule, the expectations being taken with respect to the staff member's private predictive density function. The scoring rule is said to be proper if, in such a situation, it induces each staff member to report truthfully his private density rather than some different predictive density with a higher expected score. The term "proper" was coined by Winkler and Murphy (1968) but the general idea dates back at least to Brier (1950) and Good (1952). An economist might say that the scoring rule provides an incentive for truthful revelation. If the scoring rule depends on $\left(Y_{t-1}^o, X_{t-1}^o\right)$ and $p_s\left(y_s; Y_{s-1}, X_{s-1}\right)$ $(s < t)$ only through $p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o\right)$ $(s < t)$ then it is said to be local (Bernardo, 1979).

The only proper local scoring rule is (5) and (trivially) monotone increasing linear transformations of (5). This was shown by Shuford et al. (1966) for the case in which the support of y_t is a finite set of at least three discrete points; for further discussion see Winkler (1969, p. 1075). The result was extended to continuous distributions by Bernardo (1979); for further discussion see Gneiting and Raftery (2007, p. 366).

3.2 Combining predictive densities

From the available collection of rules $p_s\left(y_s; Y_{s-1}, X_{s-1}, A_i\right)$ $(s \leq t;\ i = 1, \ldots, n)$ and data (Y^0_{t-1}, X^o_{t-1}) the decision-maker creates $p_t\left(y_t; Y^o_{t-1}, X^o_{t-1}\right)$. We refer to this mapping as a prediction pool, motivated by the more general descriptor opinion pool for a combination of subjective probability distributions originating with Stone (1961). There are endless ways in which the n predictive densities could be combined; see Genest et al. 1984) for a review and axiomatic approach. Restricting consideration to linear combinations leads to computations that are simple, both absolutely and in comparison with alternatives.[1] At the close of trading day $t-1$ the predictive density of a linear pool for the next trading day's return is

$$p\left(y_t; X^o_{t-1}, Y^o_{t-1}, \mathbf{w}_{t-1}\right) = \sum_{i=1}^{n} w_{t-1,i} p_t\left(y_t; Y^o_{t-1}, X^o_{t-1}, A_i\right) \tag{7}$$

where $\mathbf{w}_{t-1} = \left(w_{t-1,1}, \ldots, w_{t-1,n}\right)'$ is a weight vector satisfying

$$\sum_{i=1}^{n} w_{t-1,i} = 1;\ \ w_{t-1,i} \geq 0 \quad (i = 1, \ldots, n)\,. \tag{8}$$

These restrictions are sufficient to ensure that (7) is a density function. Applying the log scoring rule, this linear prediction pool is scored using

$$f_{t-1}\left(\mathbf{w}_{t-1}\right) = \sum_{s=1}^{t-1} \log\left[\sum_{i=1}^{n} w_{t-1,i} p_s\left(y^o_s; Y^o_{s-1}, X^o_{s-1}, A_i\right)\right] \tag{9}$$

and therefore the optimal weight vector $\mathbf{w}^*_{t-1}$ is chosen to maximize (9). The optimal weight vector is updated at the close of trading each day, reflecting the performance of the models in predicting that day's return. Geweke and Amisano (2010) shows that $f_t\left(\mathbf{w}_t\right)$ is at least weakly concave, and for $t \geq n$ $f_t\left(\mathbf{w}_t\right)$ is in general strictly concave. Maximization of f_t is therefore a regular convex programming problem and the optimal weights can be computed using conventional software.[2] Typically the vector $\mathbf{w}^*_{t-1}$ has several positive elements; that is, several models in fact enter the

[1] When prediction addresses vector y_t rather than scalar as is the case here, only linear combinations of predictive densities satisfy some basic conditions of internal consistency, as first shown by McConway (1981).

[2] Results reported in Section 4 all use the Matlab function `fmincon`.

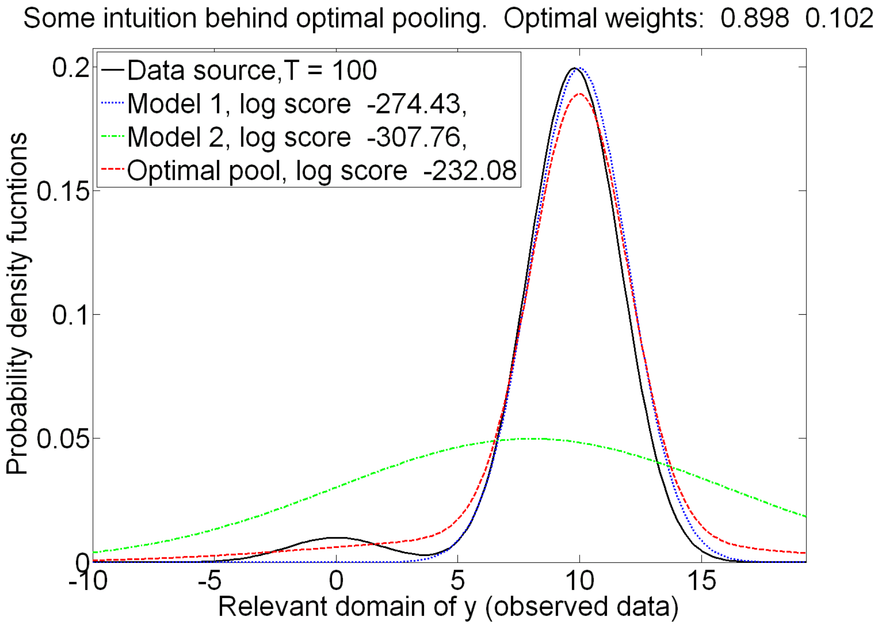

Figure 1: Constructed example illustrating optimal pooling.

pool. That turns out to be the case here, in Section 4, and it is also the case in the example provided in Geweke and Amisano (2010).

The intuition behind optimal pooling under a log scoring rule is similar to that of portfolio optimization under the constraint of no short positions. Model A_1 may have a log score that substantially exceeds that of model A_2, just as one asset may have an average return substantially higher than another. But it may also be the case that from time to time $p_t\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_1\right) \ / \ p_t\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_2\right)$ is small, much closer to zero than one, just as the asset with lower average return may from time to time substantially outperform the other. Given the concavity of the log score function, the optimal pool can (and often does) assign positive weight to both models, just as given risk aversion both assets may have positive weights in an optimal portfolio.

Figure 1 illustrates this situation for optimal pooling: Model 1 closely tracks the data generating process, except for the negative lobe that is reflected in realizations about one observation in twenty. The log score of Model 2 is much lower than that of Model 1, which will be assigned quite negligible posterior probability in a formal Bayesian approach and be rejected in favor of Model 1 in a formal sampling-theoretic test. Yet it receives positive weight in the pool, which has a substantially higher log score than Model 1, because relative to Model 1 the pool provides a very large increase in the log predictive density when realizations from the left lobe occur.

This construction of an optimal prediction pool does not invoke either the weak assumption that there exists a data generating process D generating the observed returns y_t^o, or the much stronger assumption that one of the models (unknown to us) is the data generating process so that $D = A_i$ for some $i = 1, \ldots, n$. Suppose that we grant the weak assumption with D giving rise to the time-invariant probability density function $p_t\left(y_t; Y_{t-1}, X_{t-1}, D\right)$ $(t = 1, 2, \ldots)$. It is well understood that given further weak regularity conditions $t^{-1} \sum_{s=1}^{t} \log p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right)$ tends to an almost sure limit. Geweke and Amisano (2010) shows that under these conditions both the function $t^{-1} f_t\left(\mathbf{w}_t\right)$ and the sequence of optimal weight vectors $\{\mathbf{w}_t^*\}$ have well-defined almost sure pointwise limits. In general several components of the limiting weight vector are positive. An exception is the hypothetical case $D = A_i$, for which w_{ti}^* has limiting value one (Geweke and Amisano, 2010, Theorems 1 and 2). Thus several of the competing models enter the optimal pool even in large samples, and this occurs precisely because all of the models under consideration are false.

3.3 Alternatives to pooling

This behavior contrasts markedly with virtually all conventional econometric procedures that pool the prediction models A_i. The essentials were presented fully for the first time in the econometrics literature by Gourieroux and Monfort (1989). Continuing to invoke the weak assumption, the estimate of the parameter vector $\boldsymbol{\theta}_i$ in model A_i has an almost sure limit that we may denote $\boldsymbol{\theta}_i^*$ and is known as the pseudo-true value of $\boldsymbol{\theta}_i$. In general the pseudo-true value is the same for all likelihood based methods, both Bayesian and non-Bayesian. It follows that $t^{-1} \sum_{s=1}^{t} \log p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right)$ and $t^{-1} \sum_{s=1}^{t} \log p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, \boldsymbol{\theta}_i^*, A_i\right)$ have the same almost sure limit, which we henceforth denote $L\left(A_i, D\right)$

Now introduce the stronger assumption that reality corresponds to one of the models: thus $D = A_i$ for some (unknown) $i = 1, \ldots, n$. Bayesian econometrics, which provides a logically complete theory of model combination, makes this assumption explicitly. It is implicit in non-Bayesian approaches, for example those based on model-fitting criteria like AIC and SBIC. In the Bayesian approach, for any pair of models A_i and A_j,

$$\log\left(\frac{\pi_{ti}}{\pi_{tj}}\right) = \log\left(\frac{\rho_i}{\rho_j}\right) + \sum_{s=1}^{t} \log p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right) - \sum_{s=1}^{t} \log p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_j\right),$$

where ρ_i and ρ_j are the model prior probabilities and π_{ti} and π_{tj} are the model posterior probabilities. It follows that $t^{-1}\log\left(\pi_{ti}/\pi_{tj}\right)$ has the almost sure limit $L\left(A_i, D\right) - L\left(A_j, D\right)$. Unless this limit is zero — a fortuitous case — $\log\left(\pi_{ti}/\pi_{tj}\right)$ tends either to $+\infty$ or $-\infty$ as $t \to \infty$. In general there is one model A_i for which $\log\left(\pi_{ti}/\pi_{tj}\right) \to +\infty$ $(i \neq j)$. In the Bayesian model averaging pool the weight on the predictive densities of model A_i tends to 1 and the weights on the predictive densities of all other models tend to 0 as $t \to \infty$.

The intuition behind this result is straightforward. Granted the assumption that there exists a data generating process corresponding exactly to one of the models A_i, as evidence accumulates that a particular model A_i is superior to all the others, one is driven to the conclusion that $A_i = D$ and predictions should be based on that model alone. That is what happens with Bayesian model averaging, as well as with non-Bayesian procedures working from the same assumptions. Optimal pooling does not make this assumption, to very different effect: the limiting positive weights on several models reflect the accumulated evidence that some models perform well in prediction when others perform poorly, and vice versa, as reflected in increments to their log scores.

The manifestly false assumption that one of the available models is literally true, inherent in conventional econometric procedures, leads to a zero-sum game in which one model must dominate all the others. This "winner take all" implication dominates academic discourse from time to time. The more realistic and humble assumption that all models are false, which underlies optimal pooling, leads to a procedure that trades off the strengths and weaknesses of the models available. It seems to us that this condition characterizes the situation of actual decision-makers, who typically consult several models even in the face of conventional econometric interpretations of the evidence implying that all but one model should be discarded. In the case studied here, it turns out (Section 4) that pooling indeed leads to predictions that are markedly superior to those of any of the 42 models at hand.

4 Results

The application uses S&P 500 Index (SPX) log returns from January 2, 1990 through March 31, 2010. Models based on option-implied volatility use the VIX index. SPX and VIX data were obtained directly from the Chicago Board Options Exchange (CBOE). The high-frequency models use a volatility signal extracted from five-minute intraday SPX returns, obtained from TickData.com.

Since the VIX begins with the first trading day of 1990, estimation samples for all of our models begin with t corresponding to the second trading day of 1990. For each model A_i we evaluate predictive densities $p\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_i\right)$ recursively, beginning with $t = 1$ corresponding to the first trading day of 1992 and ending with $t = T = 4596$ corresponding to March 31, 2010. This requires re-estimation of each model for each t as $\mathbf{Y}_{t-1}^o$ expands. Since there are 4596 days in the recursion and 42 models, the result is a 4596×42 matrix $\mathbf{P}$ of predictive densities. These computations are relatively time consuming.[3] All of our findings derive from $\mathbf{P}$.

4.1 Model performance and comparison

Table 1 provides the (full sample) log predictive score (5) of each model, $LS\left(Y_T^o; X_T^o, A_i\right)$ $(i = 1, \ldots, n)$. For legibility we subtract the log predictive score of the `hifreq_010` model, which is 14,783.55, from the log scores reported here and throughout Section 4. Differences in log scores, not their levels, matter. From (6), the difference $\Delta\left(A_i, A_j\right) = \exp\left\{\left[LS\left(Y_T^o; X_T^o, A_i\right) - LS\left(Y_T^o; X_T^o, A_j\right)\right]/T\right\}$ corresponds to a geometric average proportional difference in predictive densities. For example in the case of `hifreq_122` and `hifreq_010` this difference is $\exp(310.24/4596) = 1.0698$. That is, the predictive densities from model `hifreq_122` render observed events on average almost 7% more probable than do the predictive densities from model `hifreq_010`. More generally, a difference of 45.73 in log scores corresponds to a 1% increment in probability, a difference of 4.59 to a 0.1% increment.

In interpreting the results, it is essential to recall that the log predictive score is an out-of-sample criterion. Unlike in-sample criteria, out-of-sample criteria inherently penalize overfitting. If model A_i is nested in model A_j, the predictive likelihood of

[3] The stochastic volatility models required the most time, about 12 CPU days for the one-factor model and 4 CPU weeks for the 2-factor model. Models with two-factor Kalman filters took about 5 CPU hours. The other models required about 15 minutes on average. In each case the time stated is the total over all 4596 samples.

Daily models						
	`sv_1`	`sv_2`	`egarch_11`	`egarch_12`	`egarch_21`	`egarch_22`
	276.24	297.60	215.77	286.50	256.10	**323.38**
High frequency models (`hifreq_kjp`)						
	$j = 1$			$j = 2$		
	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
$k = 0$	0.00	117.91	145.50	143.01	198.30	219.35
$k = 1$	255.57	254.24	261.27	306.18	303.52	**310.24**
$k = 2$	249.25	250.73	259.73	305.78	299.11	308.45
Options models (`vix_kjp`)						
	$j = 1$			$j = 2$		
	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
$k = 0$	216.19	239.52	235.17	271.21	298.00	297.40
$k = 1$	209.74	239.17	234.90	266.31	**298.08**	297.57
$k = 2$	205.95	234.26	229.75	263.77	294.56	293.78

Table 1: Log scores of models relative to egarch010. Boldface indicates the highest log score in each of the three groups of model. See Section 2 for complete model definitions.

model A_i can exceed that of model A_j; in contrast, the maximized log-likelihood (an in-sample criterion) can never be higher for the nested model. In Table 1 notice that the `vix_121` model is nested in the `vix_222` model and has the higher log score; similarly `hifreq_112` and `hifreq_212`.

As noted in Section 3, had our method of inference been formally Bayesian, then the log scores would coincide with marginalized likelihoods in which the prior distribution for each model includes the 1990-1991 data as a training sample. That is not the case here, but differences in log scores can be regarded as of the same order of magnitude as log ratios of posterior probabilities. For example, given equal prior probablilities for the models, the posterior probability odds ratio in favor of `sv_2` over `sv_1` is on the order of 10^9.

This interpretation reveals the high return to the various elaborations on the daily, high-frequency and options models detailed in Section 2. The roughly 20-point improvement for the stochastic volatility model, resulting entirely from using a mixture of normals rather than Gaussian distribution for ε_t, has just been noted. Returns for other model classes are higher. Among the EGARCH models, `egarch_22` improves over the conventional model, `egarch_11`, by over 100 points with the introduction of a second volatility factor and use of a mixture distribution for the shocks. The improvement is most dramatic for the high-frequency models, where the increase of over 300 points in log score relative to the simplest model is due primarily to the incorporation of a filtration ($k > 0$) that allows current latent volatility to depend flexibly on lagged realized volatilities and secondarily to the use of a mixture distribution for the return shocks ($j = 2$). For the options models the elaborations described in Section 2 lead to an increase of over 80 points in log score, accounted for primarily by the mixture of normals distribution for conditional returns and secondarily by the incorporation of additional flexibility in the link between IV_t and σ_t ($p = 1$ versus $p = 0$).

4.2 Conventional predictive density combination

Arguably the simplest rule for density combinations is the equally-weighted pool A^*, which has $w_{i,t-1} = n^{-1}$ $(t = 1, \ldots, T; i = 1, \ldots, n)$ and log score $LS\left(Y_T^o; X_T^o, A^*\right)$. From Jensen's inequality $LS\left(Y_T^o; X_T^o, A^*\right)$ must exceed the mean log predictive score in Table 1. Indeed it can exceed the maximum of the log predictive scores, and that is what happens here: $LS\left(Y_T^o; X_T^o, A^*\right) = 339.77$.

A modest elaboration on this procedure is first to distribute weight equally on each group of models and then equally across models within each group. Thus in this application each group has weight 1/3, so that each daily model has weight 1/18 and each of the high-frequency and options models has weight 1/54. The log score of the resulting pool is 343.32.

Equally-weighted pools provide useful benchmarks for comparisons with alternative predictive density rules. The idea is similar to the use of the market portfolio or $1/n$ rules as benchmarks for portfolio performance. Many stock pickers believe that they can reliably beat the market. Far fewer succeed. The analogy holds for model selection as well.

The `egarch_22` model dominates the other models available, with a log score over 13 points greater than the next best model. Pursuing the informal Bayesian interpretation of the results discussed above, this suggests posterior probability ratios in favor of `egarch_22` relative to any other model on the order of at least 5×10^5 and in most cases much more. An econometrician using conventional model selection rules would place all, or nearly all, weight on this model to the exclusion of all alternatives. But even the simplest equally-weighted pool beats `egarch_22` by nearly 16 points in log score.

The reality for the model picker is even worse than this. Here, we have assumed a prescient model picker who is able to choose the individual model that performs best over the entire sample. In practice, the model picker must choose the best model in real time using available information.

Bayesian model averaging (BMA) is often put forward as an appealing approach to model combination. The prescient model averager, putting weight on each model in proportion to its full-sample log score, would put essentially all weight on `egarch_22`, yielding results that are virtually indistinguishable from the econometrician that is forced to pick a single model. But it is instructive to consider the idea of constructing real-time pools using BMA in order to examine the implications for choices amongst the 42 individual models and for contrasting these implications with optimal pooling subsequently.

Identifying $p\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_i\right)$ with the Bayesian predictive likelihood, the analogue of marginal likelihood for model A_i based on the sample from periods 1 through t is

$$ML_{it} = \prod_{s=1}^{t} p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right) = \exp\left[LS\left(Y_t^o; X_t^o, A_i\right)\right].$$

Given equal model prior probabilities, the posterior probability of model i based on this sample is $\omega_{it} = ML_{it}/\sum_{j=1}^{n} ML_{jt}$. Under the Bayesian model averaging paradigm, the predictive density for y_{t+1} is

$$p\left(y_{t+1}; Y_t^o, X_t^o, B^*\right) = \sum_{i=1}^{n} \omega_{it} p\left(y_{t+1}; Y_t^o, X_t^o, A_i\right). \tag{10}$$

The procedure just described constitutes a valid prediction model, which we denote B^* in (10). Its log predictive score $LS\left(Y_T^o; X_T^o, B^*\right)$ can be evaluated directly using the 4596×42 matrix $\mathbf{P}$ of predictive likelihoods described at the beginning of this section.

Two features of this model averaging exercise are important for this study. First, consider the weights ω_{it}. Rather than report weights for all of the models individually, at each time period t we sum the weights within each of the three groups of models (daily, high-frequency and options). These are displayed in Figure 2. For most of the sample the preponderance of the weight is on the daily model group. There are reversals within the sample, as well: for example early in 1995 weight is almost

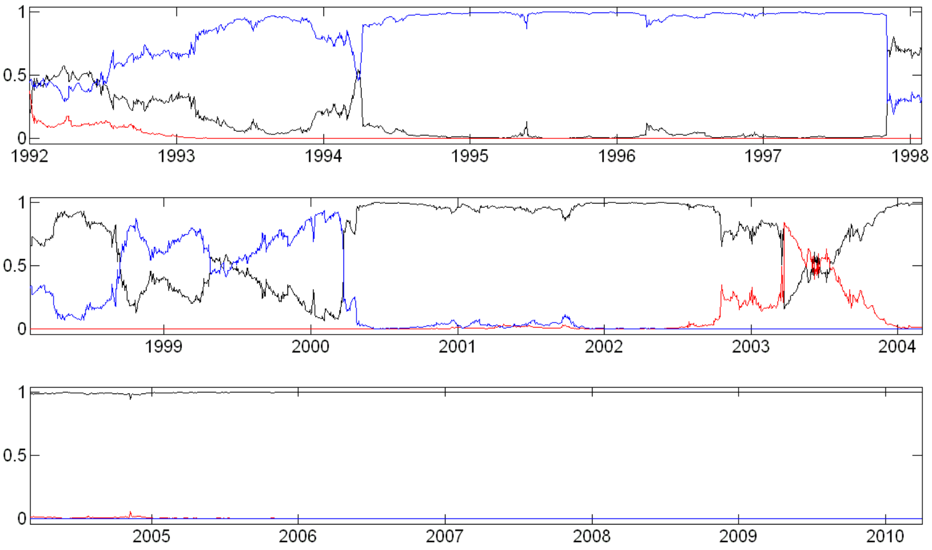

Figure 2: Bayesian model averaging weights, updated each trading day: sum of weights for daily models (black), high frequency models (red), and options models (blue).

entirely concentrated on the options group of models, corresponding to near certainty (*a posteriori*) that none of the daily or high-frequency models are true. On the other hand from 2005 onward the weights concentrate virtually entirely on the daily group of models, corresponding to near certainty that none of the high-frequency or options models are true. The latter behavior is typical of Bayesian model averaging, noted nearly two decades ago in Diebold (1991). This vacillation between near-certainties, in a procedure that starts with the premise that one of the models corresponds to the data generating process, challenges the credibility of the premise.

Second, consider the log scores. The log score of the Bayesian model averaging prediction rule is $LS\left(Y_T^o; X_T^o, B^*\right) = 316.37$. This is slightly better than the modeler who places all weight on a single model in real time, though slightly worse than the prescient model picker or model averager. But any of these fall well short of either benchmark equally-weighted pool. Thus the performance of this procedure motivated by Bayesian model averaging is poor just as its premise is not credible.

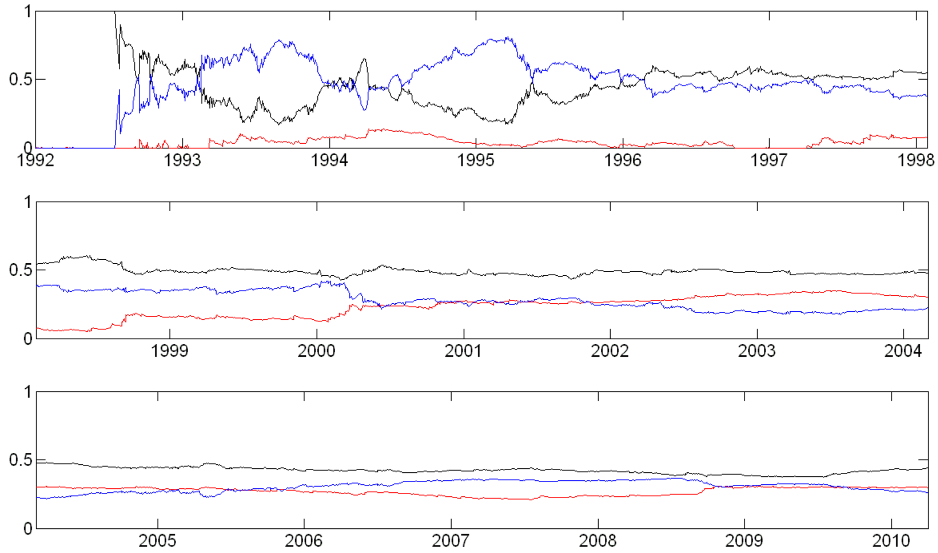

Figure 3: Optimal prediction pool weights, updated each trading day: sum of weights for daily models (black), high frequency models (red), and options models (blue).

4.3 Optimal pooling

The optimal pooling procedure implemented here reconstructs what an econometrician could have accomplished in real time. For each date t beginning with $t = 1$, which indicates the first trading day of 1992, and ending with $t = 4596$ (March 31, 2010) suppose that the econometrician has at her disposal predictive densities $p_s(y_s; Y^o_{s-1}, X^0_{s-1}, A_i)$ $(s = 1, \dots, t; i = 1, \dots, n)$ and has evaluated these densities using realized returns, $p_s(y^o_s; Y^o_{s-1}, X^0_{s-1}, A_i)$. Thus, on day t, the optimizer is using the first t rows of the 4596×42 matrix $\mathbf{P}$. Using this information, she finds the optimal pooling weights $\mathbf{w}^*_t = \arg\max_{\mathbf{w}_t} f_t(\mathbf{w}_t)$ where $f_t(\mathbf{w}_t)$ is defined in (9).

Figure 3 displays the optimal pool weights w^*_{it} in the same way that Figure 2 did for the Bayesian model averaging weights. Initially the optimal pool consists entirely of daily models. Options models enter the optimal pool mid-way through the first year and high-frequency models enter later that year. The gradual entry of models at the start of the exercise is characteristic of optimal prediction pools: notice from the calculus of optimization of a concave function on the unit simplex in (7)-(8) that at most t models will have positive weight in an optimal pool when $t < n$. As the number of predictions over which the optimal pool weights are evaluated continues to increase, optimal weights stabilize. From midway through the exercise (2001) forward the distribution of weights across the groups of models does not change substantially.

At the end of the exercise, which is the close of trading on March 31, 2010, the total weight on the group of daily models is 0.4435, all arising from the `egarch_22` model. The total weight on the high-frequency models is 0.2974, comprised of the sum of the weights on `highfreq_020` (0.0349), `highfreq_110` (0.1790), `highfreq_112` (0.0106), and `highfreq_122` (0.0729). The options models garner the remaining weight of 0.2591, allocated among `vix_022` (0.0470), `vix_111` (0.1075) and `vix_121` (0.1046). Consulting Table 1, note that the eight models with positive weights include those with the largest log score in each group, and those three weights sum to 0.621. On the other hand fully half the models (21) have log scores exceeding that of `vix_111` but have no weight in the optimal pool

Whether or not a model enters the pool with positive weight depends on its record in providing a higher density to observed returns when other models with positive weights provide lower densities. These conditions are analogous to those that prevail when an asset enters a portfolio under a constraint of no short positions, and arise for essentially the same reason. The pooling rule places a premium on diversity of models, even if some of those included have relatively low scores. For the high-frequency and options models, the number of components in the mixture distributions appears to be key. Consulting Table 1 once again, there is at least one representative from $j = 1$ and $j = 2$ in both cases, and the representatives include the model with the highest log score in each case, with the minor exception of `vix_111` whose log score is 0.35 points lower than `vix_011`.

Having computed the optimal weight vector $\mathbf{w}_t^*$ at the end of trading day t, based on rows 1 through t of $\mathbf{P} = [p_{ti}]$, our hypothetical econometrician uses the optimal pool as the predictive density for y_{t+1}. Evaluating this density at the realized return y_{t+1}^o provides the log score

$$\sum_{t=0}^{T-1} \log \left[\sum_{i=1}^{n} w_{it}^* p\left(y_{t+1}^o; Y_t^o, X_t^o\right) \right] = \sum_{t=1}^{T} \log \left(\sum_{i=1}^{n} w_{i,t-1}^* p_{ti} \right), \tag{11}$$

which may be compared directly with the entries in Table 1. The log score of the optimal pool is 346.48, about 23 points higher than the best of the constituent models, `egarch_22`. The improvement is even greater relative to the either the BMA pooling

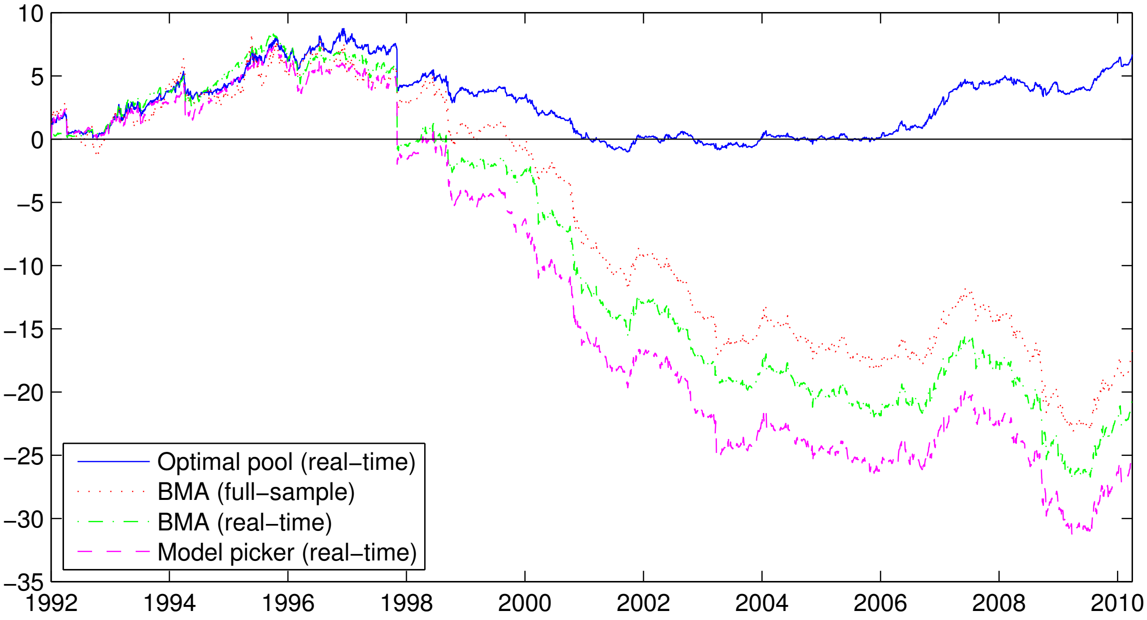

Figure 4: Log scores, differences relative to equally-weighted pool.

rule or the econometrician forced to place all weight on a single model using real-time information. It also exceeds the two equally-weighted benchmarks described in Section 4.2.

Figure 4 shows log scores relative to the equally-weighted pool at each date t in the sample period for the optimal pool, BMA pool using weights computed using the full-sample information (prescient modeler), BMA pool using real-time weights, and the pool comprised of the single model chosen in real-time by the model picker. Whereas the conventional model averager and model picker both substantially underperform the equally-weighted benchmark, the optimal pool outperforms it.

The sums of model weights across groups exhibited in Figure 3 provide one indication of the contribution of each group to the optimal pool. An indication more directly related to performance can be constructed as follows. First evaluate the real-time log score (11) at the end of each period t, yielding the sequence of real-time log scores

$$\lambda_t = \sum_{s=0}^{t-1} \log \left[\sum_{i=1}^{n} w_{is}^{*} p \left(y_{s+1}^{o}; X_{s}^{o}, Y_{s}^{o} \right) \right] = \sum_{s=1}^{t} \log \left(\sum_{i=1}^{n} w_{i,s-1}^{*} p_{si} \right) \quad (t = 1, \ldots, T).$$

Now repeat the optimization exercise, but omitting all of the daily models, and denote

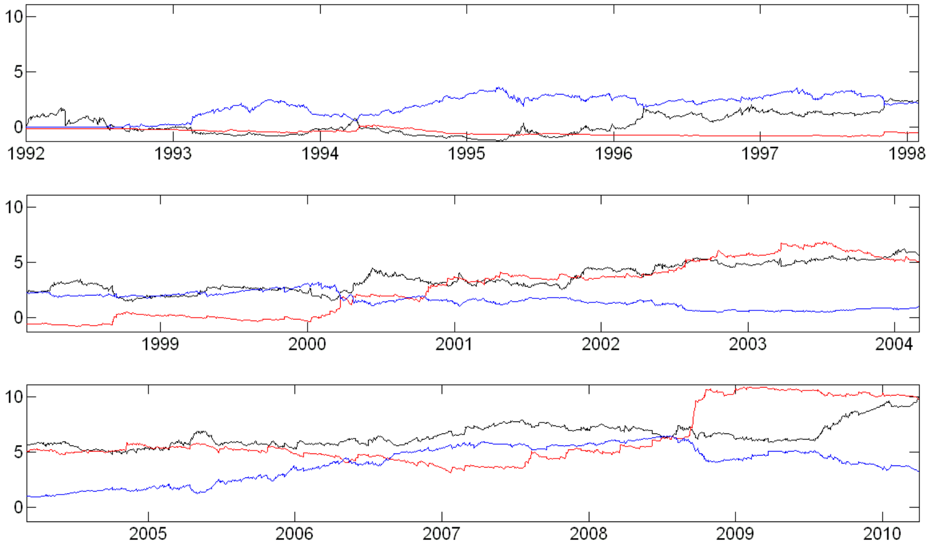

Figure 5: Values of the group of daily models (black), high frequency models (red), and options models (blue).

the resulting sequence of log scores $\left\{\lambda_t^{(1)}\right\}$. Because of the real-time nature of the exercise it is not necessarily the case that $\lambda_t^{(1)} \leq \lambda_t$, and both prior considerations and the weights displayed in Figure 3 suggest that this condition is more likely to be violated for smaller than for larger t. We refer to $\lambda_t - \lambda_t^{(1)}$ as the value of the daily model group at time t. Similarly form the sequence of values $\left\{\lambda_t - \lambda_t^{(2)}\right\}$ for the group of high-frequency models and $\left\{\lambda_t - \lambda_t^{(3)}\right\}$ for the group of options models. Unlike sums of weights within groups, group values will tend to drift with time. For any group with a limiting positive sum of weights, the drift will be upward.

Figure 5 shows the group values constructed in this way. The value of the group of daily models is nonnegative throughout the sample period. The options models exhibit some negative values in the first four years and are positive thereafter. The high-frequency models have negative or negligibly positive value until early 2000, and remain positive thereafter. The latter model group increases in value in striking fashion during September and October, 2008, the height of the global financial crisis, and remains the most valuable group until early 2010. At the end of the sample period (March 31, 2010) the value of the daily model group is 9.976, the high-frequency model group 9.721, and the options group 3.207.

5 Conclusion

This study took up the practical problem of constructing predictive densities for S&P 500 returns from a collection of models, all of which are false. The constituents of the collection were chosen with respect to alternative information sets for predictions of future volatility: daily returns, observed intraday volatility, and the VIX index obtained from options prices. The metric of evaluation was the log scoring rule, equivalent to the geometric average probability assigned to observed returns. This and all comparisons made in the study are strictly out of sample, arising from real-time procedures that could have been employed in prediction at the start of each trading day from January 2, 1992, through March 31, 2010.

Beginning with conventional base models within each of the three groups, we took several steps to improve predictions: replacing conditional Gaussian distributions with normal mixture distributions provided predictive distributions with more credible shapes; including multiple volatility factors provided increased flexibility in how the history of realized returns impacted estimates of the latent volatility state; and in the case of the high frequency and options models we used a flexible mapping from the extracted volatility state to spot volatility (the scaling factor for daily returns). This led to two stochastic volatility (SV) models, four EGARCH models, 18 high frequency models and 18 options prices, for a total of 42 models.

Quantitatively this was the most important step in improving predictive densities for the S&P 500 return series from 1992 through the first quarter of 2010, as indicated in Table 2. The *Improved model* column compares the base model in each group (e.g. `highfreq_010`) with the best model in each group (e.g. `highfreq_122`) using the entries from Table 1 and the metric shown in (6). As discussed in Section 4.1 differences across model groups arise more from disparity among base models than among the best models in each group.

Next we considered pools of all 42 models. The simple step of forming an equally-weighted pool of models led to the improvements in the *Equal weight* pool column of Table 2. Since the pool is the same for all model groups, differences across model groups in this column are due entirely to differences in the log predictive scores of the best model in each group. If it were not the case that all models are false—that is, some one of the 42 models in our collection corresponded to the data generating process for returns—then the expected incremental change in this column would be

	Incremental changes in prediction probability (percent)			
Model group	Improved model	Equal weight pool	Optimal pool	Total
Daily (SV)	0.466	0.916	0.077	1.470
Daily (EGARCH)	2.369	0.357	0.077	2.814
High frequency	6.983	0.645	0.077	7.756
Options prices	1.798	0.911	0.077	2.725

Table 2: Improvements in the geometric mean average probability assigned to observed returns, moving from left to right in each row.

negative for the group containing the true model. That is far from the case. The optimal pool provides further increases in prediction probability.

Conventional econometric model combination procedures, most highly developed in the Bayesian literature, work from the condition that one of the models is true. As an alternative to optimal pooling we examined Bayesian model averaging (BMA). Whereas optimal pools lead to stable positive weights on all three groups of models, BMA weights tend to eliminate all models but one. Furthermore, the model so identified as being almost certainly true changes from time to time over the sample period. The log score of the BMA pool was lower even than that of the simple equally-weighted pool. Prediction probabilities were on average 0.588% lower for the BMA pool than for the optimal pool. The poor performance of BMA complements the incredibility of the assumption that truth resides somewhere in the collection of models.

All dimensions of the study bear out the importance of the fact that no matter what the collection of models, they are all false. Therefore improved models exist, and in this study improvement of individual models yielded the greatest returns. But even with a set of improved models, the fact that all still remain false indicates a further improvement from model pooling (Geweke and Amisano, 2010, Theorems 1 and 2). That potential was borne out in this study. This latter improvement significantly recasts model comparison from a horse race in which there is typically little role for any but the winning model to a more cooperative situation in which many models have relative strengths and weaknesses leading to important roles for several models in improving predictive performance. In this setting an optimal pool bears strong resemblance to optimal portfolio allocation with a restriction of no short positions and the familiar gains from diversification in that setting.

Our study addressed one-step-ahead predictions of a single return, the S&P 500 index return, which in turn is the most thoroughly addressed prediction problem in financial econometrics. In contrast the most important prediction problems involve multiple returns and prediction horizons of several steps. The fundamental principles in this work, log scoring and optimal pooling, apply directly to these extensions. The case of multiple returns is straightforward, e.g. O'Doherty et al. (2010). Moreover for multivariate prediction there are compelling axiomatic arguments requiring pools to be linear (McConway, 1981) as they were in this study. Predicting several steps into the future is more demanding to the extent that covariates (in thus study, X_{t-1}, the indicators of volatility in the high frequency and options models) must also be predicted. In econometric terms these covariates are then no longer exogenous but instead must themselves be modelled. There are no fundamental difficulties, here, just the significant work of creating and improving credible models. We plan to address these issues in future research.

References

Andersen TG, Bollerslev T, Diebold FX, Labys P (2001a). The distribution of realized exchange rate volatility. Journal of the American Statistical Association 96: 42-55.

Andersen TG, Bollerslev T, Diebold FX, Ebens P (2001b). The distribution of realized stock return volatility. Journal of Financial Economics 61: 43-76.

Andersen TG, Bollerslev T, Diebold FX, Labys P (2003). Modeling and forecasting realized volatility. Econometrica 71: 579-625.

Barndorff-Nielsen OE, Shephard N (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society Series B 64: 253-280.

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

Bolleslev, T, Mikkelsen H (1996). Modelling and pricing long-memory in stock market volatility. Journal of Econometrics 73: 151-184.

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

Britten-Jones M, Neuberger A (2000). Option prices, implied price processes, and stochastic volatility. Journal of Finance 55: 839-866.

Corsi F (2009). A simple long-memory model of realized volatility. Journal of Financial Econometrics 7: 174-196.

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

Diebold FX (1991). On Bayesian forecast combination procedures. In Westlund A, Hackl P (eds.) Economic Structural Change: Analysis and Forecasting, Chapter 15, 225-232. New York: Springer-Verlag.

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

Durham G (2006). Monte Carlo methods for estimating, smoothing and filtering one- and two-factor stochastic volatility models. Journal of Econometrics 133: 273-305.

Durham G (2007). SV mixture models with application to S&P 500 index returns. Journal of Financial Economics 85: 822-856.

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

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

Geweke J, Amisano G (2010). Optimal prediction pools. Journal of Econometrics, forthcoming.

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

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

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.

Jiang GJ, Tian YS (2005). The model-free implied volatility and its information content. Review of Financial Studies 18: 1305-1342.

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

McLachlan G, Peel D (2000). Finite Mixture Models. New York: Wiley.

O'Doherty MS, Savin NE, Tiwari A (2010). Modeling the Cross Section of Stock Returns: A Model Pooling Approach.

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1536050.

Richard JF, Liesenfeld R (2006). Classical and Bayesian analysis of univariate and multivariate stochastic volatility models. Econometric Reviews 25: 335-360.

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.

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.