Predictive gains from forecast combinations using time varying model weights *

Francesco Ravazzolo — *Econometric Institute and Norges Bank*

Herman K. van Dijk — *Econometric and Tinbergen Institute Rotterdam*

Marno Verbeek — *RSM Erasmus University*

ECONOMETRIC INSTITUTE REPORT 2007-26

August 17, 2007

Abstract

Several frequentist and Bayesian model averaging schemes, including a new one that simultaneously allows for parameter uncertainty, model uncertainty and time varying model weights, are compared in terms of forecast accuracy over a set of simulation experiments. Artificial data are generated, characterized by low predictability, structural instability, and fat tails, which is typical for many financial-economic time series. Sensitivity of results with respect to misspecification of the number of included predictors and the number of included models is explored. Given the set up of our experiments, time varying model weight schemes outperform other averaging schemes in terms of predictive gains both when the correlation among individual forecasts is low and the underlying data generating process is subject to structural locations shifts. In an empirical application using returns on the S&P 500 index, time varying model weights provide improved forecasts with substantial economic gains in an investment strategy including transaction costs.

Keywords: Stock return predictability, time varying weight combination, forecast combination, Bayesian model averaging
JEL Classification: C11, C53, G11

*We thank seminar participants at the Conference of the 50-th Anniversary of the Econometric Institute 2006; the Nake Research Day 2006; and the Conference on Computational Economics and Finance, Geneva, 2007, for helpful comments on a preliminary version of this paper. Any errors are the authors responsibility.
E-mail addresses: ravazzolo@few.eur.nl, corresponding author (Francesco Ravazzolo), hkvandijk@few.eur.nl (Herman K. van Dijk), mverbeek@rsm.nl (Marno Verbeek).

1 Introduction

When a wide set of forecasts of some future economic event is available, decision makers usually attempt to discover which is the best forecast, then accept this and discard the other ones. However, the discarded forecasts may have some independent valuable information and including them in the forecasting process may provide more accurate results. An important explanation is related to the fundamental assumption that in most cases one cannot identify *a priori* the exact true economic process or the forecasting model that generates smaller forecast errors than its competitors. An alternative reasonable assumption appears to be one where different models may play a - possibly temporary - complementary role in approximating the data generating process. Furthermore, perhaps due to the presence of private information such as forecasters' subjective judgements or differences in modelling approaches, it may not be possible to pool the underlying information sets and construct a 'super' model that nests each of the underlying forecasting models. In these situations, forecast combinations are viewed as a simple and effective way to obtain improvements in forecast accuracy.

Since the seminal article of Bates and Granger (1969) several papers have shown that combinations of forecasts can outperform individual forecasts in terms of symmetric loss functions. For example, Stock and Watson (2004) find that forecast combinations to predict output growth in seven countries generally perform better than forecasts based on single models. Marcellino (2004) has extended this analysis to a large European data set with broadly the same conclusion. However, several alternative combination schemes are available and it is not clear which is the best scheme, either in a classical or a Bayesian framework. For example, Hendry and Clements (2004) and Timmermann (2006) show that simple combinations[1] often give better performance than more sophisticated approaches. Further, using a frequentist approach, Granger and Ramanathan (1984) propose the use of coefficient regression methods, Hansen (2007) introduces a Mallows' criterion, which can be minimized to select the empirical model weights, and Terui and van Dijk (2002) generalize the least squares model weights by reformulating the linear regression model as a state space specification where the weights are assumed to follow a random walk process. Guidolin and Timmermann (2007) propose a different time varying weight combination scheme where

[1] In this paper simple combinations are defined as combinations that do not require estimating parameters; arithmetic averages constitute a simple example. Complex combinations are defined as combinations that rely on estimating weights that depend on the full variance-covariance matrix and, possibly, allow for time varying weights.

weights have regime switching dynamics. Stock and Watson (2004) and Timmermann (2006) use the inverse mean square prediction error (MSPE) over a set of the most recent observations to compute model weights. In a Bayesian framework, Madigan and Raftery (1994) revitalize the concept of Bayesian model averaging (BMA) and apply it in an empirical application dealing with Occam's Window. Recent applications suggest its relevance for macroeconomics (Fernández *et al.*, 2001 and Sala-i-Martin *et al.*, 2004). Strachan and van Dijk (2007) compute impulse response paths and effects of policy measures using BMA in the context of a large set of vector autoregressive models. Geweke and Whiteman (2006) apply BMA using predictive and not marginal likelihoods.

This paper contributes to the research on forecast combinations by investigating the relative merits of eight combination schemes in simulation exercises where the data generating process is subject to low predictability, structural instability, in the sense that the relevance of forecasting factors varies over time, and fat tails. Sensitivity of results with respect to misspecification of the number of included predictors and the number of included models is explored.

The different combination schemes are summarized as two simple schemes, which do not require parameter estimates; two schemes that involve OLS weight regressions, and a more advanced time varying weight scheme due to Terui and van Dijk (2002). Next, we include two Bayesian model averaging schemes: the original one first proposed in an empirical application by Madigan and Raftery (1994), and a more recent one in terms of predictive densities given by Geweke and Whiteman (2006)[2]. Finally, we propose a new Bayesian scheme which allows for parameter uncertainty, model uncertainty and time varying model weights simultaneously.

As in Aiolfi and Timmermann (2006) we use an adequate long out-of-sample period to evaluate the forecasting performance of the different combination schemes.
Our results indicate that when correlation among forecasts of individual models is low, simple and Bayesian averaging strategies using marginal likelihoods perform poorly, while unconstrained OLS and time varying model weight schemes provide more accurate results. Moreover, when structural instability is high, we explain asymptotically and in a simulation experiment that the time varying combination schemes give the most accurate forecasts.

A second contribution of this paper is to provide an empirical illustration, where we consider forecasting the returns on the S&P 500 index by combining individual forecasts from two competing models. The first one assumes that a set of financial and macroeconomic

[2]Alternative BMA's exist such as MC^3, or frequentist approaches that share similar features as BACE or thick modelling; but we omit them to simplify the analysis.

variables have explanatory power, the second one is based on the popular market saying "Sell in May and go away", also known as the "Halloween indicator", see for example Bouman and Jacobsen (2002). Low predictability of stock market return data is well documented, see for example Marquering and Verbeek (2004) and so is structural instability in this context, see for example Pesaran and Timmermann (2002) and Ravazzolo *et al.* (2007). We confirm these results, and show that the two models, taken individually, perform poorly and in a differential way over time. We continue by applying model averaging and find that the two time varying weight schemes that we apply give the best forecasts in term of symmetric loss functions, confirming the results of the simulation exercises. Moreover, as an investor is more interested in the economic value of a forecasting model than in its forecast accuracy, we test our findings in an active short-term investment exercise, with an investment horizon of one month. Again, the time-varying weight schemes provide the highest economic gains.

The contents of this paper are organized as follows. In Section 2 we describe the eight different forecast combination schemes. In Section 3 we report results from simulation exercises in predicting future values. In Section 4 we give results from an empirical application to US stock returns and show that forecast combinations give economic gains. Section 5 concludes. In the Appendices some technical details are presented.

2 Forecast combination schemes

Two schemes are based on simple constant weights; three are frequentist approaches based on estimated (time varying) model weights; two are "known" Bayesian averaging schemes, the final one is a new Bayesian averaging scheme that allows for time varying weights. We note that the vast majority of studies on forecast combination deals with point forecasts, and we also focus on this.

We start with a brief description of the basic set up of the simulation experiments. Suppose two time series $y_1 = \{y_{s,1}\}_{s=1}^{S}$ and $y_2 = \{y_{s,2}\}_{s=1}^{S}$ are generated from the following models:

$$y_{s,1} = \alpha_1 + x_{s,1}^{'}\beta_1 + \epsilon_{s,1} \tag{1}$$

$$y_{s,2} = \alpha_2 + x_{s,2}^{'}\beta_2 + \epsilon_{s,2} \tag{2}$$

where $x_{s,1}$ and $x_{s,2}$ are $(k_1 \times 1)$ and $(k_2 \times 1)$ vectors of predictor variables respectively, where α_1, α_2 are two scalar parameters and β_1, β_2 are $(k_1 \times 1)$ and $(k_2 \times 1)$ vectors of parameters, and where $\epsilon_{s,1}$ and $\epsilon_{s,2}$, $s = 1, ..., S$, are two zero mean i.i.d. disturbances with variances σ_1^2

and σ_2^2, respectively. The simulated data generating process (DGP) is a linear combination of the previous two models:

$$y_s = y_{s,1}c_{s,1} + y_{s,2}c_{s,2}, \tag{3}$$

where $c_{s,1}$ and $c_{s,2}$ are two possibly time varying scalars. We refer to $c_{s,1}$ and $c_{s,2}$ as DGP weights.

Equations (1) and (2) are estimated over the sample period $[1, ..., T]$ with $T < S$ to compute two independent one-step ahead forecasts $\widehat{y}_{T+1,1}$ and $\widehat{y}_{T+1,2}$, combined to compute a forecast of y_{T+1}. We let $\widehat{y}_{T+1} = g(\widehat{y}_{T+1,1}, \widehat{y}_{T+1,2}, w_{T+1})$ be the combined point forecast as a function of the underlying single forecasts $\widehat{y}_{T+1,1}$ and $\widehat{y}_{T+1,2}$, the forecast combination scheme g, and the vector of the parameters of the combination w_{T+1}[3]. The values of the optimal combination $\widehat{w}_{T+1}$ solve the problem:

$$\min_{w_{T+1}} E[L(e_{T+1}(w_{T+1}))|\widehat{y}_{T+1,1}, \widehat{y}_{T+1,2}], \tag{4}$$

where $e_{T+1} = y_{T+1} - g(\widehat{y}_{T+1,1}, \widehat{y}_{T+1,2}, \widehat{w}_{T+1})$ is the forecast error from the combination, and where L is the loss function, which for simplicity we assume to depend only on the forecast error. We emphasize that the vector $\widehat{w}_{T+1}$ is not necessarily an estimate of the vector $[c_{T+1,1}, c_{T+1,2}]'$, but refers to estimated weights that minimize the loss function. The general class of combination schemes in (4) comprises non-linear as well as time-varying methods.

In most cases there is no closed form solution of equation (4), but analytical results may be computed imposing restrictions on the loss function and making distributional restrictions on the forecast errors. Often it is simply assumed that the objective function is the mean squared error (MSE) loss function:

$$L(e_{T+1}(w_{T+1})) = \theta(\widehat{y}_{T+1} - y_{T+1})^2 \quad \theta > 0. \tag{5}$$

For this case the combined forecasts choose a combination of the individual forecasts that best approximates the conditional expectation, $E(y_{T+1}|\widehat{y}_{T+1})$. In the five frequentist approaches that we apply we assume the MSE loss function and we fix $\theta = 1$. Different distributional restrictions, for example assuming a time varying θ imply different estimation techniques in equation (4).

As a next step we expand the sample period with the observation y_{T+1} and we compute new individual and combination forecasts for the value y_{T+2}. We repeat the procedure to compute H forecasts where $T + H = S$.

[3]Note that w_{T+1} may also be a vector of constants.

2.1 Simple combination schemes

Following Timmermann (2006) we define simple combination schemes as cases that do no require estimating (many) parameters, in particular do not require estimating the full variance-covariance matrix. Moreover, these schemes are distinguished by the restriction that the weight coefficients add up to unity.

The forecasts on y_{T+1} given by simple combination schemes can be written as:

$$\widehat{y}_{T+1}^{(j)} = \widehat{y}_{T+1,1}\widehat{w}_{T+1,1}^{(j)} + \widehat{y}_{T+1,2}\widehat{w}_{T+1,2}^{(j)}, \tag{6}$$

where $(\widehat{w}_{T+1,1}^{(j)}, \widehat{w}_{T+1,2}^{(j)})$, $j = 1, 2$, are computed following schemes 1 and 2 below.

Scheme 1: Equal weights

$$\widehat{w}_i^{(1)} = 1/n \tag{7}$$

where $i = 1, 2$. Extension to more general case with n individual models is straightforward. Equal weights are optimal in situations when the individual forecast errors have the same variance and identical pair-wise correlations, see Timmermann (2006).

Scheme 2: Inverse Mean Square Prediction Error (MSPE) weights

Scheme 2 derives weights from the models' relative inverse MSPE performances computed over a window of the previous υ periods, see Timmermann (2006). Estimation errors in combination weights tend to be particularly large due to the difficulties in precisely estimating the covariance matrix of the forecast error. One answer to this problem is to ignore correlation across forecast errors and making combination weights that reflect performance of each individual model relative to the performance of the average model. The MSPE at time T over the previous υ forecasts for model $i = 1, 2$ is defined as:

$$MSPE_{T,i}^{\upsilon} = \frac{\sum_{j=0}^{\upsilon-1}(\widehat{y}_{T-j,i} - y_{T-j})^2}{\upsilon} \tag{8}$$

The weights are computed as:

$$\widehat{w}_{T+1,i}^{(2)} = \frac{(1/MSPE_{T,i}^{\upsilon})}{\sum_{j=1}^{2}(1/MSPE_{T,j}^{\upsilon})} \tag{9}$$

2.2 Estimated weight combination schemes

The next three combination schemes estimate the weights in regression form, add a constant term, and do not impose that the weights add to 1.

Scheme 3: Constant OLS weights

The weights are equal to the OLS estimators of the weights (w_0, w_1, w_2) in equation:

$$y_t = w_0 + \widehat{y}_{t,1} w_1 + \widehat{y}_{t,2} w_2 + u_t; \quad u_t \sim N(0, s^2) \tag{10}$$

where $t = 1, .., T$, and w_0 is a constant term[4]. The estimation of the weights, while attractive in the sense of minimizing forecast errors, introduces parameter estimation errors. Therefore, one may estimate weights for the first forecast and then fix these as constant over the remaining out-of-sample period.

The forecast on y_{T+1} given by the estimated combination scheme is given as:

$$\widehat{y}^{(3)}_{T+1} = \widehat{w}^{(3)}_0 + \widehat{y}_{T+1,1} \widehat{w}^{(3)}_1 + \widehat{y}_{T+1,2} \widehat{w}^{(3)}_2 \tag{11}$$

where $(\widehat{w}^{(3)}_0, \widehat{w}^{(3)}_1, \widehat{w}^{(3)}_2)$ are the OLS estimates of the parameters (w_0, w_1, w_2) in (10). To compute the following $H - 1$ forecasts, the same estimated weights $(\widehat{w}^{(3)}_0, \widehat{w}^{(3)}_1, \widehat{w}^{(3)}_2)$ are applied.

Scheme 4: Recursive OLS weights

The estimated weights are equal to the recursive OLS estimators of the weights in (10). The estimated weights are updated every time when a new observation becomes available.

Scheme 5: Time varying weights

When the conditional distribution of $(y_{T+1}, \widehat{y}_{T+1})$ varies over time, it may be effective to let the combination weights also change over time. Terui and van Dijk (2002) have proposed a method that extends the OLS weight combination. The weights satisfy the following recursions:

$$y_t = w_{t,0} + \widehat{y}_{t,1} w_{t,1} + \widehat{y}_{t,2} w_{t,2} + u_t; \quad u_t \sim N(0, s^2) \tag{12}$$

$$w_t = w_{t-1} + \xi_t; \quad \xi_t \sim N(0, \Sigma) \tag{13}$$

where $w_t = [w_{t,0}, w_{t,1}, w_{t,2}]'$; $t = 1, .., T$; and Σ is a diagonal matrix. The weights are time varying and follow a random walk process. The time varying weight combination may be interpreted as a state space model, where (12) is the measurement equation which defines the distribution of y_t, and where (13) is the state equation which defines the distribution of the weights for every t. The Kalman filter algorithm can be applied to compute the estimators

[4]Granger and Ramanathan (1984) explain that the constant term must be added to avoid biased forecasts. They also conclude that this strategy is often more accurate than restricted OLS weights.

$\widehat{w}^{(5)}_{t|t-1}$. Appendix A gives details of the computation and explains the difference with the recursive OLS estimator.
The forecasts on y_{T+1} given by schemes 4 and 5 are:

$$\widehat{y}^{(j)}_{T+1} = \widehat{w}^{(j)}_{T+1,0} + \widehat{y}_{T+1,1}\widehat{w}^{(j)}_{T+1,1} + \widehat{y}_{T+1,2}\widehat{w}^{(j)}_{T+1,2} \tag{14}$$

where $j = 4, 5$.

2.3 Bayesian model averaging

Bayesian approaches have been widely used to construct forecast combinations, see for example Leamer (1978), Hodges (1987), Draper (1995), Min and Zellner (1993), and Strachan and van Dijk (2007). In this approach one does not estimate regression weights and uses those to compute forecasts, but one derives the posterior probability for any individual model and combines these. The predictive density accounts then for model uncertainty by averaging over the probabilities of individual models. Since the output is a complete density, point prediction (for example by taking the mean), distribution and quantile forecasts can be easily derived.
We choose three BMA schemes: the original one proposed in an empirical application by Madigan and Raftery (1994), a more recent one discussed in Geweke and Whiteman (2006), and a new one to be introduced below.

Scheme 6: BMA using marginal likelihood
The predictive density of y_{T+1} given the data up to time T, F_T, is computed by averaging over the conditional predictive densities given the individual models with the posterior probabilities of these models as weights:

$$p(y_{T+1}|F_T) = \sum_{i=1}^{n} P(m_i|F_T)p(y_{T+1}|F_T, m_i) \tag{15}$$

where n is the number of individual models; $p(y_{T+1}|F_T, m_i)$ is the conditional predictive density given F_T and model m_i; $P(m_i|F_T)$ is the posterior probability for model m_i. The conditional predictive density given F_T and model m_i is defined as:

$$p(y_{T+1}|F_T, m_i) = \int p(y_{T+1}|\theta_i, F_T, m_i)p(\theta_i|F_T, m_i)d\theta_i \tag{16}$$

where $p(y_{T+1}|\theta_i, F_T, m_i)$ is the conditional predictive density of y_{T+1} given F_T, the model parameters $\theta_i = (\alpha_i, \beta_i, \sigma_i^2)'$, and model m_i in (1) or (2); $p(\theta_i|F_T, m_i)$ is the posterior density

for parameter θ_i. The posterior probability for model m_i is:

$$P(m_i|F_T) = \frac{p(y|m_i)p(m_i)}{\sum_{j=1}^{n} p(y|m_i)p(m_j)} \tag{17}$$

where $y = \{y_t\}_{t=1}^{T}$; $p(m_i)$ is the prior density for model m_i; and $p(y|m_i)$ is the marginal likelihood for model (m_i) given by:

$$p(y|m_i) = \int p(\theta_i|F_T, m_i)p(\theta_i)d\theta_i, \tag{18}$$

$p(\theta_i)$ is the prior density for the parameter θ_i. The integral in equation (18) can be evaluated analytically in the case of linear models, but not for more complex forms. Chib (1995), for example, has derived a method to compute the expression also for nonlinear examples. Proper priors for θ_i are usually applied, otherwise the Bartlett paradox may hold and models with less parameters preferred. The point forecast is computed by taking the mean of the predictive density in (15).

We note that an alternative Bayesian procedure to compute model weights is presented below under scheme 8.

Scheme 7: BMA using predictive likelihood

Geweke and Whiteman (2006) propose a BMA based on the idea that a model is good as its predictions. The predictive density of y_{T+1} conditional on F_T has the same form as equation (15), but the posterior density of model m_i conditional on F_T is now computed as:

$$P(m_i|F_T) = \frac{p(y_T|F_{T-1}, m_i)p(m_i)}{\sum_{j=1}^{n} p(y_T|F_{T-1}, m_j)p(m_j)} \tag{19}$$

where $p(y_T|F_{T-1}, m_i)$ is the predictive likelihood for model m_i, e.g. the density derived by substituting the realized y_T in the predictive density of y_T conditional on F_{T-1} given model m_i. We compute the predictive density for month T using information until month $T-1$ and we evaluate the *realized* value for time T using the same density. The resulting probability is then applied to compute the weight for model m_i in constructing the forecast for $T+1$ made at time T[5]. Similar to scheme 6, the point forecast is computed by taking the mean

[5]Eklund and Karlsson (2007) evaluate the fit of the predictive density over some more observations, by means of the predictive likelihood, and then update the probability density for the forecasts. The latter approach results in weights which are based more on the fit of the model, even when using out-of-sample data, than on the probability of out-of-sample realized values. Our approach incorporates the uncertainty that

of the predictive density in (15).

Scheme 8: BMA using time varying model weights

We present a new combination scheme that extends the time varying weight scheme 5 by adding parameter uncertainty and model uncertainty. We reformulate equations (12) and (13) by substituting the means of the conditional predictive densities $p(y_T|F_{T-1}, m_i)$ given models m_i, $i = 1, 2$ for the point forecasts $\widehat{y}_{T,i}$. Then we apply Bayesian inference using Gibbs sampling to estimate w_t; for details we refer to Appendix C. The result is a set of posterior densities for the model weights given the data F_T, $p(w_{T+1,i}|F_T)$. These posterior densities are used to average over the conditional predictive densities given F_T and model m_i

$$p(y_{T+1}|F_T) = p(w_{T+1,0}|F_T) + \sum_{i=1}^{n} p(w_{T+1,i}|F_T)p(y_T|F_{T-1}, m_i) \tag{20}$$

in order to derive the predictive density of y_{T+1} given F_T. The point forecast is computed by taking the mean of the predictive density in (20).

Scheme 8 allows for parameter uncertainty by applying Bayesian analysis to individual models m_i, for model uncertainty by combining the conditional predictive densities given F_T and model m_i, and for time varying patterns by assuming a pattern for model weights as in (13). It also extends scheme 5 by providing a density forecast and not only a point forecast. Thus, for instance, forecasting and policy measures with respect to risk management can be performed in a more flexible way.

We emphasize that special cases of this proposed scheme may be constructed as Bayesian versions of schemes 3 and 4. More details are presented in Appendix C.

3 Simulation exercises

In this section we describe ten simulation exercises to evaluate the eight forecast combination schemes presented in Section 2. In exercises I-III the correlation between predictors varies

future out-of-sample values may differ from historical out-of-sample realizations. It would be more natural to compute the predictive likelihoods as product of the predictive likelihood made for last υ successive forecasts. Some computational problems may arise because any predictive likelihood is in the interval [0,1]. Then, it might be difficult to work with possible small numbers, or if only one predictive value of the υ averaged is close to zero the weight on the respective model will be zero independently by performances in the other periods.

from low to high; in exercises IV-VII misspecification with respect to the number of included predictors and number of included models is explored; exercises VIII-IX deal with structural change; finally exercise X considers the case of fat tailed generated data patterns.

Following previous notation, we simulate DGPs in a range of settings from equations (1)-(3). We fix $T = 240$ and $H = 120$, that is the genuine out-of-sample period has 120 one-step ahead forecasts. The last 60 observations of the in-sample period ($t = 181, .., 240$) are used as initial training period for the combination schemes. We repeat each exercise 1000 times. In all examples we assume that the predictor variables (x) are normally distributed with values for the means (μ), variances (σ^2) and covariances (ϱ) that are specified in Table 1. The disturbances are assumed to be i.i.d normal (0,1). We restrict the DGP weights $c_{s,1}$ and $c_{s,2}$ to add to 1 for any s in order to exclude shifts in the unconditional mean of the DGP. In exercises I-VII $\{c_{s,1}\}_{s=1}^{360}$ and $\{c_{s,2}\}_{s=1}^{360}$ are time invariant and the DGP is stationary. In exercises VIII-X time-variation is added. In Bayesian analysis we generally use diffuse proper priors for the model parameters.

For any simulation we compute the MSPE's of the individual forecasts and forecast combinations over the 120 "genuine" one-step ahead forecasts, and its decomposition in bias and variance of the forecast errors. In Table 2 we report the average of 1000 MSPE's, bias and variance of the forecasts. For completeness, we also give the same statistics for the correctly specified models (labelled as "correct" model), and the forecast combination where the vector $\widehat{w}_{T+1}$ is identical to $[c_1, c_2]$ (labelled as "given" weights).

3.1 Varying correlations between predictors

In exercises I-III a stationary DGP is simulated; c_1 and c_2 are plotted at the top-right corner in Figure 1: c_1 is set almost two times c_2. The difference in exercises relates to the degree of correlation between the individual forecasts.

Exercise I: zero correlation between predictor variables We first give some analytical results that may help the analysis. With the parameter values from Table 1, it is easy to derive that

$$y_{s,i} = 2 + \epsilon^*_{s,i} \qquad with \;\; \epsilon^*_{s,i} \sim iidN(0,3)$$

with $i = 1, 2$. Then,

$$y_s = 0.7y_{s,1} + 0.3y_{s,2} = 2 + 0.7\epsilon^*_{s,1} + 0.3\epsilon^*_{s,1} \tag{21}$$

Accordingly, the expected value of y_s is $E(y_s) = 2$ and its variance is $V(y_s) = 1.74$. We also

notice that the coefficients of the variables $(x_{s,1}, x_{s,2})$ in the simulated DGP are $(\beta_1 c_{s,1}) = 0.7$ and $(\beta_2 c_{s,2}) = 0.3$ for any s.

By computing the probability limit of the OLS estimator $\widehat{\beta}_1$ in model (1) we find that $\widehat{\beta}_1$ is a consistent estimator of $(\beta_1 c_{s,1})$, its estimate is close to 0.7 for any $s = 181, .., 360$, and $\widehat{\beta}_2$ is a consistent estimator of $(\beta_2 c_{s,2})$, its estimate is close to 0.3 for any $s = 181, .., 360$. Moreover, both $(\widehat{\alpha_1} + \widehat{\beta}_1)$ and $(\widehat{\alpha_2} + \widehat{\beta}_2)$ add to 2 implying that the forecasts of the single models are unbiased, since $E(y_s) = 2$. In term of accuracy (MSPE), equation (1) does much better than equation (2), but the difference with the correct model, in which both (x_1, x_2) are included, is substantial. As the forecasts of both models are unbiased, the difference in accuracy is only due to the variance of the prediction errors. The variance of the prediction error of model (2) is more than double than that of the prediction error of model (1), reflecting the choice of (c_1, c_2)[6].

We find that the forecasts from the individual models and frequentist combination schemes can be approximated respectively as:

Model 1	$\widehat{y}_{T+h,1} = 1.3 + 0.7x_{T+h,1}$
Model 2	$\widehat{y}_{T+h,1} = 1.7 + 0.3x_{T+h,2}$
True model	$\widehat{y}_{T+h} = 1 + 0.7x_{T+h,1} + 0.3x_{T+h,2}$
Given weights	$\widehat{y}^{(g)}_{T+h} = 1.42 + 0.49x_{T+h,1} + 0.09x_{T+h,2}$
Case 1	$\widehat{y}^{(1)}_{T+h} = 1.5 + 0.35x_{T+h,1} + 0.15x_{T+h,2}$
Case 2	$\widehat{y}^{(2)}_{T+h} = 1.42 + 0.49x_{T+h,1} + 0.09x_{T+h,2}$
Case 3	$\widehat{y}^{(3)}_{T+h} = 1 + 0.7x_{T+h,1} + 0.3x_{T+h,2}$
Case 4	$\widehat{y}^{(4)}_{T+h} = 1 + 0.7x_{T+h,1} + 0.3x_{T+h,2}$
Case 5	$\widehat{y}^{(5)}_{T+h} = 1 + 0.7x_{T+h,1} + 0.3x_{T+h,2}$

where $h = 1, .., 120$.

The estimators of β_1, β_2 are consistent for the products $(\beta_1 c_{s,1})$, $(\beta_2 c_{s,2})$. Therefore, esti-

[6]We compute forecasts also by applying Bayesian inference (with diffuse priors). We do not report results because they are very similar to the previous ones.

mating $(c_{s,1}, c_{s,2})$ both equal to 1 is the optimal solution to reduce the variance of the prediction errors. Combination schemes 3, 4 and 5 are the only methods to provide estimates of (c_1, c_2) equal to vectors of 1, providing the best statistics. Recursive and time-varying weight schemes, which allow for time varying estimates of (c_1, c_2), do not improve results compared to constant OLS weight scheme as (c_1, c_2) are time-invariant in the simulation. Other combination approaches (given weights, case 1 and 2) provide different estimates of (c_1, c_2), implying that the products $(\widehat{\beta}_1 \widehat{w}_1^j)$ and $(\widehat{\beta}_2 \widehat{w}_2^j)$ are not consistent estimator of $(\beta_1 c_1)$ and $(\beta_2 c_2)$. The forecasts given by those combination schemes are still unbiased but the variance of the prediction errors is higher. For example, assigning weights to single models based on the inverse of the MSPE well approximates the variance of the noises of the single models, $\epsilon_{s,1}^*$ and $\epsilon_s, 2^*$ respectively. Indeed, weight estimates of this scheme are very similar to the original values $c_1 = 0.7\iota$ where ι is a (120×1) vector of ones and $c_2 = 0.3\iota$ such as in the given weight combination. But this is not optimal in this exercise.
To sum up, model (1) predicts the part of the DGP related to $x_{s,1}$, model (2) predicts the part of the DGP related to $x_{s,2}$. Therefore, the optimal averaging strategy is adding with weight 1 the forecasts of the individual models and inserting a constant term to avoid biases. As Table 2 confirms, both the OLS weights and Terui and van Dijk (2002)'s time varying extension model this providing very accurate forecasts.

The Bayesian averaging scheme using marginal likelihood requires a different explanation. What is important in Bayesian averaging is assigning the right probability to individual models. BMA based on marginal likelihood does not do this job well: it gives almost all the probability to model (1) and zero probability to model (2). The problem apparently relates to the use of un-normalized marginal likelihoods. To derive the marginal likelihood given by the individual models we compute the log marginal likelihood. Figure 2 plots the average of the log marginal likelihood for the two individual models for $s = 181, .., 360$ over the 1000 simulations. When we take the exponential to compute posterior weights the two numbers are not anymore comparable. And since (1) has higher log marginal likelihood all the probability is given to it. We note that more sophisticated ways of computing marginal likelihoods may exist, but we do not pursue this further. Instead we present a group of "simple to compute" Bayesian schemes under scheme eight.
BMA based on predictive likelihood gives on average probabilities similar to the original values 0.7 and 0.3. But its performance is not up to the level of estimated weight schemes. Bayesian results depend on the priors that we apply. We assume diffuse proper priors for model parameters, which imply parameter posterior means around OLS estimates (for

derivation see, e.g., Koop, 2003, p. 37). The priors for $(\alpha_1, \alpha_2, \beta_1, \beta_2)$, however, could be chosen very informative around the true values 1. Then, averaging models (1) and (2) with predictive likelihoods would provide forecasts very similar to the correct model[7]. We think that it is in practice not easy to find such accurate priors and not all agents may agree on these precise priors, therefore we have applied diffuse priors that allow direct comparison to frequentist inference, but these diffuse priors apparently reduce forecast accuracy.

The use of diffuse priors does not reduce the forecast accuracy of scheme 8 compared to that of schemes 3-5. In scheme 8 a Gibbs sampling procedure is applied to combine predictive densities of individual models. This Gibbs procedure is a Bayesian extension of scheme five. Results may be even more accurate when informative priors are applied.

Exercise II: medium correlation In the second exercise the correlation of the individual forecasts is increased and a medium positive (0.5) correlation between $x_{s,1}$ and $x_{s,2}$ is assumed for any s.

Model (1) performs better than model (2) due to the magnitude of the weights. Estimated weight schemes and Bayesian time varying weight scheme provide again better statistics than other averaging schemes, with results very similar to the correct model. However, simple combination schemes and BMA based on the predictive likelihood also give quite accurate forecasts. In this exercise model (1) and model (2) do not provide consistent estimate of $(\beta_1 c_{s,1})$ and $(\beta_2 c_{s,2})$, therefore weight estimates achieve this result. BMA based on marginal likelihood still selects only model (1).

Exercise III: high correlation In this exercise, the correlation of the individual forecasts is substantially increased (around 0.9). As in Timmermann (2006) in this framework equal weights are an appropriate choice. All the schemes forecast accurately and very similar to the correct model, since the individual models (1) and (2) give accurate and highly correlated results. Note that the time varying weight combinations are robust in this case.

3.2 Misspecification

In Exercises IV-VII the number of predictors and individual models varies. The DGP is still assumed stationary.

[7]Results for this exercise are available upon request.

Exercise IV: included irrelevant variable In exercise IV an irrelevant variable (x_6) is included as additional regressor in model (1); its coefficient β_6 is given in Table 1. Due to the long series and the number of repetitions of the simulations β_6 is correctly estimated to be zero and results are very similar to exercise I.

Exercise V: omitted relevant variable In exercise V, a new variable, $x_{s,6}$, is added in the simulation of the DGP in equation (3). This variable is excluded in both models (1) and (2). All the forecasts are less accurate than in exercise I and the difference with the forecasts of correct model is substantial. However, estimated weight and Bayesian time varying weight schemes still give better statistics than individual models and other combination schemes. Results given by simple schemes and BMA schemes 6 and 7 are marginally worse than those of model (1).

Exercise VI-VII: 3 and 5 individual models The analysis is extended to include three and five individual models in the simulation exercise. Individual series are combined with weights given in Figure 3. In exercise VI $c_4 = \{c_{s,4}\}_{s=1}^{360}$ and $c_5 = \{c_{s,5}\}_{s=1}^{360}$ are vector of zeros.

In both examples, the estimated weight and Bayesian time varying weight schemes give the best forecasts. These schemes provide forecasts very similar to the correct model, and are the only ones to outperform the best individual model. Simple combination schemes do perform worse than the best individual model and Bayesian model averaging using marginal likelihoods. In the exercises where the misspecification of individual models is more substantial, allowing for parameter uncertainty is beneficial, even if parameter priors are not precise.

3.3 Structural change

In the following two exercises, VIII and IX, the vectors c_1 and c_2 in equation (3) are subject to instability. For exercise VIII, Figure 4 shows that a shift happens at the beginning of the out-of-sample period. The weights assigned to models (1) and (2) are exactly reversed. In exercise IX, two shifts are plotted in Figure 5, at different times, with one of them in the in-sample period, and of opposite direction.

Exercise VIII: one shift The recursive OLS weight and (Bayesian) time-varying weight schemes dramatically outperform individual models, other combination schemes, and the correct model. The weight estimates of these three schemes capture the signal of instability,

and react faster to it, partially reducing the inefficiency of parameter estimates of the individual models, which do not allow for instability in estimation. Rejecting instability may cause serious mistakes and, indeed, the correct model[8] gives marginally worse statistics than model (2). However, the instability, and therefore its signal, is quite moderate due to the fact that we have a unique break over the full sample. As Appendix A shows, this explains why recursive OLS and the Kalman Filter produce very similar weight estimates. Bayesian time varying weigh scheme 8 produces results very similar or marginally superior to scheme 5 again due to the use of diffuse priors.
BMA with predictive likelihood now provides quite accurate forecasts, even though it gives too high probability to model (2). BMA with marginal likelihood does not seem adequate even in this exercise. It assigns all the weight to model (1).

Exercise IX: two shifts The correct model gives the most accurate forecasts. The second shift partially correct the first one and moves the weight patterns close to their in-sample average value. The time varying weight schemes provides the lowest statistics comparing to individual models and other averaging schemes. The instability is higher than in exercise VIII therefore the difference between the recursive OLS and the Kalman filter is evident, following the derivations in Appendix A. Simple combination schemes provide less accurate results. BMA based on predictive likelihoods copes with instability quite efficiently, but the diffuse type of priors chosen for individual model parameters reduce the forecast accuracy. Interestingly, the other BMA method initially assigns positive probability to both models, but when the number of observation increases, it converges to assign all the weight to model (1).

3.4 Fat tails

The DGP from exercise IX is changed by assuming fat tailed errors. In particular, $\epsilon_{s,1}$ and $\epsilon_{s,2}$ in (1)-(2) are assumed to be Student t distributed with mean, variance and ν degree of freedom in Table 1. The DGP weights are still as in Figure 5. All forecasts are less accurate than in exercise IX, but the results are qualitatively similar to the previous example. Again, the time varying weight schemes provide the lowest statistics among the averaging schemes and provides results very close to the correct model. Adding parameter uncertainty seems beneficial as scheme 8 gives marginally superior results that scheme 5. As

[8]We remember that the "correct" model does not account for instability.

in the previous case, several averaging schemes give more accurate forecasts than individual models, confirming that averaging is in our set up of experiments a simple and attractive way to cope with instability.

3.5 Summary of findings

The results in Table 2 indicate that it is not easy to find a general rule how to average individual models in an optimal way, and elements as the degree of correlation of the individual forecasts, data predictability, structural instability and model (mis)specification, play a strategic role in the process of combining forecasts of individual models. In particular, we find that in situations of low predictability and high noise, and almost no correlation of a limited set of individual forecasts, combination schemes that estimate model weights and their extension in a Bayesian framework give the most accurate forecasts. Intuitively, when individual forecasts contain complementary information, the best averaging strategy is to add this independent information. Simple combination schemes are not adequate schemes as they average individual models instead of adding with weight 1 the independent information of different models. Bayesian model averaging based on marginal likelihood has some computational problems due to the fact of deriving un-normalized marginal likelihoods for a relative small set of individual models. Bayesian model averaging based on predictive likelihood assigns precise weights to individual models, but using diffuse priors in model parameters as we do reduce the forecast accuracy.

If the DGP is also subject to structural instability, in the sense that the relevance of the predictors varies over time, time varying weight schemes give the highest predictive gains. Simple combination schemes and recursive OLS weight schemes do not learn (efficiently) from the signals of instability, and therefore do not react fast to it. Bayesian model averaging based on predictive likelihood copes better with instability, but inadequate priors can reduce forecast accuracy. Results are qualitative similar when the distribution has fatter tails than the standard normal case, and adding more sources of uncertainty as the Bayesian time varying weight scheme does seems to be beneficial.

4 Empirical illustration

We extend our study by investigating the forecasting performance and economic gains obtained by applying the eight forecast combination schemes to the case of US stock index returns, defined as the discretely compounded monthly return on the S&P 500 index in

excess of the 1-month T-Bill rate, from January 1976 to December 2005, for a total of 360 observations; see Figure 6. We use two linear non-nested forecasting models. The first one is based on the idea that a set of financial and macroeconomic variables are potentially relevant factors for forecasting stock returns. Among others, Pesaran and Timmermann (1995), Cremers (2002), Marquering and Verbeek (2004) have shown that such variables have predictive power. We label this forecasting model "Leading factor" (LF). The second forecasting model is a simple linear regression model with a constant and a dummy for November-April. It is based on the popular market saying "Sell in May and go away", also known as the "Halloween indicator" (HI), and it based on the assumption that stock returns can be predicted simply by deterministic time patterns. This suggests to buy stock in November and sell it in May. Bouman and Jacobsen (2002) show that this strategy has predictive power.

4.1 Data and evaluation

The source of the S&P 500 index is the CRSP database and the 1-month T-Bill rate is from Ibbotson and Associates. We include as predictors the S&P 500 index price-earnings ratio (PE), the S&P 500 index dividend yield (DY) defined as the ratio of dividends over the previous twelve months and the current stock price, the 3-month T-Bill rate ($I3$), the monthly change in the 3-month T-bill rate ($DI3$), the term spread (TS) defined as the difference between the 10-year T-bond rate and the 3-month T-bill rate, the credit spread (CS) defined as the difference between Moody's Baa and Aaa yields, the yield spread (YS) defined as the difference between the Federal funds rate and the 3-month T-bill rate, the annual inflation rate based on the producer price index (PPI) for finished goods (INF), the annual growth rate of industrial production (IP), the annual growth rate of the monetary base (MB), and the log monthly realized volatility of the S&P 500 index ($LVol$). The monthly realized volatility is computed using daily returns, where we follow French *et al.* (1987) and Marquering and Verbeek (2004) by assuming that daily returns are appropriately described by a first-order autoregressive process. In particular, we use the following estimate for realized volatility

$$\hat{\sigma}_t^2 = \sum_{t=1}^{N_s} (y_{i,t} - \bar{y}_t)^2 \left[1 + \frac{2}{N_t} \sum_{j=1}^{N_t - 1} (N_t - j) \hat{\phi}_t^j, \right]$$

where $y_{i,t}$ is the return on day i in month t which has N_t trading days, $\bar{y}_t$ is the average daily return in month t, and $\hat{\phi}_t$ denotes the first-order autocorrelation estimated using daily returns within month t. We take into account the typical publication lag of macroeconomic

variables in order to avoid look-ahead bias. We therefore include inflation and the growth rates of industrial production and the monetary base with a two-month lag. As the financial variables are promptly available, these are included with a one-month lag. Finally, the "Halloween indicator" (HI) model is specified as a simple linear regression with a constant and a dummy for November-April.

We evaluate the statistical accuracy of the individual models and the eight forecast combinations schemes in terms of MSPE, and its decomposition in square bias and variance of the forecast errors. Again Bayesian predictive densities are computed for the BMA schemes. Moreover, as an investor is more interested in the economic value of a forecasting model than its precision, we test our conclusions in an active short-term investment exercise, with an investment horizon of one month. The investor's portfolio consists of a stock index and riskfree bonds only. At the start of month $T+1$, the investor decides upon the fraction of her portfolio to be invested in stocks $w_{p,T+1}$, based upon a forecast of the excess stock return y_{T+1}. The investor is assumed to maximize a mean-variance utility function

$$\max_{w_{T+1}} u(E_T(y_{p,T+1}), Var_T(y_{p,T+1})) \tag{22}$$

where $y_{p,T+1}$ is the return of the investor's portfolio return at time $T+1$, which is equal to

$$y_{p,T+1} = W_T((1 - w_{p,T+1})(y_{f,T+1}) + w_{p,T+1}(y_{f,T+1} + y_{T+1})) \tag{23}$$

where W_T denotes the wealth at time T, where y_{T+1} is the excess returns on S&P500, and where $y_{f,T+1}$ is the riskfree rate.

Without loss of generality we set initial wealth equal to one, $W_T = 1$. Further, we assume the following utility function:

$$E_T(y_{p,T+1}) - \frac{1}{2}\gamma Var_T(y_{p,T+1}) \tag{24}$$

where γ is the coefficient of relative risk aversion. Solving the maximization problem shows that the optimal portfolio weight for the investor is given by:

$$w^*_{p,T+1} = \frac{E_T(y_{T+1}) - r_{y,T+1}}{\gamma Var_T(y_{T+1})}. \tag{25}$$

If the expected excess return on the risky asset increases, it is optimal for the investor to increase her weight on the risky asset. The conditional variance $Var_T(y_{T+1})$, which represents a measure of the risk involved, is negatively related to this weight. We forecast $E_T(y_{T+1})$ with nine different approaches: two individual models, the 'leading factor' one (LF), and the

'Halloween indicator' one (HI), and the eight averaging schemes discussed in this paper. Each individual forecasting approach corresponds to an investment strategy which is defined in the same way. We approximate the conditional variance with the 60-month moving window average of the realized variances computed as above[9]. We also assume that short selling and borrowing at the riskfree rate are not allowed, therefore we restrict the portfolio weights to be between 0 and 1. For purposes of comparison we consider a passive investment strategy where the total wealth is invested in the risky market (RW).

We evaluate the different investment strategies by computing the average return, the standard deviation of the portfolio return, and the Sharpe ratio, defined as the ratio of the mean excess return on the (managed) portfolio and the standard deviation of the portfolio return. Since the Sharpe ratio overestimates risk in case of time varying volatility, we also compute the *ex post* utility levels - in order to estimate the economic value of the strategy - by substituting the realized return of the portfolios at time $T+1$ in (24)

$$U^*_{p,T+1} = y_{p,T+1} - \frac{1}{2}\gamma w^2_{p,T+1} Vol_{T+1} \tag{26}$$

where Vol_{T+1} denotes the ex post realized volatility of the risky return on month $T+1$. Total utility is then obtained as the sum of U^*_p across all H investment periods. The above approach enables us to compare alternative investment strategies by calculating the associated average utility levels.

Finally, as the portfolio weights in the active investment strategies change every month, the portfolio must be rebalanced accordingly. Hence, transaction costs play a non-trivial role and should be taken into account when evaluating the relative performance of different strategies. Rebalancing the portfolio at the start of month $T+1$ means that the weight invested in the risky asset is changed from w_T to w_{T+1}. We assume that transaction costs amount to a fixed percentage c on each traded dollar. Setting the initial wealth W_T equal to 1 for simplicity, transaction costs at time $T+1$ are defined as equal to

$$c_{T+1} = 2c|w_{T+1} - w_T| \tag{27}$$

where the multiplication by 2 follows from the fact that the investor rebalances her investments in both stocks and bonds. The net portfolio return is then given by $r_{T+1} - c_{T+1}$. We

[9]We also forecast the conditional variance $Var_T(y_{T+1})$ using an AR(1), an AR(12), an Heterogeneous Autoregressive (HAR) model similar to Corsi (2004), and an EGARCH model as in Marquering and Verbeek (2004). Results are qualitative similar. We prefer the 60-month moving window average because most investors use similar simple schemes, in particular at beginning of our sample period.

apply three scenarios with transaction costs of 0.1%, 0.5% and 1%[10]. Note that for the passive investment strategy where the total wealth is invested in the risky market the inclusion of transaction costs matters only in setting up the portfolio at time T_0.

4.2 Empirical Results

The analysis for the active investment strategies is implemented for the period from January 1996 until December 2005, involving 120 one month ahead excess stock return forecasts. The models are estimated recursively using an expanding window of observations. The period January 1991 to December 1995 is used to start up the forecast combination schemes. The investment strategies are implemented for three levels of relative risk aversion, $\gamma = 2, 5$ and 10. Before we analyze the performance of the different portfolios, we summarize the statistical accuracy of the excess return forecasts.

4.2.1 Statistical accuracy

The statistical accuracy of the individual models and forecast combination is evaluated by MSPE, and its decomposition in square bias and variance as in Section 3. Results are reported in Table 3. In the market column, labelled RW, we report the statistics of the Random Walk model.

We notice that both the individual models provide much lower evaluation criteria than the RW. In particular, the Halloween Indicator model has the lowest MSPE error and both the mean and the variance of the forecast errors are lower than for the other individual models. However, both series of forecasts have a quite different pattern than the very noise excess return series in Figure 7. The HI model has a seasonal pattern given by the particular strategy with a positive unconditional mean, and few negative forecasts only in 2002. The LF generates forecasts which are more volatile, and in particular too low at the end of 1990's and at beginning of 2000, and too high in 2001. In term of sign prediction the HI strategy performs very well in 90's. The 60 month moving average sign hit ratios, which are the proportions of correctly predicted signs of the excess return over the previous 60 months, shown in Figure 8, are higher than 0.7 and close to 0.8. But after 1998, the ratios start to deteriorate and stabilize at hit ratios around 0.5 for the final years of the sample period. The higher percentage of positive returns in 90's, and the almost always positive forecasts

[10]We think that 10 basis points is an average transaction cost to buy a 1-month future on S&P500 or a 1-month future on 1-month Treasury Bill.

given by model HI may explain the result. The hit ratios given by the LF model are more stable and on average just above 0.5. In term of MSPE, Figure 9 show similar predictive patterns of the set of forecasts, but after middle of 1996 the HI model always provides lower mean square errors than the LF model.

When averaging schemes are applied, the results are intriguing; see the top of Table 3 for details. The MSPEs of schemes 1, 2, 3, 4, 6 are all higher than that of model HI. Moreover, constant OLS and recursive OLS schemes have a positive bias[11]. The time varying weight schemes, however, provide the best statistics. If we investigate the weight estimates, we find that there is an indication of a break in the weight for model HI in the training period at year 1995, moving from a lower value to values very stable around 1. At the same time, the weight on model LF decreases and stabilizes around -0.5. This confirms ex-post instability evidence in Figure 9 that model HI provides more accurate forecasts than the alternative model after 1996. The dramatic boom of stock prices at the end of 90's and well documented lower predictability of macroeconomic and financial indicators can explain this result. It may also indicate that strategy HI captures some seasonal stylized facts of stock index returns and assigning weight 1 to it is beneficial in term of forecasting performance.

The BMA with predictive likelihood also gives a marginal lower MSPE than the individual model HI. These results are similar to the ones from exercise IX, which shows that the BMA scheme 7 copes with possible instability better than simple combination schemes.

Summarizing, the forecast statistics of the combination schemes are rather similar; the largest difference between schemes is less than 5%. However, because predictability of stock market is very low, small improvements in MSPE may have substantial economic value. To investigate this we implement a portfolio exercise, reported n the next section.

4.2.2 Economic value

Panel B of Table 3 provides performance measures for the different investment strategies based on the ten different forecasting methods presented in the previous sections. Over the forecasting period, January 1996 to December 2005, the average return on the stock portfolio is 10%, the standard deviation is 16%, and the Sharpe ratio is 0.12. The strategies based on forecasting returns with one of the two individual models give lower mean returns for a moderately risk averse ($\gamma = 5$) investor, but also lower standard deviation, which results in a higher Sharpe ratio for the Halloween strategy. Accounting for possible time

[11]We emphasize that their bias is insignificant with respect to the MSPE, and it is less than 0.2% of the unconditional mean return.

varying volatility, and evaluating strategies with the *ex-post* realized utility shows that the Halloween indicator performs better than the leading indicator and the market. The leading factor strategy gives very low mean portfolio returns, which implies a low Sharpe ratio and utility level.

Next, consider the strategies based on forecasting excess returns with the eight averaging schemes. Strategy 5 and 8, based on time varying model weights, give the highest mean returns among all the active strategies, among the lowest standard deviations, and the highest Sharpe ratios and utility levels. In particular, the Bayesian time varying weight scheme has marginally higher mean return but also standard deviation. Strategy 7, based on BMA with predictive likelihood, provides also marginally superior results in terms of portfolio measures than the strategy HI, but substantially lower than the previous strategy. Again, more precise priors may be chosen, but we omit this "subjective" exercise. All other strategies have lower economic values, in particular, give lower mean portfolio returns. Results are qualitative similar for a risk seeking investor ($\gamma = 2$) and a risk averse investor ($\gamma = 10$). Moreover, adding transaction costs does not change the quality of the results, and even with substantial transaction costs of 100 basis points, strategies 5 and 8 give higher levels of utility compared to a random walk strategy of investment. We notice that their Sharpe ratios are lower, confirming that the Sharpe ratio may overestimates risk in case of time varying volatility.

To conclude, the results indicate than the individual models HI and LF provide different forecasts. Moreover, instability in the relation between realized excess returns and individual forecasts seems to be relevant. As in the simulation exercises, in the empirical example the time varying weight schemes give the highest predictive gains both in statistical measures and economic gains.

5 Conclusions

Investors often have a set of forecasts on asset returns available from different models. Such investors may attempt to discover which is the best forecasting model and use it to allocate their portfolios, or they may consider all forecasts and take decisions by averaging forecast information from the individual models. In this paper we explained in a simulation experiment that when data is subject to low predictability, low correlation among individual forecasts, and structural instability, the Terui and van Dijk (2002) time varying model weight scheme and its extension in a Bayesian framework to incorporate parameter uncertainty provides

the most accurate forecasts compared to other frequentist and Bayesian model averaging (with diffuse priors on model parameters) schemes. We applied the different model averaging schemes also to forecast the index of US stock returns. As in the simulation exercise, stylized facts of stock index data are low predictability and possible structural instability. We considered two forecasting models that represent different views on predicting the US stock index. We have shown, firstly, that averaging strategies can give higher predictive gains than selecting the best model; secondly, that time varying model weights have higher statistical and economic values than other averaging schemes considered. An interesting topic for further research is to compare our results to other time varying weight combination schemes, such as regime switching, see e.g. Guidolin and Timmermann (2007), or schemes that carefully model breaks, see e.g. Ravazzolo *et al.* (2007). Moreover, combination schemes can be applied to the analysis of density forecasts. Market operators, such as financial investors or central bank decision makers, are becoming increasingly interested in knowing the complete distribution of the assets of interests for purposes of risk management. The Bayesian time varying weight scheme that we put forward seems particular adequate in this context.

References

Aiolfi, M. and A. Timmermann (2006), Persistence in Forecasting Performance and Conditional Combination Strategies., *Journal of Econometrics*, **135 (1)**, 31–53.

Bates, J. M. and C. W. J. Granger (1969), Combination of Forecasts, *Operational Research Quarterly*, **20**, 451–468.

Bouman, S. and B. Jacobsen (2002), The Halloween Indicator, Sell in May and Go Away: Another Puzzle, *American Economic Review*, **92 (5)**, 1618–1635.

Carter, C. and R. Kohn (1994), On Gibbs Sampling for State-Space Models, *Biometrika*, **81**, 541–553.

Chib, S. (1995), Marginal Likelihood from the Gibbs Output, *Journal of American Statistical Association*, **90**, 972–985.

Corsi, F. (2004), Simple Long Memory Models of Realized Volatility, *Lugano Working paper*, 26.

Cremers, K. J. M. (2002), Stock Return Predictability: A Bayesian Model Selection Perspective, *Review of Financial Studies*, **15**, 1223–1249.

Draper, D. (1995), Assessment and Propagation of Model Uncertainty, *Journal of the Royal Statistical Society Series B*, **56**, 45–98.

Eklund, J. and S. Karlsson (2007), Forecast Combination and Model Averaging using Predictive Measures, *Econometric Reviews*, **26**, 329–362.

Fernández, C., E. Ley, and M. F. J. Steel (2001), Model uncertainty in cross-country growth regressions, *Journal of Applied Econometrics*, **16**, 563–576.

French, K. R., G. W. Schwert, and R. F. Stambaugh (1987), Expected Stock Returns and Volatility, *Journal of Financial Economics*, **19**, 3–29.

Geman, S. and D. Geman (1984), Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J. and C. Whiteman (2006), Bayesian Forecasting, in G. Elliot, C. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, North-Holland.

Granger, C. W. J. and R. Ramanathan (1984), Improved Methods of Combining Forecasts, *Journal of Forecasting*, **3**, 197–204.

Guidolin, M. and A. Timmermann (2007), Forecasts of US Short-term Interest Rates: A Flexible Forecast Combination Approach, *forthcoming in Journal of Econometrics*.

Hansen, B. E. (2007), Least Squares Model Averaging, *Econometrica*, **75(4)**, 1175–1189.

Harvey, A. C. (1993), *Time Series Models*, Pearson Education.

Harvey, A. C., T. M. Trimbur, and H. K. van Dijk (2006), Trends and cycles in economic time series: A Bayesian approach, *forthcoming in Journal of Econometrics*.

Hendry, D. F. and M. P. Clements (2004), Pooling of Forecasts, *Econometric Reviews*, **122**, 47–79.

Hodges, J. (1987), Uncertainty, Policy Analysis and Statistics, *Statistical Science*, **2**, 259–291.

Koop, G. (2003), *Bayesian Econometrics*, John Wiley & Sons Ltd, West Sussex, England.

Leamer, E. (1978), *Specification Searches*, New York: Wiley.

Madigan, D. and A. Raftery (1994), Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window, *Journal of the American Statistical Association*, **89**, 1335–1346.

Marcellino, M. (2004), Forecasting Pooling for Short Time Series of Macroeconomic Variables, *Oxford Bulletin of Economic and Statistics*, **66**, 91–112.

Marquering, W. and M. Verbeek (2004), The Economic Value of Predicting Stock Index Returns and Volatility, *Journal of Financial and Quantitative Analysis*, **39 (2)**, 407–429.

Min, C. and A. Zellner (1993), Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates, *Journal of Econometrics*, **56**, 89–118.

Pesaran, M. H. and A. Timmermann (1995), Predictability of Stock Returns: Robustness and Economic Significance, *Journal of Finance*, **50**, 1201–1228.

Pesaran, M. H. and A. Timmermann (2002), Market Timing and Return Predictability Under Model Instability, *Journal of Empirical Finance*, **9**, 495–510.

Ravazzolo, F. (2007), *Forecasting Financial Time Series Using Model Averaging*, Tinbergen Institute Research Series 415, Rotterdam.

Ravazzolo, F., R. Paap, D. van Dijk, and P. H. Franses (2007), Bayesian Model Averaging in the Presence of Structural Breaks, in M. Wohar and D. Rapach (eds.), *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Elsevier.

Stock, J. H. and M. Watson (2004), Combination Forecasts of Output Growth in a Seven-country Data Set, *Journal of Forecasting*, **23**, 405–430.

Strachan, R. and H. K. van Dijk (2007), Bayesian Model Averaging in Vector Autoregressive Processes with an Investigation of Stability of the US Great Ratios and Risk of a Liquidity Trap in the USA, UK and Japan, *Econometric Institute Report 2007-09*, 47.

Tanner, M. A. and W. H. Wong (1987), The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, **82**, 528–550.

Terui, N. and H. K. van Dijk (2002), Predictability in the Shape of the Term Structure of Interest Rates, *International Journal of Forecasting*, **18**, 421–438.

Sala-i-Martin, X., G. Doppelhoffer, and R. Miller (2004), Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach, *American Economic Review*, **94**, 813–835.

Timmermann, A. (2006), Forecast Combinations, in G. Elliot, C. W. J. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, North-Holland.

A Comparison of Recursive Least Squares and time varying model weight combinations

The model weights of the OLS averaging scheme 4 can be computed by Recursive Least Squares. Consider (10) and rewrite it as

$$y_t = z_t^{'} w + u_t; \quad u_t \sim N(0, s^2) \tag{28}$$

where $z_t^{'}$ is a $(1 \times q)$ row vector and where w is a $(q \times 1)$ vector of unknown constant parameters. The recursive least squares estimator of the weight w is given as

$$b_t^{(4)} = b_{t-1}^{(4)} + (Z_{t-1}^{'} Z_{t-1})^{-1} z_t (z_t^{'} (Z_{t-1}^{'} Z_{t-1})^{-1} z_t + 1)^{-1} (y_t - z_t^{'} b_{t-1}^{(4)}) \tag{29}$$

$b_t^{(4)}$ is defined recursively as equal to its previous value plus a weighted value of the prediction error $(y_t - z_t^{'} b_{t-1}^{(4)})$ times the observed value of z_t. A minimum of k observations are needed to compute a starting value for the estimator. For details of the derivation see, e.g., Ravazzolo (2007)

The model weights of the time varying averaging scheme 5 are defined as

$$y_t = z_t^{'} w_t + u_t; \quad u_t \sim N(0, s^2) \tag{30}$$

$$w_t = w_{t-1} + \xi_t; \quad \xi_t \sim N(0, \Sigma) \tag{31}$$

where w_t is a $(q \times 1)$ vector of random variables, and u_t and ξ_t are independently and identical distributed for $t = 1, ..., T$, and uncorrelated for all lags, $E(\xi_t, u_\tau) = 0$ for all t and τ, $t \neq \tau$, and where Σ is a diagonal matrix. We make use of the Kalman Filter technique to compute estimators for the model weights w_t. Following Harvey (1993, section 4.3), the distribution of w_t conditional on y_t is multivariate normal with mean

$$b_t^{(5)} = b_{t|t-1}^{(5)} + P_{t|t-1} z_t (z_t^{'} P_{t|t-1} z_t + s^2)^{-1} (y_t - z_t^{'} b_{t|t-1}^{(5)}) \tag{32}$$

and covariance matrix

$$P_t = P_{t|t-1} - P_{t|t-1} z_t (z_t^{'} P_{t|t-1} z_t + s^2)^{-1} z_t^{'} P_{t|t-1} \tag{33}$$

Thus $b_t^{(5)}$, the vector of estimated model weights in (30), is defined equal to its previous value plus a term that is the weighted product of the prediction error $(y_t - z_t^{'} b_{t|t-1}^{(5)})$, the observed value of z_t, and the prediction for the variance of the latent factor estimator $P_{t|t-1}$.

Comparison Let $P_k = (Z_k^{'} Z_k)^{-1}$. Following (32), the weight estimates at time $(k+1)$ given by the Kalman Filter, $b_{k+1}^{(5)}$, can be written as:

$$b_{k+1}^{(5)} = b_k^{(5)} + \left(\frac{(Z_k^{'} Z_k)^{-1}}{s^2} + \frac{\Sigma}{s^2} \right) z_{k+1} \left(z_{k+1}^{'} \left(\frac{(Z_k^{'} Z_k)^{-1}}{s^2} + \frac{\Sigma}{s^2} \right) z_{k+1} + 1 \right)^{-1} (y_{k+1} - z_{k+1}^{'} b_k^{(5)}) \tag{34}$$

where $b_{k+1|k}^{(5)} = b_k^{(5)}$, where $(P_{k+1|k} = (Z_k^{'} Z_k)^{-1} + \Sigma)$, and where s^2 is a scaling parameter bounded from (30) as $0 < s^2 < Var(y)$. The recursive least square estimator of $w^{(4)}$ at time $k+1$ is given in (29) and repeated for convenience as

$$b_{k+1}^{(4)} = b_k^{(4)} + (Z_k^{'} Z_k)^{-1} z_{k+1} (z_{k+1}^{'} (Z_k^{'} Z_k)^{-1} z_{k+1} + 1)^{-1} (y_{k+1} - z_{k+1}^{'} b_k^{(4)}) \tag{35}$$

If Σ is a matrix of zeros and $s^2 = 1$, the weight estimates in (34) and (35) are identical. Otherwise, if k is sufficient large, the elements of the matrix $(Z_k^{'} Z_k)^{-1}$ are relative small. Then by dividing for the scalar s^2 they change marginally. What really matters in such situation for comparing the two estimators in (34) and (35) is the signal to noise ratio (SNR), that is Σ / s^2.

- If the SNR is large, meaning that one or more diagonal elements of Σ are very large comparing to s^2, the weight estimates of the two schemes will differ substantially.
- If the SNR is on contrary small, meaning that s^2 is large compared to the diagonal elements of Σ, the weight estimates in the two schemes will be almost identical.

In our simulation exercise, a large SRN corresponds to large instability in the DGP weights. Thus, our conclusion is that in cases where the data are subject to structural instability, the time varying weight scheme is preferable to the Recursive OLS scheme.

B Graphical examples

We develop few simulation exercises to explain graphically results in Appendix A. Let assume that a series is generated from the following DGP:

$$y_t = 1 + z_t w_{t,1} + u_t; \quad u_t \sim N(0, s^2) \tag{36}$$

$$w_t = w_{t-1} + \xi_t; \quad \xi_t \sim N(0, \sigma^2) \tag{37}$$

where $t = 1, .., T$, where $z = \{z_t\}_{t=1}^{T}$ is a $(T \times 1)$ normally distributed vector with mean μ_z and variance σ_z in Table 4.

We apply the Recursive Least Squares and the Kalman Filter algorithms to estimate $w = \{w_t\}_{t=k+1}^{T}$, defined as $b^{(4)}$ and $b^{(5)}$ respectively, where k are the initial observations to initialize the estimation algorithms. Precisely, we use the OLS estimate of w on the initial k observation and $P_k = (Z_k' Z_k)^{-1}$ to initialize the algorithms.

Exercise B.I: Zero SNR We fix $T = 240$, $k = 120$, $s^2 = 1$, $\sigma^2 = 0$, and $\beta_0 = 1$. Results are in Figure 10. The vector w is constant and the two estimators provide the same results.

Exercise B.II: Medium SNR In this exercise we fix $s^2 = 1$, $\sigma^2 = 0.04$, and $\beta_0 = 1$. Results are in Figure 11. The vector w has a time varying pattern. $b^{(4)}$ and $b^{(5)}$ initialize with the same value, then $b^{(4)}$ is very persistent around the value 1, $b^{(5)}$ on contrary approximates very precisely the pattern of w.

Exercise B.III: High SNR In this exercise we fix $s^2 = 1$, $\sigma^2 = 1$, and $\beta_0 = 1$. Results are in Figure 12. The vector w follows a very high volatile pattern, $b^{(5)}$ accurately estimates it, $b^{(4)}$ is on contrary a poor estimator.

C Estimation of the Bayesian time varying model weight combinations

The model weights of the time varying weights in scheme 8 are defined as in (30) and (31) (z_t may assume different values). The parameters in (30) and (31) are the variances of the residuals in the observation equation, s^2, and the variances of the residuals in the latent equation $q_0^2, \ldots, q_i^2$, where $q_0^2, \ldots, q_i^2$ are the diagonal elements of Σ. The model parameters are collected in the $((1 + i) \times 1)$ vector $\theta = (s^2, q_0^2, \ldots, q_i^2)'$. To facilitate the posterior simulation we make use of diffuse or independent conjugate priors where such values of prior parameters are chosen that we are rather diffuse. For the variance parameters we take the inverted Gamma-2 prior

$$q_j^2 \sim \text{IG-2}(\nu_j, \delta_j) \qquad \text{for } j = 0, \ldots, i \tag{38}$$

and

$$s^2 \sim \text{IG-2}(\nu_s, \delta_s), \tag{39}$$

where ν_j, δ_j, $j = 0, \ldots, i$, ν_s, and δ_s are parameters which can be chosen to reflect diffuse prior beliefs about the variances and the information in the likelihood is allowed to dominate.

Posterior results are obtained using the Gibbs sampler of Geman and Geman (1984) combined with the technique of data augmentation of Tanner and Wong (1987). The latent variables $w = \{w_t\}_{t=1}^T$ are simulated alongside the model parameters θ. The complete data likelihood function is given by

$$p(y, w|z, \theta) = \prod_{t=1}^{T} p(y_t|z_t, w_t, s^2) p(w_t|w_{t-1}, q_0^2, \ldots, q_i^2) \tag{40}$$

where $y = (y_1, \ldots, y_T)'$ and $z = (z_1', \ldots, z_T')'$. The terms $p(y_t|z_t, w_t, s^2)$ and $p(w_t|w_{t-1}, q_0^2, \ldots, q_i^2)$ are normal density functions, which follows directly from (30) and (31) respectively. If we combine (40) together with the prior density $p(\theta)$, which follows from (38)-(39), we obtain the posterior density

$$p(\theta, w|y, z) \propto p(\theta) p(y, w|z, \theta) \tag{41}$$

The sampling scheme can be summarized as follows:

1. Draw w conditional on θ.

2. Draw θ conditional on w.

The full conditional posterior density for the latent regression parameters w in step 1 is computed using the simulation smoother as in Carter and Kohn (1994). Other simulation smoothers can also be applied, see e.g. Harvey *et al.* (2006). The Kalman smoother is applied to derive the conditional mean and variance of the latent factors; for the initial value w_0 a multivariate normal prior with mean 0 is chosen as for scheme 5. To sample the parameters θ in step 2 we can use standard results in Bayesian inference. Hence, the variance parameters s^2 and q_j^2 are sampled from inverted Gamma-2 distributions.
The one-step ahead predictive density of y_{T+1} at time T conditional on y, z and z_{T+1} is given by

$$p(y_{T+1}|y, z, z_{T+1}) = \iint p(y_{T+1}|z_{T+1}, w_{T+1}, s^2) p(w_{T+1}|w_T, q_0^2, \ldots, q_i^2) \\ p(\theta, w|y, z) p(z_{T+1}|z_T) dw d\theta \tag{42}$$

Simulating y_{T+1} from the one-step ahead distribution (42) is in fact rather straightforward. In each step of the Gibbs sampler, we use the simulated values of w_T and $(q_0, \ldots, q_i^2)$, and equation (31) to simulate w_{T+1}. Equation (30) in combination with the simulated value of w_{T+1}, the current Gibbs draws of s^2, and the simulated value of z_{T+1} then provide a simulated value for y_{T+1}.

We emphasize that special cases of our algorithm are Bayesian versions of the OLS schemes 3 and 4. The Bayesian version of schemes 4 is almost identical to scheme 8. The only difference is that we make use of equation (28) (eventually partially reformulated to account for prior information, see e.g. Koop, p. 37) instead of equation (31). We note that the Bayesian version of schemes 3 and 4 do not longer deal with latent weights w_t, but w_t is constant and just a vector of parameters of the model.

Table 1: Simulation design of exercises I-X

	EXERCISES						
PARAMETERS	I,VIII,IX	II	III	IV,V	VI	VII	X
μ_{x_1}	1.00	1.00	1.00	1.00	1.00	1.00	1.00
μ_{x_2}	1.00	1.00	1.00	1.00	1.00	1.00	1.00
μ_{x_3}	-	-	-	-	1.00	1.00	-
μ_{x_4}	-	-	-	-	-	1.00	-
μ_{x_5}	-	-	-	-	-	1.00	
μ_{x_6}	-	-	-	1.00	-	-	-
$\sigma^2_{x_1}$	2.00	2.00	2.00	2.00	2.00	2.00	2.00
$\sigma^2_{x_2}$	2.00	2.00	2.00	2.00	2.00	2.00	2.00
$\sigma^2_{x_3}$	-	-	-	-	2.00	2.00	2.00
$\sigma^2_{x_4}$	-	-	-	-	-	2.00	-
$\sigma^2_{x_5}$	-	-	-	-	-	2.00	-
$\sigma^2_{x_6}$	-	-	-	2.00	-	-	-
ϱ_{x_1,x_2}	0.00	1.00	1.80	0.00	0.00	0.00	0.00
ν	-	-	-	-	-	-	4
α_1, β_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
α_2, β_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
α_3, β_3	-	-	-	-	1.00	1.00	-
α_4, β_4	-	-	-	-	-	1.00	-
α_5, β_5	-	-	-	-	-	1.00	-
β_6	-	-	-	1.00	-	-	-

Table 2: Results of simulation exercises

		Individual Models						Schemes								
Exercises		1	2	3	4	5	Correct	Given	1	2	3	4	5	6	7	8
I-III: Varying correlations between predictors																
	MSPE	0.77	1.57	-	-	-	0.58	0.76	0.88	0.78	0.62	**0.61**	**0.61**	0.76	0.79	**0.61**
0.00	$BIAS^2$	0.01	0.01	-	-	-	0.00	0.01	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.00
	VAR	0.76	1.56	-	-	-	0.58	0.75	0.87	0.77	0.60	0.60	0.61	0.75	0.78	0.61
	MSPE	0.72	1.32	-	-	-	0.58	0.65	0.72	0.66	0.62	**0.60**	**0.60**	0.72	0.71	0.61
0.50	$BIAS^2$	0.01	0.01	-	-	-	0.00	0.01	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.00
	VAR	0.71	1.31	-	-	-	0.58	0.64	0.71	0.65	0.60	0.59	0.60	0.71	0.70	0.61
	MSPE	0.62	0.77	-	-	-	0.58	**0.58**	0.60	0.60	0.62	0.60	0.60	0.63	0.62	0.60
0.90	$BIAS^2$	0.01	0.01	-	-	-	0.00	0.00	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.00
	VAR	0.61	0.76	-	-	-	0.58	0.58	0.59	0.59	0.60	0.59	0.60	0.62	0.61	0.60
IV-VII: Misspecification																
	MSPE	0.76	1.57	-	-	-	0.58	0.76	0.88	0.77	0.63	**0.61**	**0.61**	0.76	0.79	**0.61**
Irrelevant variable	$BIAS^2$	0.01	0.01	-	-	-	0.00	0.01	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.00
	VAR	0.75	1.56	-	-	-	0.58	0.75	0.87	0.76	0.61	0.60	0.61	0.75	0.78	0.61
	MSPE	1.78	2.59	-	-	-	0.58	1.78	1.89	1.83	1.70	**1.66**	1.67	1.85	1.79	**1.66**
Omiited variable	$BIAS^2$	0.02	0.02	-	-	-	0.00	0.02	0.02	0.02	0.05	0.02	0.00	0.07	0.07	0.00
	VAR	1.76	2.57	-	-	-	0.58	1.76	1.87	1.81	1.65	1.64	1.67	1.78	1.72	1.66
	MSPE	0.67	1.21	1.36	-	-	0.46	0.69	0.87	0.76	0.52	0.53	**0.50**	0.66	0.71	**0.50**
3	$BIAS^2$	0.01	0.01	0.01	-	-	0.00	0.01	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.00
	VAR	0.66	1.20	1.35	-	-	0.46	0.68	0.86	0.75	0.50	0.52	0.50	0.66	0.70	0.50
	MSPE	0.57	0.89	1.05	1.07	1.07	0.36	0.59	0.81	0.74	0.44	0.44	**0.40**	0.57	0.61	**0.40**
5	$BIAS^2$	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.00
	VAR	0.57	0.88	1.04	1.06	1.06	0.36	0.59	0.81	0.73	0.42	0.43	0.40	0.57	0.61	0.40
VIII-IX: Structural changes																
	MSPE	1.78	0.99	-	-	-	1.01	1.00	1.11	1.08	1.06	0.76	0.74	1.70	1.01	**0.73**
1	$BIAS^2$	0.01	0.01	-	-	-	0.01	0.01	0.01	0.01	0.02	0.01	0.00	0.02	0.01	0.00
	VAR	1.77	0.98	-	-	-	1.00	0.99	1.10	1.07	1.04	0.75	0.74	1.68	1.00	0.73
	MSPE	0.86	1.26	-	-	-	0.55	0.77	0.81	0.81	0.96	0.68	0.60	0.93	0.80	**0.59**
2	$BIAS^2$	0.01	0.01	-	-	-	0.01	0.01	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.00
	VAR	0.85	1.25	-	-	-	0.54	0.76	0.80	0.80	0.94	0.67	0.60	0.93	0.80	0.59
X: Fat tails																
	MSPE	1.38	1.79	-	-	-	1.08	1.29	1.33	1.33	1.53	1.23	1.15	1.47	1.31	**1.13**
	$BIAS^2$	0.01	0.01	-	-	-	0.01	0.01	0.01	0.01	0.03	0.01	0.00	0.01	0.01	0.00
	VAR	1.37	1.77	-	-	-	1.07	1.28	1.32	1.32	1.50	1.22	1.15	1.46	1.30	1.13

The table presents the mean square prediction error (MSPE), the square bias ($BIAS^2$), and the variance of the prediction errors (VAR) for the individual models and combination schemes, given in section 2, in exercises I-X.

Table 3: Empirical application - No transaction costs

	Individual Models			Strategies							
Statistics	RW	LF	HI	1	2	3	4	5	6	7	8
				Panel A: Statistical accuracy							
MSPE	40.96	21.45	20.04	20.55	20.53	20.96	20.73	19.84	20.33	19.99	**19.83**
$BIAS^2$	0.00	0.00	0.02	0.01	0.01	0.36	0.10	0.01	0.03	0.03	0.01
VAR	40.96	21.45	20.02	20.54	20.52	20.61	20.63	19.84	20.30	19.97	19.82
				Panel B: Economic value							
Criteria						$\gamma = 2$					
Mean	9.94	5.24	9.09	6.59	6.60	10.09	9.24	11.04	9.51	9.84	11.09
St dev	15.63	12.82	13.69	13.93	14.01	15.61	14.48	13.43	14.00	14.41	13.46
SR	0.12	0.04	0.12	0.06	0.06	0.12	0.11	0.16	0.12	0.13	0.16
Utility	0.77	0.35	0.77	0.48	0.48	0.79	0.76	**0.97**	0.76	0.80	**0.97**
						$\gamma = 5$					
Mean	9.94	4.51	8.94	6.45	6.51	8.24	8.12	9.76	8.60	9.57	9.87
St dev	15.63	10.16	11.01	10.21	10.22	14.41	12.89	10.35	10.83	11.03	10.52
SR	0.12	0.03	0.14	0.08	0.08	0.09	0.10	0.17	0.13	0.16	0.17
Utility	0.44	0.15	0.65	0.39	0.39	0.35	0.59	0.82	0.54	0.68	**0.83**
						$\gamma = 10$					
Mean	9.94	3.99	7.16	5.07	5.14	6.21	6.23	8.05	6.82	6.98	8.15
St dev	15.63	7.30	6.91	6.12	6.11	9.75	9.25	7.21	7.37	6.35	7.31
SR	0.12	0.02	0.15	0.07	0.07	0.08	0.08	0.18	0.13	0.15	0.18
Utility	-0.11	0.15	0.56	0.34	0.35	0.25	0.44	**0.69**	0.43	0.53	**0.69**

The table presents in panel A the mean square prediction error (MSPE), the square bias ($BIAS^2$), and the variance of the prediction errors (VAR) for the individual models and combination schemes in forecasting the S&P500 index; in panel B the average portfolio return and standard deviation (both in percentage points), the Sharpe ratio (SR), and utility for various level of risk aversion coefficients, γ.

Table 4: Empirical application - Transaction costs

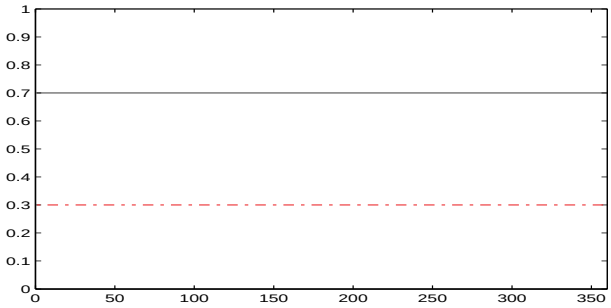
	Individual Models			Strategies							
Statistics	RW	LF	HI	1	2	3	4	5	6	7	8
						$c = 0.1\%$					
Mean	9.93	4.10	8.62	6.01	6.06	8.06	7.86	9.38	8.16	9.15	9.51
St dev	15.63	10.12	10.98	10.18	10.20	14.40	12.89	10.34	10.81	11.02	10.50
SR	0.12	0.01	0.13	0.07	0.07	0.09	0.10	0.16	0.12	0.15	0.16
Utility	0.44	0.11	0.62	0.34	0.35	0.33	0.56	0.78	0.49	0.64	**0.79**
						$c = 0.5\%$					
Mean	9.89	2.44	7.37	4.26	4.30	7.34	6.83	7.87	6.43	7.47	8.03
St dev	15.62	10.01	10.90	10.11	10.13	14.36	12.90	10.30	10.76	10.98	10.46
SR	0.12	-0.03	0.10	0.02	0.02	0.08	0.07	0.12	0.08	0.10	0.12
Utility	0.44	-0.06	0.50	0.17	0.17	0.26	0.46	0.63	0.32	0.47	**0.64**
						$c = 1\%$					
Mean	9.84	0.38	5.79	2.07	2.09	6.43	5.54	5.98	4.26	5.36	6.19
St dev	15.62	9.92	10.85	10.04	10.07	14.33	12.92	10.29	10.74	10.97	10.46
SR	0.12	-0.09	0.06	-0.04	-0.04	0.06	0.04	0.07	0.02	0.05	0.07
Utility	0.43	-0.26	0.34	-0.05	-0.05	0.17	0.33	0.44	0.10	0.26	**0.46**

The table presents the average portfolio return and standard deviation (both in percentage points), the Sharpe ratio (SR), and utility for various level of transaction costs c and coefficient of risk aversion $\gamma = 5$.

Table 5: Simulation design in exercises BI-BIII

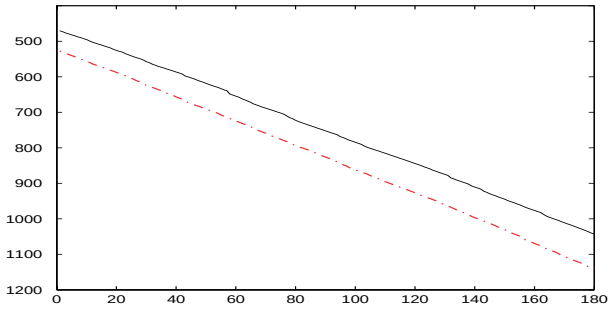
EXERCISES	I	II	III
μ_z	0.00	0.00	0.00
μ_u	0.00	0.00	0.00
μ_ξ	0.00	0.00	0.00
s^2	1.00	1.00	1.00
σ_z^2	1.00	1.00	1.00
σ_ξ^2	0.00	0.04	1.00

Figure 1: Exercise I (1)



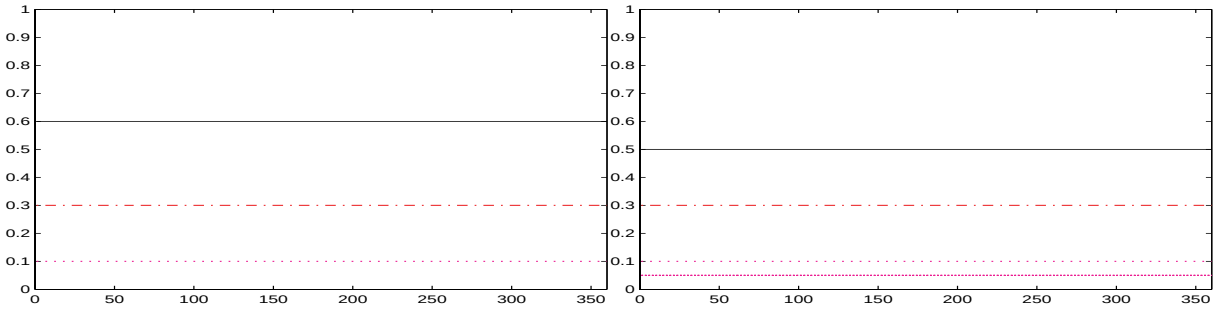
Note: The figure presents the patterns of parameters c_1 (in solid line) and c_2 (in dotted line) in equation (3) in exercises I.

Figure 2: Exercise I (2)



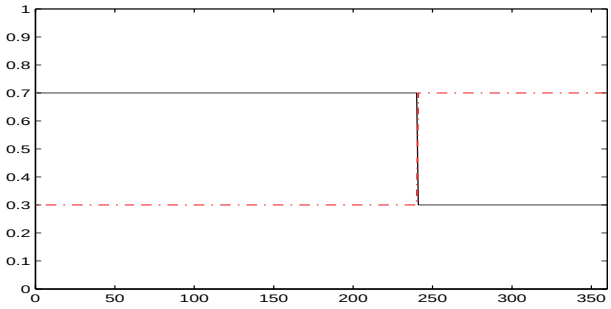
Note: The figure presents the log marginal likelihood given model 1 (in solid line) and the log marginal likelihood given model 2 (in dotted line) in exercise I.

Figure 3: Exercise VI-VII



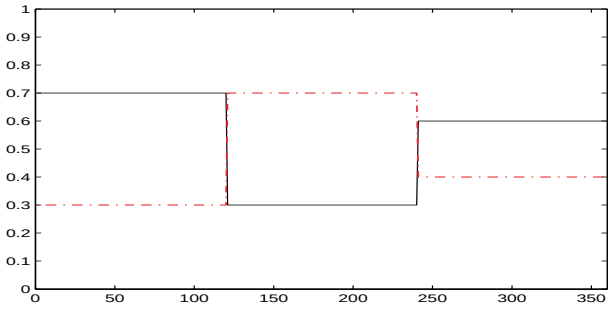
Note: The figures present in the left panel the patterns of parameters c_1 (- line), c_2 (-. line), c_3 (.. line) in equation (3) in exercises VI, and in the right panel also the parameters c_4 and c_5 (– line) in exercise VII.

Figure 4: Exercise VIII



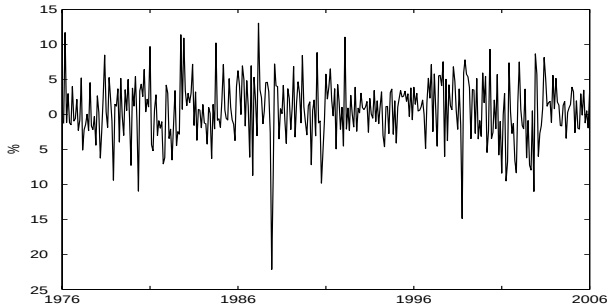
Note: The figure presents the patterns of parameters c_1 (in solid line) and c_2 (in dotted line) in equation (3) in exercises VIII.

Figure 5: Exercise IX



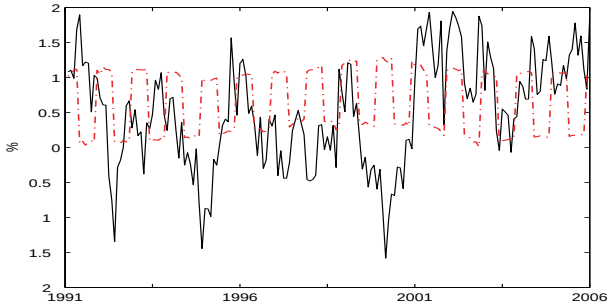
Note: The figure presents the patterns of parameters c_1 (in solid line) and c_2 (in dotted line) in equation (3) in exercises IX.

Figure 6: S&P500 Excess returns



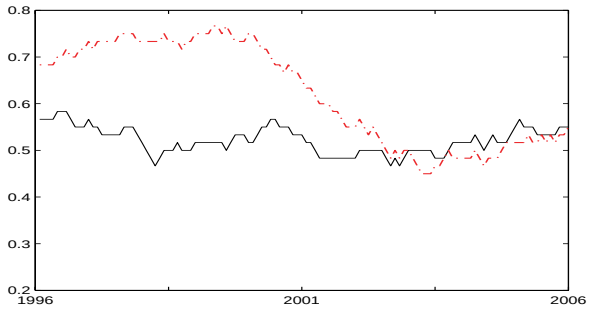
Note: The figure presents the excess returns on the S&P500 over the sample 1976:1-2005:12.

Figure 7: Individual forecasts



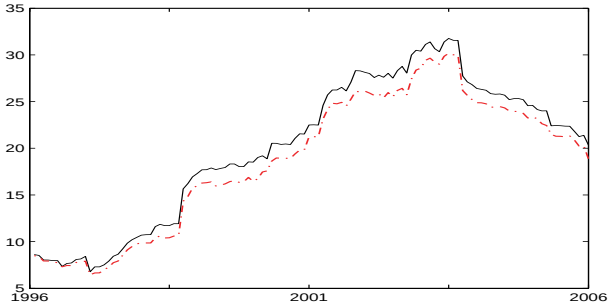
Note: The figure presents the forecasts on excess returns on the S&P500 given by the individual models 'Leading Indicator' (in solid line) and 'Halloween indicator' (in dotted line) over the sample 1996:1-2005:12.

Figure 8: 60 month moving average sign hit ratios



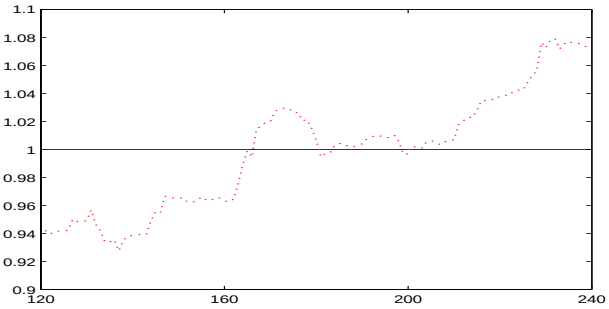
Note: The figure presents the 60 month moving average sign hit ratios given by the individual models 'Leading Indicator'(in solid line) and 'Halloween indicator' (in dotted line).

Figure 9: 60 month moving average MSPE



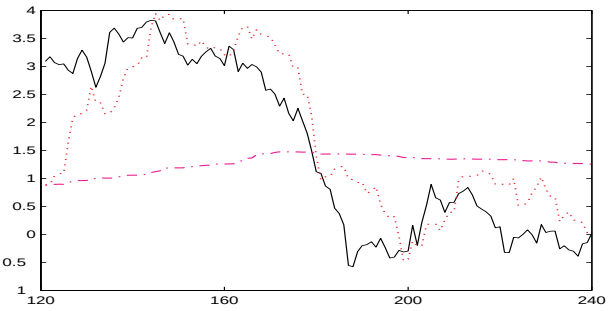
Note: The figure presents the 60 month moving average MSPE given by the individual models 'Leading Indicator'(in solid line) and 'Halloween indicator' (in dotted line).

Figure 10: Exercise B.I



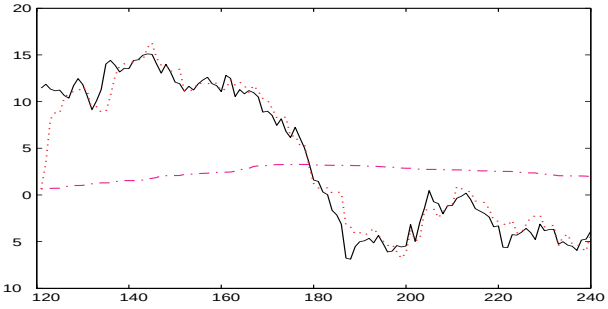
Note: The figure presents the patterns of parameter β (in - line), and estimates $\widehat{\beta}^{(4)}$ (in -. line) and $\widehat{\beta}^{(5)}$ (in .. line) in exercises B.I.

Figure 11: Exercise B.II



Note: The figure presents the patterns of parameter β (in - line), and estimates $\widehat{\beta}^{(4)}$ (in -. line) and $\widehat{\beta}^{(5)}$ (in .. line) in exercises B.II.

Figure 12: Exercise B.III



Note: The figure presents the patterns of parameter β (in - line), and estimates $\widehat{\beta}^{(4)}$ (in -. line) and $\widehat{\beta}^{(5)}$ (in .. line) in exercises B.III.