



BANK OF CANADA
BANQUE DU CANADA

Working Paper/Document de travail
2008-34

Combining Canadian Interest-Rate Forecasts

by David Jamieson Bolder and Yuliya Romanyuk

Bank of Canada Working Paper 2008-34

September 2008

Combining Canadian Interest-Rate Forecasts

by

David Jamieson Bolder and Yuliya Romanyuk

Financial Markets Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
yromanyuk@bankofcanada.ca

Bank of Canada working papers are theoretical or empirical works-in-progress on subjects in economics and finance. The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada.

Acknowledgements

We thank Scott Hendry, Greg Tkacz, Greg Bauer, Chris D'Souza, and Antonio Diez de los Rios from the Bank of Canada; Francesco Ravazzolo from the Norges Bank; Michiel de Pooter from the Econometric Institute, Erasmus University Rotterdam; and David Dickey from North Carolina State University. We retain any and all responsibility for errors, omissions, and inconsistencies that may appear in this work.

Abstract

Model risk is a constant danger for financial economists using interest-rate forecasts for the purposes of monetary policy analysis, portfolio allocations, or risk-management decisions. Use of multiple models does not necessarily solve the problem as it greatly increases the work required and still leaves the question “which model forecast should one use?” Simply put, structural shifts or regime changes (not to mention possible model misspecifications) make it difficult for any single model to capture all trends in the data and to dominate all alternative approaches. To address this issue, we examine various techniques for *combining* or *averaging* alternative models in the context of forecasting the Canadian term structure of interest rates using both yield and macroeconomic data. Following Bolder and Liu (2007), we study alternative implementations of four empirical term structure models: this includes the Diebold and Li (2003) approach and three associated generalizations. The analysis is performed using more than 400 months of data ranging from January 1973 to July 2007. We examine a number of model-averaging schemes in both frequentist and Bayesian settings, both following the literature in this field (such as de Pooter, Ravazzolo and van Dijk (2007)) in addition to introducing some new combination approaches. The forecasts from individual models and combination schemes are evaluated in a number of ways; preliminary results show that model averaging generally assists in mitigating model risk, and that simple combination schemes tend to outperform their more complex counterparts. Such findings carry significant implications for central-banking analysis: a unified approach towards accounting for model uncertainty can lead to improved forecasts and, consequently, better decisions.

JEL classification: C11, E43, E47

Bank classification: Interest rates; Econometric and statistical methods

Résumé

Le risque de modèle présente un écueil constant pour les économistes financiers qui appuient leurs analyses de la politique monétaire, leurs choix de portefeuille ou leur gestion du risque sur des prévisions de taux d'intérêt. Le recours à de multiples modèles de prévision ne résout pas nécessairement le problème puisqu'il alourdit grandement les calculs et ne nous dit pas quelle prévision retenir. En un mot, les changements structurels ou de régime (sans oublier les erreurs de spécification possibles) font qu'il est difficile de représenter toutes les tendances qui se dégagent des données à l'aide d'un modèle unique, susceptible de dominer tous les autres. Pour y voir plus clair, les auteurs examinent diverses méthodes qui consistent à *combinaison*, en les pondérant, les prévisions qu'ils obtiennent au sujet de la structure des taux d'intérêt canadiens à partir de

différents modèles estimés au moyen de données relatives aux rendements et aux variables macroéconomiques. Conformément à l'approche de Bolder et Liu (2007), les auteurs étudient plusieurs façons de combiner quatre modèles empiriques de la structure des taux, soit le modèle de Diebold et Li (2003) et trois généralisations associées. Leur analyse met à contribution plus de 400 observations mensuelles allant de janvier 1973 à juillet 2007. Un certain nombre des combinaisons examinées relèvent des cadres fréquentiste et bayésien et s'inspirent de la littérature dans ce domaine (p. ex., Pooter, Ravazzolo et van Dijk, 2007), et d'autres sont nouvelles. Les prévisions tirées des modèles pris isolément et des combinaisons de modèles sont évaluées de différentes façons. Les résultats préliminaires montrent que le fait de combiner plusieurs modèles contribue en règle générale à réduire le risque de modèle, et que les schémas les plus simples tendent à donner de meilleures prévisions que les plus complexes. Ces résultats ont des conséquences intéressantes du point de vue des banques centrales : une approche unifiée de prise en compte de l'incertitude des modèles pourrait aboutir à des prévisions améliorées et, partant, à des décisions plus éclairées.

Classification JEL : C11, E43, E47

Classification de la Banque : Taux d'intérêt; Méthodes économétriques et statistiques

1 Introduction and motivation

Model risk is a real concern for financial economists using interest-rate forecasts for the purposes of monetary policy analysis, strategic portfolio allocations, or risk-management decisions. The issue is that one's analysis is always conditional upon the model selected to describe the uncertainty in the future evolution of financial variables. Moreover, using an alternative model can, and does, lead to different results and possibly different decisions. Selecting a single model is challenging because different models generally perform in varying ways on alternative dimensions, and it is rare that a single model dominates along all possible dimensions.

One possible solution is the use of multiple models. This has the advantage of diversifying away, to a certain extent, the model risk inherent in one's analysis. It does, however, have some drawbacks. First of all, it is time consuming insofar as one must repeat one's analysis with each alternative model. In the event one uses a simulation-based algorithm, for example, this can also substantially increase one's computational burden. A second drawback relates to the interpretation of the results in the context of multiple models. In the event that one employs n models, there will be n separate sets of results and a need to determine the appropriate weight to place on these n separate sets of results. The combination of these two drawbacks reduces the appeal of employing a number of different models.

Perhaps a better approach, that has some theoretical and empirical support, involves combining, or averaging, a number of alternative models to create a single combined model. This is not a new idea. The concept of model averaging has a relatively long history in the forecasting literature. Indeed, there is evidence dating back to Bates and Granger (1969) and Newbold and Granger (1974) suggesting that combination forecasts often outperform individual forecasts. Possible reasons for this are that the models may be incomplete, they may employ different information sets, and they may be biased. Combining forecasts, therefore, acts to offset this incompleteness, biasedness, and variation in information sets. Combined forecasts may also be enhanced by the covariances between individual forecasts. Thus, even if misspecified models are combined, the combination may, and often will, improve the forecasts (Kapetanios, Labhard and Price (2006)).

Another motivation for model averaging involves the combination of large sets of data. This application is particularly relevant in economics, where there is a literature describing management of large numbers of explanatory variables through factor modelling (see, for example, Moench (2006) and Stock and Watson (2002)). We can also combine factor-based models to enrich the set of information used to generate forecasts, as suggested in Koop and Potter (2003) in a Bayesian framework. There is vast literature on Bayesian model averaging; for a good tutorial on Bayesian model averaging, see Hoeting et al. (1999). Draper (1995) is also a useful reference. A number of papers investigate the predictive performance of models combined in a Bayesian setting and find that there are accuracy and economic gains from using combined forecasts (for example, Andersson and Karlsson (2007), Eklund and Karlsson (2007), Ravazzolo, van Dijk and Verbeek (2007), and de Pooter, Ravazzolo and van Dijk (2007)).

However, model averaging is not confined to the Bayesian setting. For example, Diebold and Pauly (1987) and Hendry and Clements (2004) find that combining forecasts adds value in the presence of structural breaks in the frequentist setting. Kapetanios, Labhard and Price (2005) use a frequentist information-theoretic approach for model combinations and show that it can be a powerful alternative to both Bayesian and factor-based methods. Likewise, in a series of experiments Swanson and Zeng (2001) find that combinations based on the Schwartz

Information Criterion perform well relative to other combination methods. Simulation results in Li and Tkacz (2004) suggest that the general practice of combining forecasts, no matter what combination scheme is employed, can yield lower forecast errors on average.

It appears, therefore, that there is compelling evidence supporting the combination of multiple models as well as a rich literature describing alternative combination algorithms. This paper attempts to explore the implications for the aforementioned financial economist working with multiple models of Canadian interest rates. This work asks, and attempts to answer, a simple question: does model averaging work in this context and, if so, which approach works best and most consistently? While the model averaging literature finds its origins in Bayesian econometrics, our analysis considers both frequentist and Bayesian combination schemes. Moreover, the principal averaging criterion used in determining how the models should be combined is their out-of-sampling forecasting performance. Simply put, we generally require that the weight on a given model should be larger for those models that forecast better out of sample. This is not uniformly true across the various forecasting algorithms, but it underpins the logic behind most of the nine combination algorithms examined in this paper.

The rest of the paper is organized in four main parts. In Section 2, we describe the underlying interest-rate models and review their out-of-sample forecasting performance. Next, in Section 3, we describe the alternative combination schemes. Section 4 evaluates the performance of the different model averaging approaches when applied to Canadian interest-rate data, and Section 5 concludes.

2 Models

The primary objective of this paper is to investigate whether combined forecasts improve the accuracy of out-of-sample Canadian interest-rate forecasts. The first step in attaining this objective is to introduce, describe and compare the individual interest-rate models that we will be combining. Min and Zellner (1993) point out that if models are biased, combined forecasts may perform worse than individual models. Consequently, it is critically important to appraise the models and their forecasts carefully before combining them. The models used in this work are empirically motivated from previous work in this area. In particular, Bolder (2006) and Bolder and Liu (2007) investigate a number of models, including affine (see, for example, Dai and Singleton (2000), Duffie, Filipovic and Schachermayer (2003), Ang and Piazzesi (2003)), in which pure-discount bond prices are exponential-affine functions¹ of the state variables, and empirical-based (such as those in Bolder and Gusba (2002) and the extension of the Nelson-Siegel model by Diebold and Li (2003)). The results indicate that forecasts of affine term-structure models are inferior to those of empirically-motivated models.

Out of these models, we choose those with the best predictive ability, in the hope that their combinations will further improve term-structure forecasts. The four models examined in this paper, therefore, are the Nelson-Siegel (NS), Exponential Spline (ES), Fourier Series (FS) and a state-space approach (SS). It should be stressed that none of these models are arbitrage-free; in our experience, the probability of generating zero-coupon rate forecasts that admit arbitrage is very low.² An attractive feature of the selected models is that they allow us to easily incorporate macroeconomic factors into our analysis of the term structure, assuming a

¹More complex mappings are considered by Leippold and Wu (2000), Cairns (2004), among others.

²If such outcomes occur, there are a number of possible solutions. For example, one could substitute for the arbitrage forecast the previous forecast or some combination of previous forecasts.

unidirectional effect from macroeconomic factors to the term structure. This has a documented effect of increasing forecasting efficiency. We do not model feedback between macro and yield factors, since Diebold, Rudebusch and Aruoba (2006) and Ang, Dong and Piazzesi (2007) find that the causality from macroeconomic factors to yields is much higher than that from yields to macro factors.

The models have the following basic structure:

$$\begin{aligned} Z(t, \tau) &= G(t, \tau)Y_t, \\ Y_t &= C + \sum_{l=1}^L F_l Y_{t-l} + \nu_t, \quad \nu_t \sim N(0, \Omega). \end{aligned} \tag{1}$$

Here $Z(t, \tau)$ denotes the zero-coupon rate at time t for maturity τ , $(\tau - t)$ the term to maturity, and G the mapping from state variables (factors) Y to zero-coupon rates. We model the vector Y_t by a VAR(L) with $L = 2$, which we find works best for our purposes. For ES and FS models, $Z(t, \tau) = -\frac{\ln(P(t, \tau))}{\tau - t}$ and $P(t, \tau) = \sum_{k=1}^n Y_{k,t} g_k(\tau - t)$, where $P(t, \tau)$ is the price of a zero-coupon bond at t for maturity τ . In the ES model, $g_k(\tau - t)$ are orthogonalized exponential functions; in the FS model, they are trigonometric basis functions (see Bolder and Gusba (2002) for details).

For all models except SS, we find the factors Y_t at each time t by minimizing the square distance between $P(t, \tau)$ above and the observed bond prices. We augment the factors with three macroeconomic variables—the output gap x_t , consumer price inflation π_t , and the overnight rate r_t —and collect these to form a time series. This procedure and the estimation of model-specific parameters for the NS, ES and FS models are given in Bolder and Liu (2007) and the references therein. In the SS model, we simply regress the vector of zero-coupon rates Z_t on the first three principal components, extracted from the observed term structure up to time t , and the three contemporaneous macro variables. Note that only the SS model allows for a direct connection between the macro factors and the zero-coupon rates. In the other three models, only the term-structure factors determine the yields or bond prices: in the mapping from state variables to bond prices or zero-coupon rates, the coefficients for macro factors are set to zero.³

2.1 A few words about Bayesian framework

The task of selecting appropriate parameters for the prior distributions is not a trivial one, and a number of papers discuss this issue (see, for instance, Litterman (1986), Kadiyala and Karlsson (1997), Raftery, Madigan and Hoeting (1997), Fernandez, Ley and Steel (2001)). We have tried a variety of specifications, including those in the references above as well as some calibrated ones. We have found that for our purposes, the g-prior (Zellner (1986)) appears to produce the most satisfactory results. We estimate the parameters for the g-prior from the in-sample data. While this may not be the most optimal way to estimate a prior distribution, and ideally we would like to set aside a part of our data just for this purpose, we are constrained by the length of the available time series. First, we have to forecast for relatively long horizons and thus set aside a large proportion of the time series for the out-of-sample testing. Second, we have to leave some part of the time series to train model combinations. Third, our models are multidimensional and require a sizeable portion of the data just for estimation. Finally, it

³Using the state-space (Diebold, Rudebusch and Aruoba (2006)) adaptation of the Nelson-Siegel model, de Pooter, Ravazzolo and van Dijk (2007) account for the effects of macro variables in a similar manner.

is difficult to have a strong independent (from observed data) prior belief about the behaviour of parameters in high-dimensional models. For these reasons, we estimate the g-prior and the posterior distribution using the same in-sample data.

While our models have the general structure of state-space models, there are differences. We assume that zero-coupon rates Z in observation equations are observed without error for all models except the SS. To estimate the models in a full Bayesian setup, we could have introduced an error term in each of these equations and then we would have had to use a filter to extract the unobserved state variables Y . However, because FS and ES models are highly nonlinear (and the dimensions of the corresponding factors are high), such a procedure would be very computationally heavy and may not be optimal.⁴ Instead of this, we take the state variables as given (from Bolder (2006)) and estimate the transition VAR(2) equations in the Bayesian framework for each of the models. This facilitates computations greatly, because we can use existing analytic results for VAR(L) models. Please see the Appendix for more details about Bayesian estimation of VAR(L) models.

We use transition equations to determine weights for Bayesian model averaging schemes. For consistency with the other models, we compute the weights based on the transition equation of the SS model, even though the observation equation for the SS model is a regression with an error term. Technically speaking, this approach does not give proper Bayesian posterior model probabilities for the four models that are competing to explain the observed term structure, since the data y has to be the same (the same observed zero-coupon rates Z) and the explanatory variables different depending on the model \mathcal{M}_k . In our case, the y data differs for each transition equation: it is the NS, ES, FS or SS factors. So in effect we are assigning weights to each model in the forecast combination based on how well the transition equations capture the trends in the underlying factors of each model. In light of our assumption that observation equations do not contribute any new information since they have no error term,⁵ this approach appears reasonable.

2.2 Forecasts of individual models

In practice, we do not observe zero-coupon rates. We do not even observe prices of pure-discount bonds. We must use the observed prices of coupon-bearing bonds and some model for zero-coupon rates to extract the zero-coupon term structure. A number of alternative approaches for extracting zero-coupon rates from government bond prices are found in Bolder and Gusba (2002). Figure 1 shows the Canadian term structure of zero-coupon rates from January 1973 to August 2007. As in many industrialized economies, the Canadian term structure is characterized by periods of high volatile rates in the late 1980s and the 1990s. Moreover, starting in 2005, the term structure becomes rather flat. Any single model will generally have difficulties describing and forecasting both volatile and stable periods equally well.

To evaluate the forecasts of the four models, we use monthly data for bond prices for different tenors and macroeconomic variables (output gap, consumer price inflation, and overnight rate) from January 1973 to August 2007. This constitutes 416 observations. We take the first 120

⁴de Pooter, Ravazzolo and van Dijk (2007) discuss issues that arise in the Bayesian inference of affine models, whose parameters are highly nonlinear, similarly to our models.

⁵While some may argue that such assumption is not realistic, we feel that it is justified by the tangible benefits of greatly reduced estimation complexity and computational effort. We think that such benefits would not be outweighed by the advantages from introducing error into the observation equations to make the already stylized models more realistic.

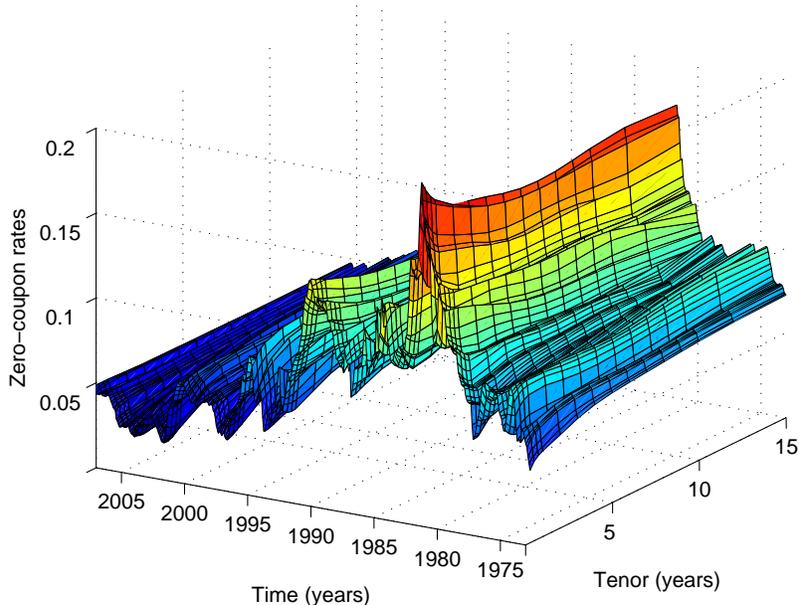


Figure 1: Zero-Coupon Rates from January 1973 to August 2007, extracted from Government of Canada treasury bill and nominal bond prices using a nine-factor exponential spline model described in Bolder and Gusba (2002).

points as our initial in-sample estimation data. Once the models are estimated, we make out-of-sample interest rate forecasts for horizons $h = 1, 12, 24, 36$ months at time $T = 120$ (the information set up to time T will be denoted by filtration \mathcal{F}_T). Next, for each model \mathcal{M}_k , $k = 1, \dots, 4$, we evaluate the vector of N tenors of forecasted zero-coupon rates $\hat{Z}_{T+h} = \mathbb{E}(Z_{T+h} | \mathcal{F}_T, \mathcal{M}_k)$ against the actual zero-coupon rates Z_{T+h} , $N \times 1$, extracted from observed bond prices:

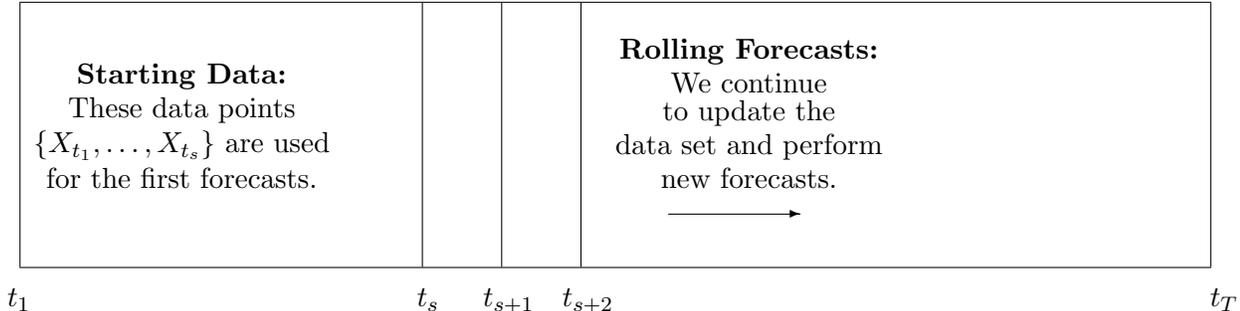
$$e_{T+h}^{\mathcal{M}_k} = \sqrt{\left(\frac{\left(Z_{T+h} - \hat{Z}_{T+h}^{\mathcal{M}_k} \right)' \left(Z_{T+h} - \hat{Z}_{T+h}^{\mathcal{M}_k} \right)}{N} \right)}. \quad (2)$$

A schematic describing the various steps in the determination of these overlapping forecasts is found in Figure 2.

We subsequently re-estimate each model for each $T \in [121, 416 - h]$ in-sample points, calculating the corresponding forecast errors for each model. Figure 3 shows the root mean squared deviations between the actual and forecasted zero-coupon rates relative to the errors from random walk forecasts using a rolling window of 48 observations.⁶ We include the RMSE for the random walk model as a reference because, in the term-structure literature, it is frequently used

⁶The random walk is scaled to one. Consequently, values higher than one imply worse, and lower than one better, performance than the random walk. We opt for graphs with relative root mean squared forecast errors as opposed to the commonly reported tables with the same information, because we have found graphs easier to read.

Figure 2: **Forecasting Interest Rates:** This schematic describes the steps involved in generating rolling interest-rate forecasts, which in this work, act as the principal input for the parametrization of our model-averaging schemes.



0. Set $i = s$ and $k = 1$;
1. Formulate $\mathbb{E}_{\mathcal{M}_k}(Z_{t_{i+h}} | \mathcal{F}_{t_i})$;
2. Observe $Z_{t_{i+h}}$;
3. Compute $\epsilon_{t_{i+h}}^{\mathcal{M}_k} = Z_{t_{i+h}} - \mathbb{E}_{\mathcal{M}_k}(Z_{t_{i+h}} | \mathcal{F}_{t_i})$;
4. Repeat steps 1-3 for $k = 2, \dots, n$ models;
5. Repeat steps 1-4 for $i = s + 1, \dots, T - h$ observations.
6. Repeat steps 1-5 for $h = 1, \dots, H$ months.

as a benchmark model and it is not easy to beat, at least for affine models (see, for example, Duffee (2002) and Ang and Piazzesi (2003)). Note that the forecasts of the random walk are just the last observed zero-coupon rates.

From Figure 3, we observe that for all horizons, there are periods when the models outperform the random walk, but none of the models seem to outperform the random walk on average (over the sample period). As one would expect, the forecasting performance of all four models deteriorates as the forecasting horizon increases. For horizons beyond one month, all models have difficulties predicting interest rates during the period of high interest rates in the early 1990s. The models also struggle to capture the flat term structure observed in the early 2000s; however, the FS and the ES models appear to be more successful at this than the NS and the SS models. While all models perform similarly for the short-term horizon, certain patterns emerge at longer horizons: the NS and SS models tend to move together, as do the FS and ES models. This result is confirmed in Figure 4: the correlation between forecast errors from the NS and SS models is very close to one beyond six months. The correlation between the ES and FS models is also quite high.

The heterogeneity between the models is a strong motivating factor for model averaging. In particular, it suggests that there is some potential for combining models to complement the information carried by each model and thereby produce superior forecasts.

Figure 5 shows the performance of our models estimated in the Bayesian setting relative to the random walk. Comparing with Figure 3, we see that Bayesian forecasts are virtually identical to frequentist forecasts. We do not test whether the Bayesian forecasts are statistically significantly different from the frequentist ones, since we are not comparing frequentist vs. Bayesian estimation methods. We estimate the models in the Bayesian setting only because we

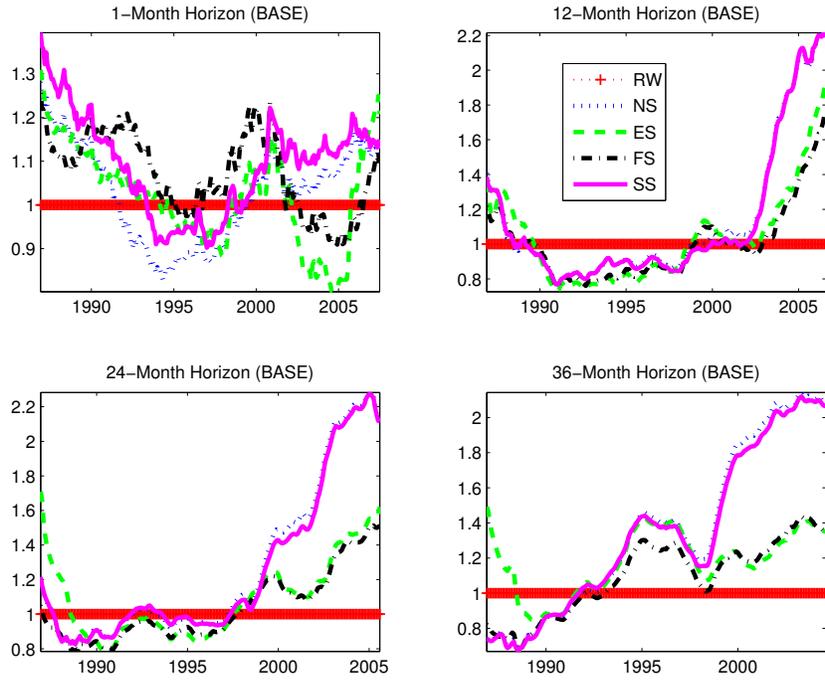


Figure 3: Predictive performance for frequentist forecasts relative to random walk.

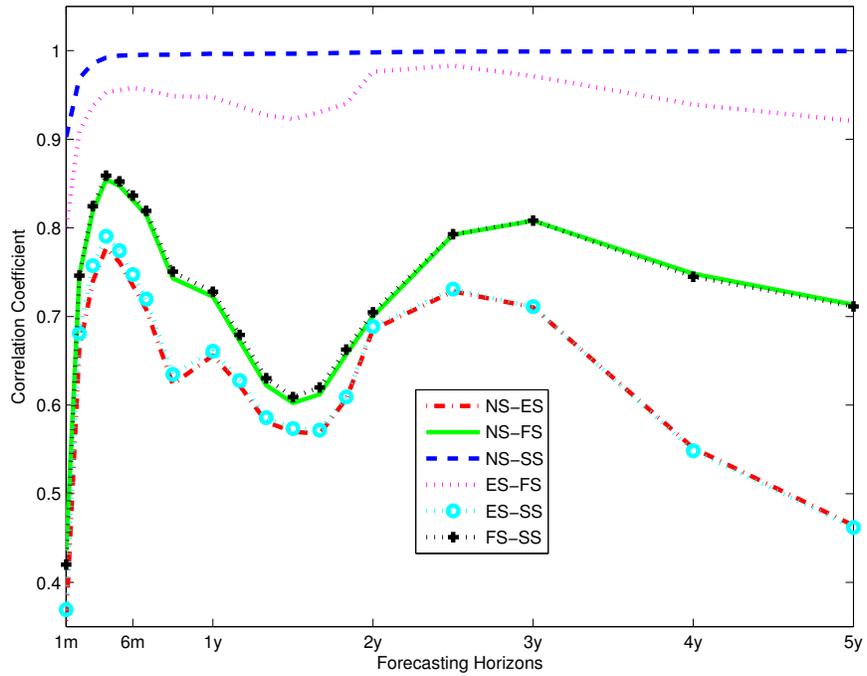


Figure 4: Correlation structure of frequentist forecasts for different horizons.

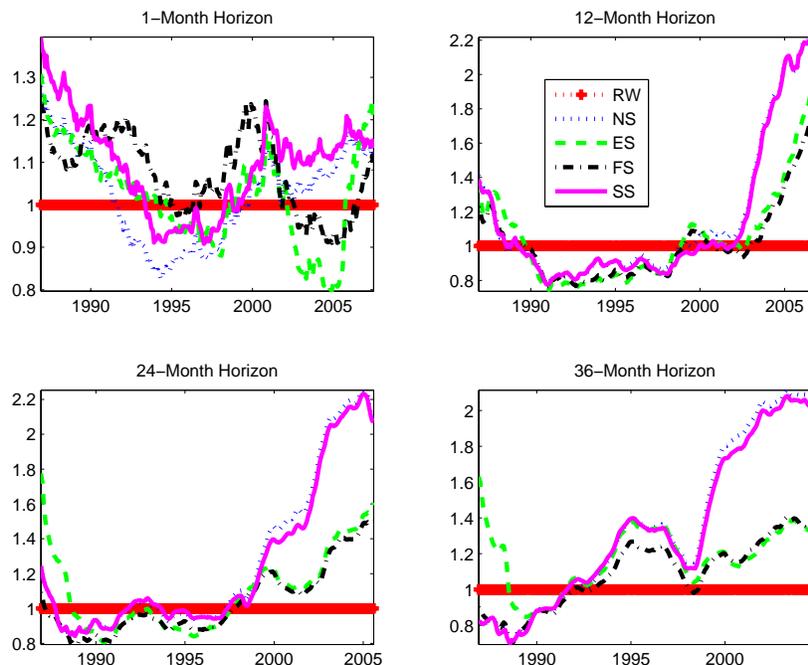


Figure 5: Predictive performance for Bayesian forecasts relative to random walk.

need Bayesian forecast distributions to obtain weights for Bayesian model averaging schemes.

3 Model combinations

In this work, we investigate nine alternative model combination schemes, which we denote $\mathcal{C}_1 - \mathcal{C}_9$. They are Equal Weights, Inverse Error, Simple OLS, Factor OLS, MARS, Predictive Likelihood, Marginal Model Likelihood, Log Predictive Likelihood, and Log Marginal Model Likelihood. We refer to the first five schemes as *ad-hoc*, and the last four as *Bayesian*.⁷ Our goal is to calculate weights for each model \mathcal{M}_k , horizon h , and combination \mathcal{C}_j : $w_{k,h}^{\mathcal{C}_j}$, $k = 1, \dots, 4$, $j = 1, \dots, 9$, $h = 1, 12, 24, 36$ months. Conceptually, therefore, different model averaging schemes merely amount to alternative methods for determining the amount of weight (i.e., the ω 's) to place on each individual forecast.

Models estimated in the frequentist setting produce point forecasts, whereas in the Bayesian setting we obtain forecast densities. There are two approaches to combine Bayesian forecasts: the first refers to averaging the individual densities directly (Mitchell and Hall (2005), Hall and Mitchell (2007), and Kapetanios, Labhard and Price (2005)), while the second to combining the *moments* of individual densities (Clyde and George (2004)). For example, as indicated in the latter article, a natural point prediction at time T for a zero-coupon rate vector h -steps

⁷The difference between the two types of schemes is that ad-hoc combinations can be applied to forecasts generated in either frequentist or Bayesian setting, whereas Bayesian combination schemes should be applied to Bayesian forecasts.

ahead is

$$\hat{Z}_{T+h} = \mathbb{E}(Z_{T+h}|\mathcal{F}_T) = \sum_{k=1}^4 w_{k,h}^{C_j} \mathbb{E}(Z_{T+h}|\mathcal{F}_T, \mathcal{M}_k) = \sum_{k=1}^4 w_k \hat{Z}_{T+h}^{\mathcal{M}_k}, \quad (3)$$

where $\hat{Z}_{T+h}^{\mathcal{M}_k}$ are the means of individual forecast densities.

3.1 C_1 : Equal Weights

This is the simplest possible combination scheme. Each individual forecast receives an equal weight as follows

$$w_{k,h}^{C_1} = \frac{1}{n}. \quad (4)$$

While Equal Weights combination is very simple, it is a standard benchmark for the evaluation of alternative model-averaging algorithms precisely because it performs quite well relative to individual forecasts and more complicated schemes (see, for example, Hendry and Clements (2004) and Timmermann (2006)).

3.2 C_2 : Inverse Error

In this combination scheme, we assign higher weights to models whose forecasts perform better out of sample. We set aside M points from our sample to evaluate the predictive performance of each model, and then we average the forecast errors over these M points. More specifically, we estimate the models using $T = 120$ initial points, make h -step forecasts and evaluate each model's performance by calculating the forecast error (2). Then we repeat these steps for each $T \in [121, 120 + M - h]$. This procedure yields $M - h + 1$ forecast errors, which we average. The resulting weights are given by

$$w_{k,h}^{C_2} = \frac{1 / \left(\sum_{T=120}^{120+M-h} e_{T+h}^{\mathcal{M}_k} / (M - h + 1) \right)}{\sum_{k=1}^4 \left[1 / \left(\sum_{T=120}^{120+M-h} e_{T+h}^{\mathcal{M}_k} / (M - h + 1) \right) \right]}. \quad (5)$$

This combination scheme is also simple, but it differs from the Equal-Weights approach in that it requires data. We use M observations to train the weights for this and all subsequent model combinations depending on the evaluation approach. Indeed, the Equal Weights combination is the only technique that does not require a training period.

3.3 C_3 : Simple OLS

Here we combine the forecasts from individual models using simple OLS regression coefficients as weights. First, we estimate the models and make h -step forecasts for each $T \in [120, 120 + M - h]$. We treat these $M - h + 1$ forecasts $\hat{Z}_{T+h}^{\mathcal{M}_k}$ as realizations of four predictor variables, and for each tenor $i \in [1, N]$, we regress⁸ the actual zero-coupon rates Z_{T+h} against these individual

⁸This can be done with or without the intercept $\beta_{0,h}$ and/or forcing $\beta_{k,h}$ to add up to one. We have found (in studies unreported here) that unconstrained regression without an intercept works best in our case.

forecasts for the respective tenor i :

$$Z_{T+h}(i) = \beta_{0,h}(i) + \sum_{k=1}^4 \beta_{k,h}(i) \hat{Z}_{T+h}^{\mathcal{M}_k}(i). \quad (6)$$

The weights for the simple OLS scheme are given by

$$w_{k,h}^{\mathcal{C}_3}(i) = \beta_{k,h}(i). \quad (7)$$

This type of combination scheme is very flexible, since the weights are unconstrained. What this implies is that one can place negative weights on certain forecasts and significant positive weights on other forecasts. As a consequence of this flexibility, this approach turns out to be our best-performing combination. Its flexibility is not, however, without a cost since we find the approach can be sensitive to the training period. We discuss these points later in the discussion.

3.4 \mathcal{C}_4 : Factor OLS

A drawback of the simple OLS scheme is that we estimate the weights separately for a set of pre-specified zero-coupon tenors and then interpolate for the remaining tenors. This leads to a fairly large number of regressions. To reduce the number of parameters, therefore, we construct a lower-dimensional alternative, which we term the factor OLS scheme.

First, we perform a basic decomposition of the zero-coupon term structure as follows:

$$\underbrace{Y_t(1)}_{\text{Level}} = Z_{t,15\mathbf{y}}, \quad \underbrace{Y_t(2)}_{\text{Slope}} = Z_{t,15\mathbf{y}} - Z_{t,3\mathbf{m}}, \quad \underbrace{Y_t(3)}_{\text{Curve}} = 2Z_{t,2\mathbf{y}} - (Z_{t,3\mathbf{m}} + Z_{t,15\mathbf{y}}). \quad (8)$$

Here $3\mathbf{m}$, $2\mathbf{y}$ and $15\mathbf{y}$ refer to the 3-month bill, and 2- and 15-year bonds respectively. Clearly, this approach is motivated by the notions of well-known level, slope and curvature variables stemming from principal components analysis.

Now we have only three components from which we build the term structure of zero-coupon yields. To obtain the OLS weights, we regress⁹ the actual l -th factor $Y_{T+h}(l)$, $l = 1, 2, 3$, on the factors forecasted by each model, $Y_{T+h}(l)^{\mathcal{M}_k}$:

$$Y_{T+h}(l) = \beta_{0,h}(l) + \sum_{k=1}^4 \beta_{k,h}(l) \hat{Y}_{T+h}^{\mathcal{M}_k}(l). \quad (9)$$

The weights for the factor OLS scheme are

$$w_{k,h}^{\mathcal{C}_4}(l) = \beta_{k,h}(l). \quad (10)$$

Once we have the combined forecasted factors $\hat{Y}_{T+h}(l)$, we invert the decomposition itera-

⁹As with the simple OLS combination scheme, we can do this with or without an intercept or forcing the coefficients to add up to one, and we obtain better results for the specification with no intercept and no restrictions.

tively as follows:

$$Z_{t,15y} = Y_t(1), \quad Z_{t,3m} = Y_t(1) - Y_t(2), \quad Z_{t,2y} = \frac{Y_t(3) + 2Y_t(1) - Y_t(2)}{2}. \quad (11)$$

The advantage of this averaging approach is that it reduces the number of regressions and thus estimated parameters. Its disadvantage is that we are forced now to interpolate the entire curve from on only three points. In some cases, this error with such an approximation may be substantial.

3.5 \mathcal{C}_5 : MARS

The previous four schemes are relatively straightforward. For the purposes of comparison, however, we opted to include a more mathematically complex approach to combine the forecasts from individual models. The approach we selected is termed Multiple Adaptive Regression Splines (MARS), which is a function-approximation technique based on the recursive-partitioning algorithm. The basic idea behind this technique is to define piecewise linear spline functions on an overlapping partition of the domain (Bolder and Rubin (2007) provide a detailed description of the MARS algorithm). As such, the MARS combination scheme can be considered an example of a mathematically complicated nonparametric, nonlinear aggregation of our four alternative models.

The combination is trained on a set of $M + h - 1$ realized zero-coupon rates Z_{T+h} and their forecasts $\hat{Z}_{T+h}^{\mathcal{M}_k}$, $T \in [120, 120 + M - h]$, for all tenors, horizons and models. Once trained, we combine the individual forecasts according to the MARS algorithm. Note that, unlike in the previous four schemes, we cannot write the combined forecast \hat{Z}_{T+h} as a linear combination of weights $w_{k,h}^{\mathcal{C}_5}$ and individual forecasts $\hat{Z}_{T+h}^{\mathcal{M}_k}$ due to the nonlinearity and complexity of the MARS scheme.

3.6 \mathcal{C}_6 : Predictive Likelihood

In our Bayesian model averaging schemes, the weights are some version of posterior model probabilities. Theoretically, the posterior model probabilities $\mathbb{P}(\mathcal{M}_k|Y)$ are

$$\begin{aligned} \mathbb{P}(\mathcal{M}_k|Y) &= \frac{p(Y, \mathcal{M}_k)}{\underbrace{\sum_{j=1}^4 p(Y, \mathcal{M}_j)}_{p(Y)}} \quad (12) \\ &= \frac{p(Y|\mathcal{M}_k)\mathbb{P}(\mathcal{M}_k)}{\sum_{j=1}^n p(Y|\mathcal{M}_j)\mathbb{P}(\mathcal{M}_j)}. \end{aligned}$$

We think that all of the models are equally likely, so we take prior model probabilities $\mathbb{P}(\mathcal{M}_k) = \frac{1}{n}$.

The quantity $p(Y|\mathcal{M}_k)$ is the marginal model likelihood for model \mathcal{M}_k , which measures in-sample fit and fit to prior distribution only. However, out-of-sample forecasting ability is our main criterion for selecting models and evaluating model combinations (Geweke and Whiteman (2006) indicate that “a model is as good as its predictions”). This and other recent papers (for example, Ravazzolo, van Dijk and Verbeek (2007), Eklund and Karlsson (2007), Andersson and Karlsson (2007)) use predictive likelihood, which is the predictive density

evaluated at the realized value(s), instead of the marginal model likelihood, to average models in a Bayesian setting.¹⁰ Following this stream of literature to obtain the weights for combination \mathcal{C}_6 , for each model \mathcal{M}_k and horizon h , we (a) formulate $\mathbb{E}_{\mathcal{M}_k}(Y_{T+h}|\mathcal{F}_T) = \hat{Y}_{T+h}^{\mathcal{M}_k}$; (b) formulate $p(Y_T|\mathcal{M}_k, \mathcal{F}_{T-h})$; (c) observe Y_T and evaluate $p(Y_T|\mathcal{M}_k, \mathcal{F}_{T-h})$; and (d) use $p(Y_T|\mathcal{M}_k, \mathcal{F}_{T-h})$ to combine $\mathbb{E}_{\mathcal{M}_k}(Y_{T+h}|\mathcal{F}_T)$.

Substituting the predictive likelihood into (12) in place of the marginal model likelihood, we obtain the weights for the predictive likelihood combination. Similarly to the previous four combinations, we calculate the weights for each $T \in [120, 120 + M - h]$ and average the resulting $M - h + 1$ weights to get the fixed weights that will be used to evaluate model combinations out of sample:

$$w_{k,h}^{\mathcal{C}_6} = \frac{\sum_{T=120}^{120+M-h} \left(\frac{p(Y_T|\mathcal{M}_k, \mathcal{F}_{T-h})}{\sum_{j=1}^4 p(Y_T|\mathcal{M}_j, \mathcal{F}_{T-h})} \right)}{M - h + 1}. \quad (13)$$

Strictly speaking, such weights are not proper posterior model probabilities, but their advantage is measuring the out-of-sample predictive ability.

3.7 \mathcal{C}_7 : Marginal Model Likelihood

Even though marginal model likelihood evaluates in-sample fit only, we use it as one of our model combination schemes, since this is the classical Bayesian model averaging approach (see, for instance, Madigan and Raftery (1994) and Kass and Raftery (1995)). To generate a combined forecast, we calculate the marginal model likelihood $p(Y_T|\mathcal{M}_k)$ for model \mathcal{M}_k using T in-sample data points. The weight for each model is its posterior probability. Then we average the weights for each $T \in [120, 120 + M - h]$, as with previous model combinations, to obtain the weights for the marginal model likelihood combination:

$$w_k^{\mathcal{C}_7} = \frac{\sum_{T=120}^{120+M-h} \left(\frac{p(Y_T|\mathcal{M}_k)}{\sum_{j=1}^4 p(Y_T|\mathcal{M}_j)} \right)}{M - h + 1}. \quad (14)$$

Unlike with weights based on the predictive likelihood, the weights based on the marginal model likelihood do not depend on the forecasting horizon h (Figure 7).

3.8 \mathcal{C}_8 and \mathcal{C}_9 : Log Likelihood Weights

It turns out that in practice the weights based on marginal model likelihood or predictive likelihood vary significantly depending on the estimation period. This is shown in Figures 6 and 7 for observations at $T \in [120, 120 + M - h]$, $M = 120$. To obtain a smoother set of weights based on the marginal model (or predictive) likelihood, we take the logarithms of the marginal model (predictive) likelihood values and transform them linearly into weights. We want these weights w_k , $k = 1, \dots, 4$, to satisfy $w_k \in (0, 1)$, $\sum_{k=1}^4 w_k = 1$, and the relative distance between the weights should be preserved by the transformation.

¹⁰Model averaging based on predictive likelihood methods is not limited to Bayesian framework. Kapetanios, Labhard and Price (2006) use predictive likelihood, as opposed to the likelihood of observed data, to construct weights based on information criteria in a frequentist setting.

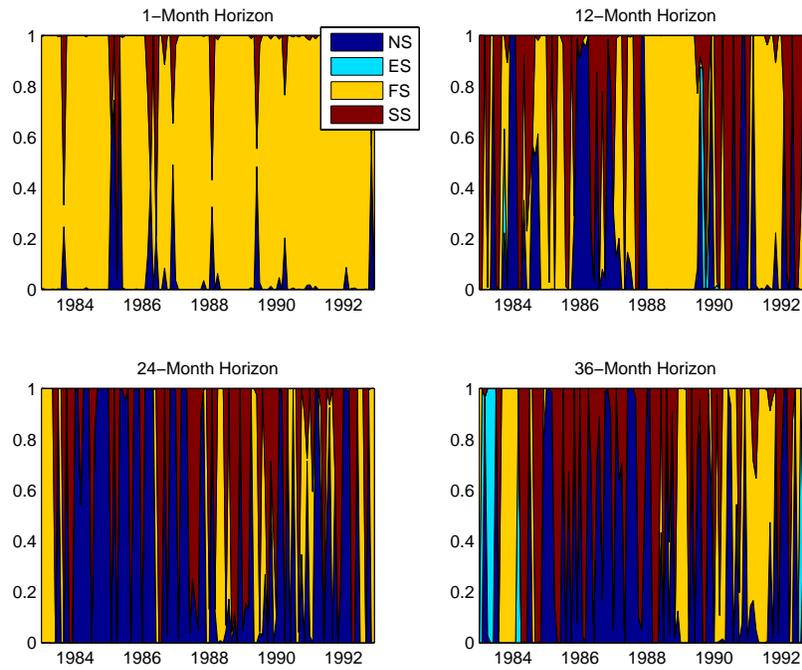


Figure 6: Predictive likelihood weights over the training period of 120 points.

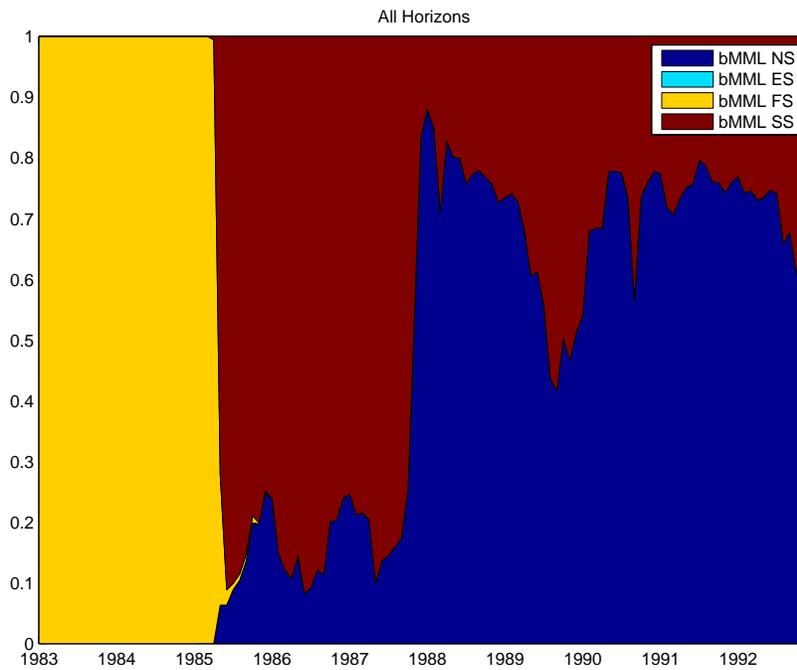


Figure 7: Marginal model likelihood weights over the training period of 120 points.

One possibility for such transformation is to let a be the lower bound of the interval on which our observed log likelihoods lie, order the log likelihoods in ascending order, and specify that $\frac{\log(p(Y_T|\mathcal{M}_i))-a}{\log(p(Y_T|\mathcal{M}_j))-a} = \frac{w_{i,T}}{w_{j,T}}$ for $i = 1, 2, 3$, $j = 2, 3, 4$, with $\sum_{k=1}^4 w_k = 1$. For marginal model likelihoods (alternatively, we could have used logs of predictive likelihoods), the set of weights

$$w_{k,T} = \frac{\log(p(Y_T|\mathcal{M}_k)) - a}{\sum_{j=1}^4 (\log(p(Y_T|\mathcal{M}_j)) - a)} \quad (15)$$

solve the linear system and satisfy the desired properties for weights stated above. Now the only tricky part is to choose a appropriately.¹¹ We take $a = \log(p(Y_T|\mathcal{M}_1)) - s$, where s is the standard deviation of the log marginal model (predictive) likelihoods from their mean.

We find that the weights calculated in such a manner are much more stable, as shown in Figures 8 and 9 for marginal model likelihoods and predictive likelihoods, respectively, for $T \in [120, 120 + M - h]$ and $M = 120$. Note that in Figure 9 the weights are the same for all four forecasting horizons, since log marginal model likelihood weights are independent of the forecasting horizon (the same situation as with marginal model likelihood weights in Figure 7).

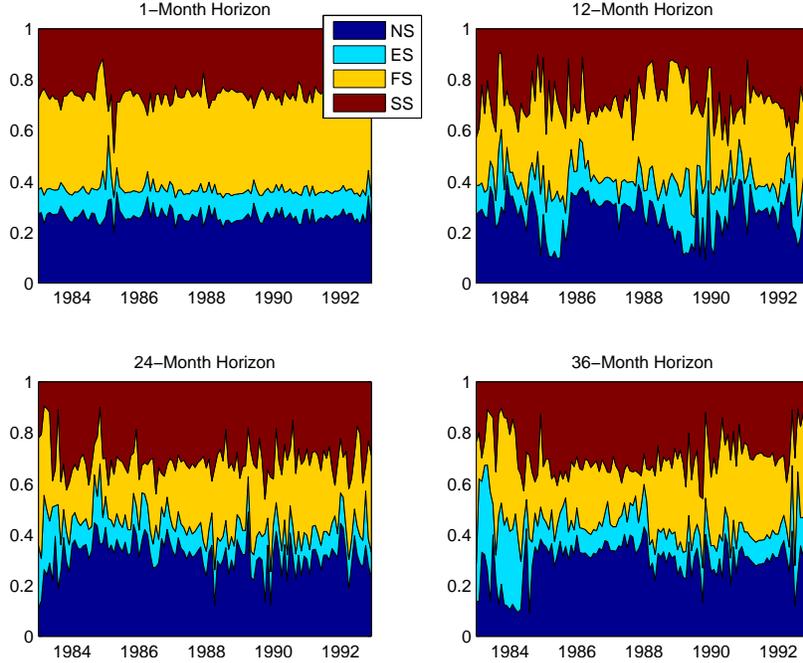


Figure 8: Log predictive likelihood weights over the training period of 120 points.

Finally, we average the weights over the training period. For log marginal model likelihood combination, the weights are

$$w_k^{\mathcal{C}_s} = \frac{\sum_{T=120}^{120+M-h} \left(\frac{\log(p(Y_T|\mathcal{M}_k)) - a}{\sum_{j=1}^4 (\log(p(Y_T|\mathcal{M}_j)) - a)} \right)}{M - h + 1}. \quad (16)$$

¹¹There are many ways to do this. We are not claiming that our suggested method is superior in any way; it is just a way to measure dispersion in the observed data.

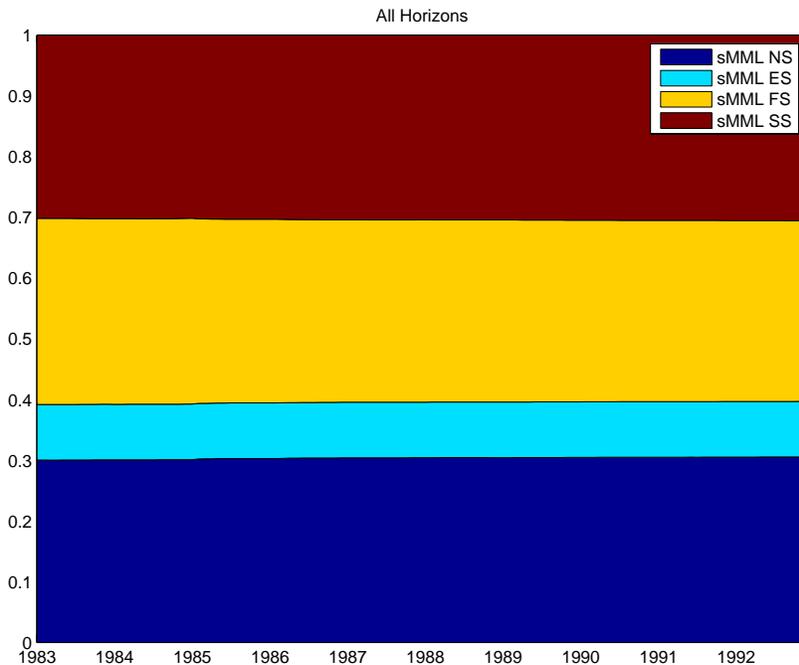


Figure 9: Log marginal model likelihood weights over the training period of 120 points.

For log predictive likelihood combination, we have

$$w_{k,h}^{\mathcal{C}_9} = \frac{\sum_{T=120}^{120+M-h} \left(\frac{\log(p(Y_T|\mathcal{M}_k, \mathcal{F}_{T-h})) - a}{\sum_{j=1}^4 (\log(p(Y_T|\mathcal{M}_j, \mathcal{F}_{T-h})) - a)} \right)}{M - h + 1}. \quad (17)$$

4 Evaluating Model-Combination Schemes

We use two methods to evaluate the performance of the nine previously described model combinations schemes. We call these approaches *dynamic* and *static* model averaging. For both we require the following ingredients: forecasts from individual models to be combined, a subset of the data to train the weights for model combinations, and the remainder of the data to evaluate the out-of-sample forecasts of different model combinations.

We generate individual forecasts for our models $Z_{T+h}^{\mathcal{M}_k}$, $k = 1, \dots, 4$, for $T \in [120, 416 - h]$, as described in Section 2.2, and set these aside. Next we take a subset of these forecasts of length M to evaluate the predictive ability of the models and use this information to obtain the weights for model combinations. In Section 3 we refer to this as training the weights. The last observation used in the training period to evaluate individual forecasts is $120 + M$. Starting at this point $T = 120 + M$, we can combine the models using their respective weights and evaluate the out-of-sample predictive ability of the combinations using the remainder of the

sample. That is, we calculate the forecast error

$$e_{T+h}^{\mathcal{C}_j} = \sqrt{\left(\frac{\left(Z_{T+h} - \hat{Z}_{T+h}^{\mathcal{C}_j} \right)' \left(Z_{T+h} - \hat{Z}_{T+h}^{\mathcal{C}_j} \right)}{N} \right)} \quad (18)$$

for $j = 1, \dots, 9$ model combinations at points $T \in [120 + M, 416 - h]$. Schematics with a graphic description of the dynamic and static forecasting approaches are found in Figures 10 and 11.

The key difference between the two methods for evaluating the combinations is their treatment of the training period. In the dynamic approach, the parameters of the model averaging scheme are updated gradually as we move forward in time. In this way, the most recent information regarding the forecasting performance of the models is incorporated in the model-averaging algorithm. The static approach, however, involves only a single computation of the model-combination parameters. As we move through time, therefore, the parameters are not updated to incorporate the most recent forecasting performance. Such evaluation is not the typical approach used in the forecasting literature, but it nonetheless appropriate for examining the usefulness of a given model-combination scheme for simulation analysis, where one does not have the liberty of updating continuously one's information set. We expect that with a limited training set, the static forecast combinations should underperform their dynamic counterparts.

4.1 Dynamic model averaging

The idea with dynamic model averaging is to use as much recent information as possible to train the weights for model combinations. We consequently update the training period as new information arrives: starting with $M = 120$, we increase the training period until we run out of data (the last value for M is $416 - h$). The steps involved are given in Figure 10.

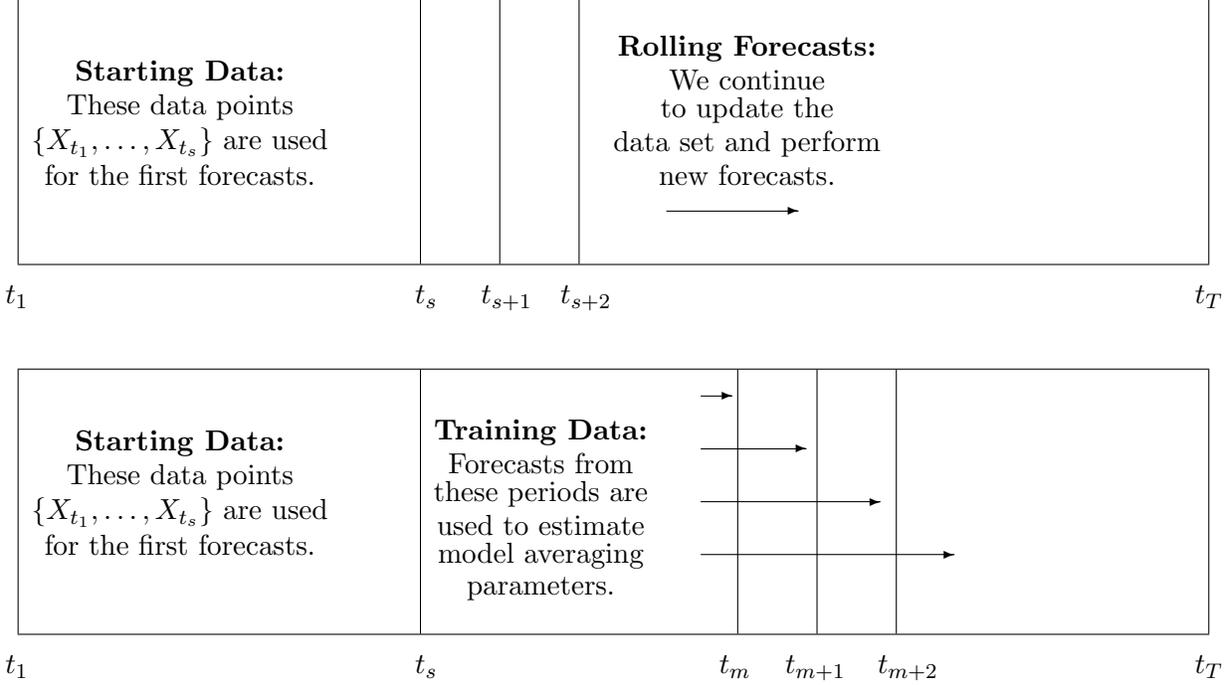
Figure 12 shows the predictive performance of frequentist combinations ($\mathcal{C}_1 - \mathcal{C}_5$) relative to the random walk using a rolling window of 48 observations. With the exception of factor OLS, all combinations beat the random walk on average for one-month horizon. As the horizon increases, the performance of Inverse Error, Equal Weights and especially MARS combinations worsens,¹² while factor scheme OLS improves significantly. Past the one-month horizon, the simple OLS scheme outperforms all other frequentist combinations, approaching the random walk at one- and two-year horizons, and beating the random walk for the entire out-of-sample evaluation period at the three-year horizon. An interesting result is that the predictive performance of Inverse Error and Equal Weights is almost identical in our setting.

Figure 13 shows the performance of the Bayesian model averaging schemes \mathcal{C}_6 and \mathcal{C}_7 relative to the random walk, as well as Equal Weights and simple OLS, for comparison with the frequentist combinations. We see that our Bayesian schemes do not beat the frequentist ones in the dynamic-evaluation approach.

Figure 14 compares Bayesian log combinations \mathcal{C}_8 and \mathcal{C}_9 to the random walk. Equal Weights and simple OLS schemes are also displayed for reference. We observe that using weights based on the logs of marginal model and predictive likelihoods improves the performance of Bayesian schemes significantly: they beat the random walk and the simple OLS scheme at the one-month horizon and get close to the Equal Weights combination at longer horizons.

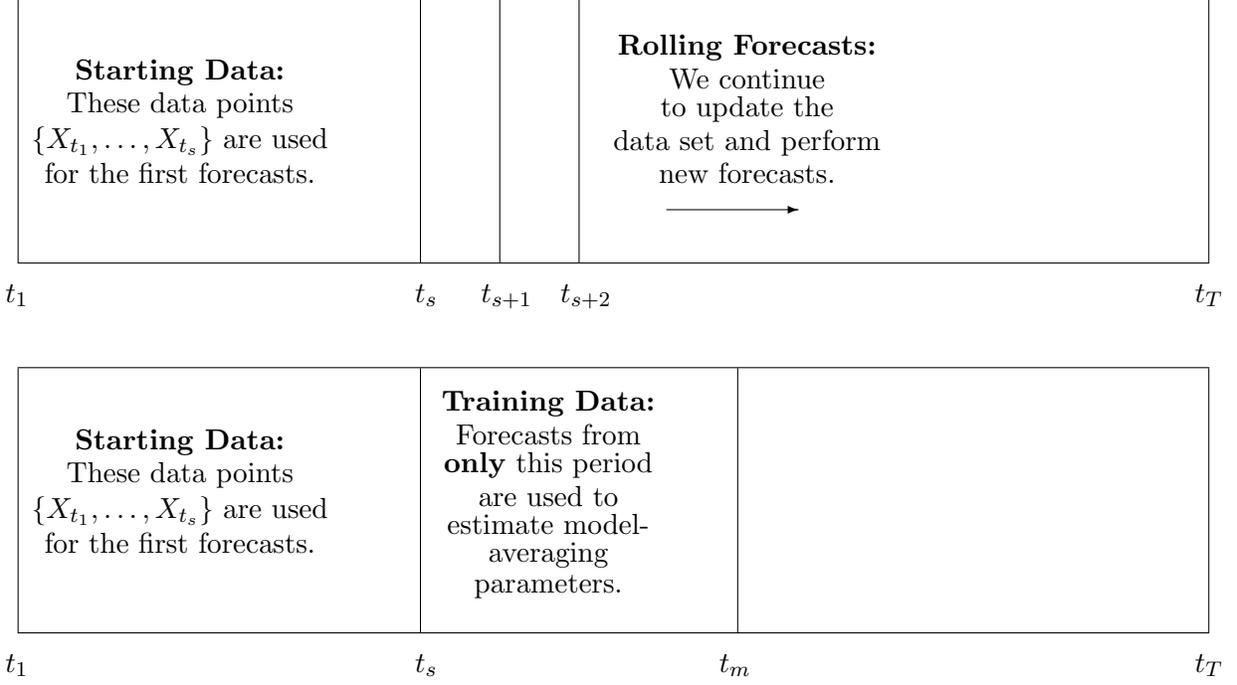
¹²The MARS result is not surprising: as shown in Sephton (2001), MARS scheme is very promising in-sample, but its out-of-sample performance is not entirely accurate.

Figure 10: **Dynamic Model Averaging:** This schematic describes the steps involved in dynamic model averaging whereby the parameters for each model-averaging algorithm are updated as new information becomes available.



0. Set $i = m$, $j = 1$, and $h = 1$;
1. Estimate $\mathbb{P}_{\mathcal{C}_j}(\mathcal{M}_k | \mathcal{F}_{t_i})$ for $k = 1, \dots, n$;
2. Apply weights to $\{\hat{Z}_{t_{i+h}}^{\mathcal{M}_k}, k = 1, \dots, n\}$ to form $\mathbb{E}_{\mathcal{C}_j}(Z_{t_{i+h}} | \mathcal{F}_{t_i})$;
3. Compute $\epsilon_{t_{i+h}}^{\mathcal{C}_j} = Z_{t_{i+h}} - \mathbb{E}_{\mathcal{C}_j}(Z_{t_{i+h}} | \mathcal{F}_{t_i})$;
4. Repeat steps 1-3 for $j = 2, \dots, \kappa$ model-averaging approaches;
5. Repeat steps 1-4 for $i = m + 1, \dots, T - h$.
6. Repeat steps 1-5 for $h = 2, \dots, H$ forecasting horizons.

Figure 11: **Static Model Averaging:** This schematic describes the steps involved in static model averaging whereby the parameters for each model-averaging algorithm are estimated only once with a fixed set of training data and *not* updated as new information becomes available.



0. Estimate **once** $\mathbb{P}_{\mathcal{C}_j}(\mathcal{M}_k | \mathcal{F}_{t_m})$ for $k = 1, \dots, n$. Note: m is fixed;
1. Set $i = m$, $j = 1$, and $h = 1$;
2. Apply **fixed** weights to $\{\hat{Z}_{t_{i+h}}^{\mathcal{M}_k}, k = 1, \dots, n\}$ to form $\mathbb{E}_{\mathcal{C}_j}(Z_{t_{i+h}} | \mathcal{F}_{t_i})$;
3. Compute $\epsilon_{t_{i+h}}^{\mathcal{C}_j} = Z_{t_{i+h}} - \mathbb{E}_{\mathcal{C}_j}(Z_{t_{i+h}} | \mathcal{F}_{t_i})$;
4. Repeat steps 2-3 for $j = 1, \dots, \kappa$ model-averaging approaches;
5. Repeat steps 2-4 for $i = m + 1, \dots, T - h$ observations;
6. Repeat steps 2-5 for $h = 2, \dots, H$ forecasting horizons.

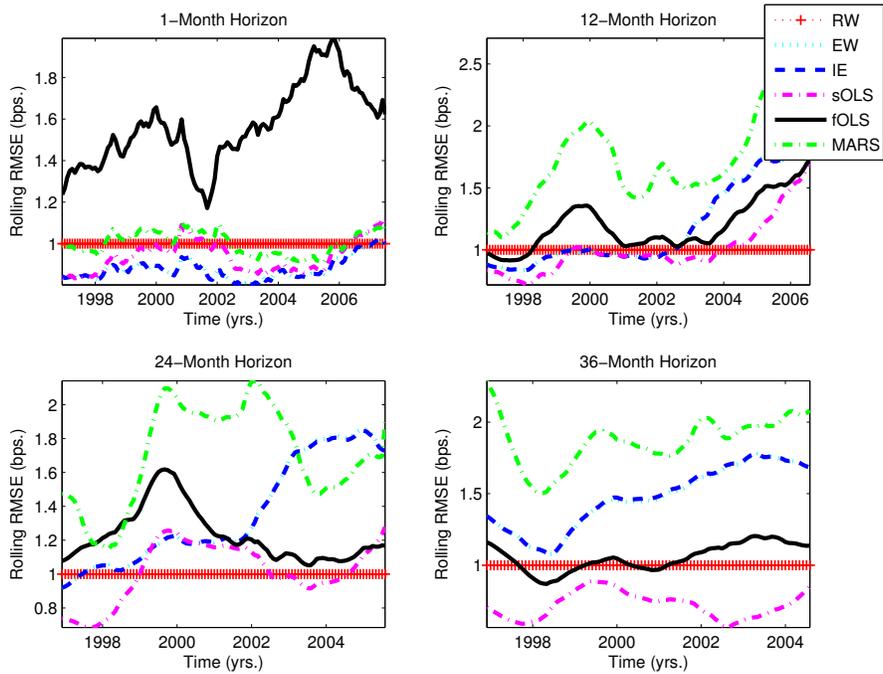


Figure 12: Dynamic predictive performance for frequentist combinations relative to random walk.

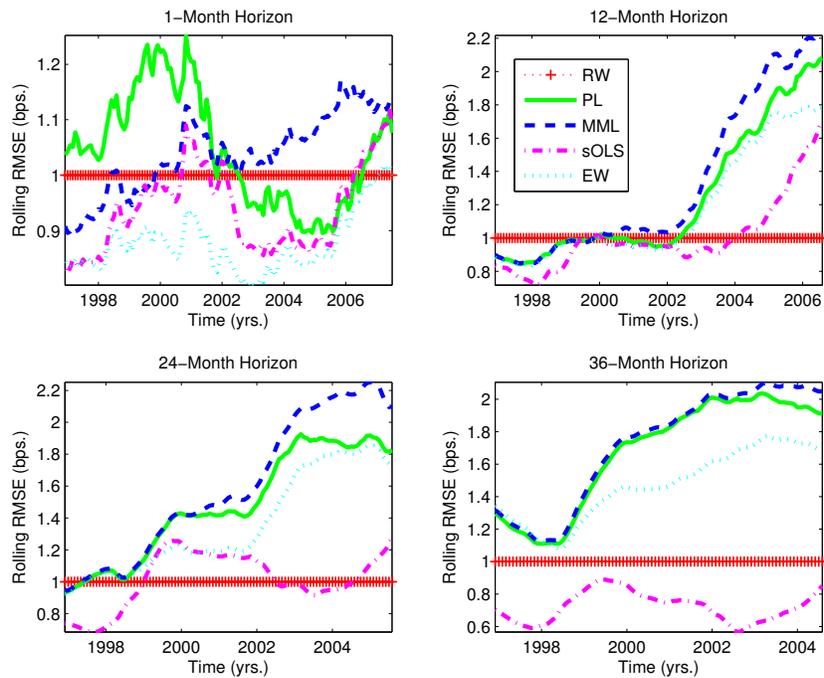


Figure 13: Dynamic predictive performance for Bayesian combinations relative to random walk.

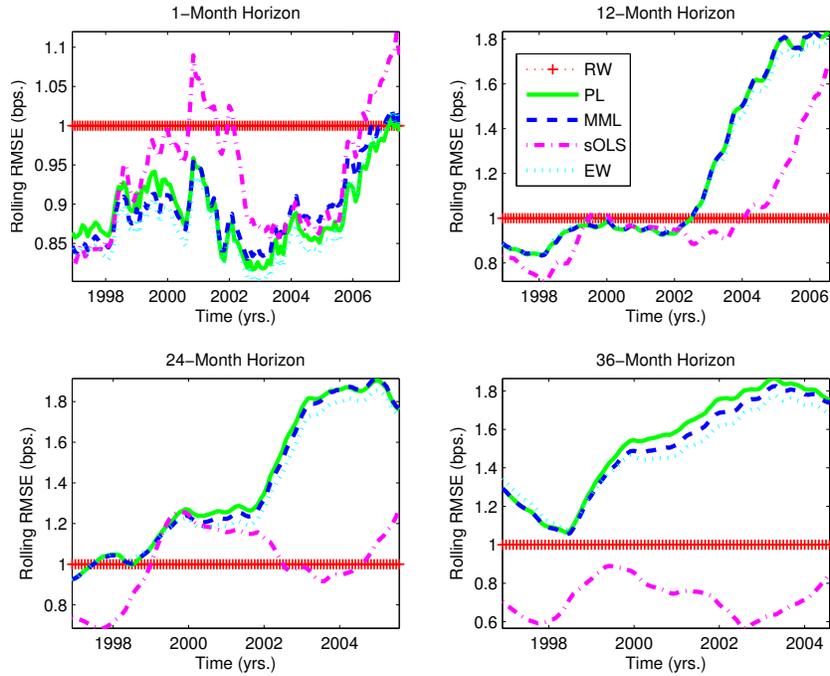


Figure 14: Dynamic predictive performance for Bayesian log combinations relative to random walk.

4.2 Static model averaging

We may not always be in the position where we can increase the training period as is done in the dynamic setting.¹³ So we have to test how well the different combinations perform if we calculate the weights over a fixed training period and apply these weights to all remaining individual forecasts out-of-sample, without updating the training period. The steps for static model averaging are given in Figure 11.

Figures 15-17 show the predictive performance of our nine combinations in the static model averaging setting. Comparing to the same figures from the dynamic setting, we see that Equal Weights, Inverse Error, and Bayesian schemes are more robust to the training period than other combinations—MARS, simple OLS, and factor OLS, in the sense that predictive performance of the former combinations is quite similar in both dynamic and static settings and thus not very sensitive to the estimation period. The performance of the latter schemes (particularly MARS) deteriorates when we estimate the weights over a fixed training period. However, the performance of the combinations relative to each other is the same in both dynamic and static settings: Equal Weights and simple OLS are still the best frequentist schemes, and Bayesian log likelihood schemes are close to the Equal Weights. Finally, for horizons beyond one month, simple OLS combination beats all other schemes and is only slightly worse than the random walk at long horizons.

¹³For instance, as debt managers in a central bank, we may have to use weights calculated over some fixed period to calculate term-structure forecasts for the purposes of managing a foreign reserves portfolio or debt issuance for the next couple of years.

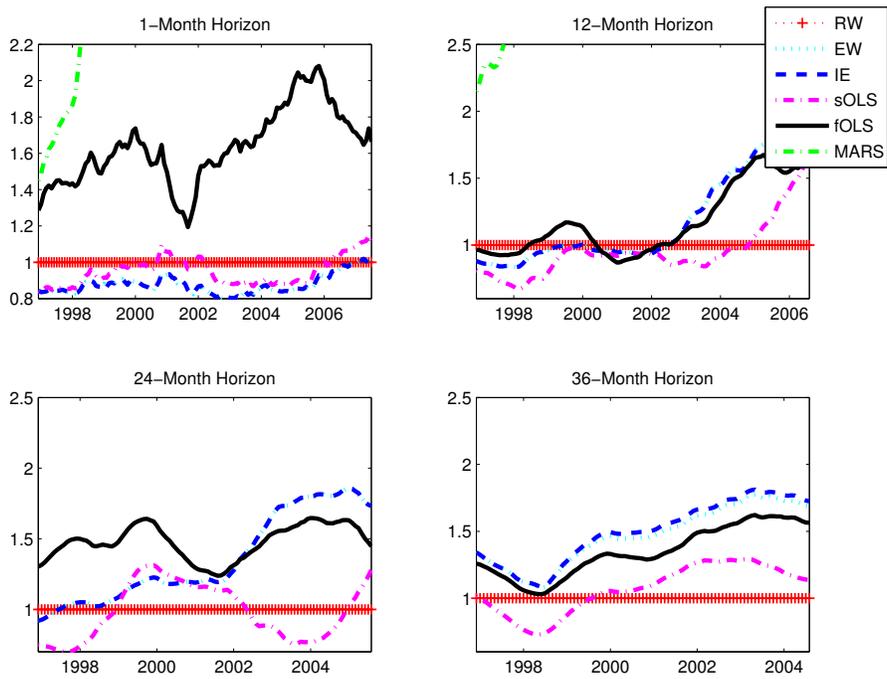


Figure 15: Static predictive performance for frequentist combinations relative to random walk.

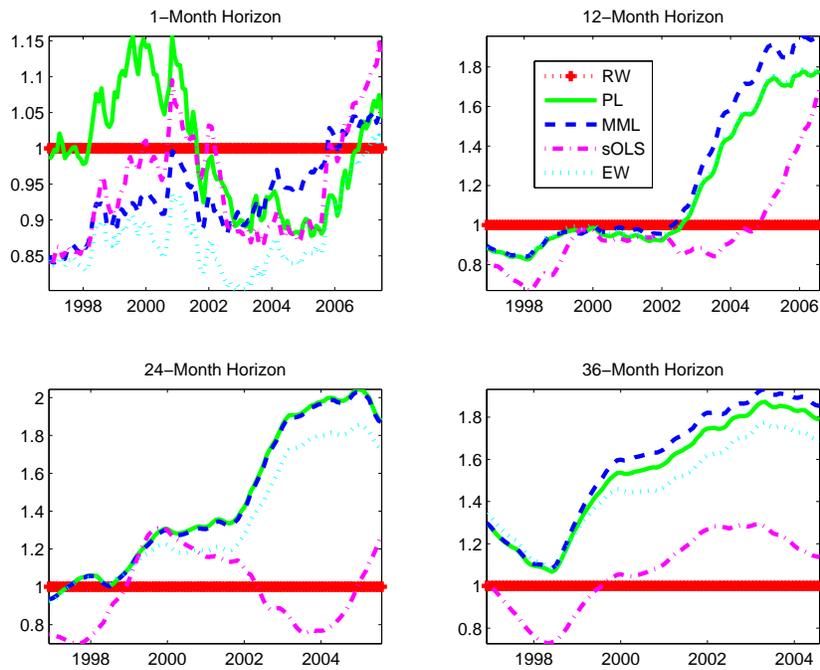


Figure 16: Static predictive performance for Bayesian combinations relative to random walk.

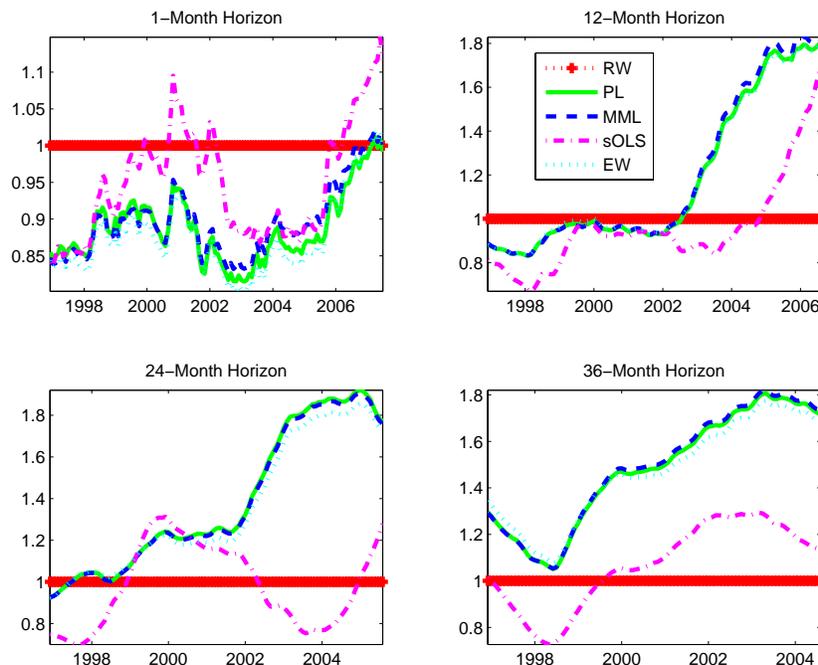


Figure 17: Static predictive performance for Bayesian log combinations relative to random walk.

4.3 Best combinations vs. best individual models

Since the objective of this paper is to answer the question of whether there is benefit from using combinations of models as opposed to a single best-performing model, it makes sense to address this question directly. From Figure 3, we see that the Nelson-Siegel model performs well for short horizons, and the Fourier Series model performs well for longer horizons. Figure 18 compares these two models, and the combination schemes that perform best in the static model averaging setting (Equal Weights, Log Predictive Likelihood, and simple OLS), to the random walk.

We can make the following observations. All of our best combinations beat the best individual models at the 1-month horizon on average. As the length of the horizon increases, Equal Weights and Log Predictive Likelihood schemes outperform the Nelson-Siegel model, but not the Fourier Series model. On average, the simple OLS combination outperforms both individual models at all horizons. While it may be tempting to conclude that the simple OLS combination should be implemented instead of a single model, we are not ready to make this conclusion. First, simple OLS is unconstrained, which means that the weights can be negative and they need not sum to one. The idea of assigning negative weights to particular forecasts may be difficult to accept for policymakers. Consequently, there may be practical obstacles to implementing this combination scheme. Also, forecasts with unconstrained OLS weights and no intercept (as is the case in our situation) may be biased, as pointed out in Diebold and Pauly (1987). Second, some preliminary testing results (not reported here) show that the simple OLS scheme is sensitive to the subset of data used for the training period and to the length of the training period, as can be expected with least squares estimation in a relatively small sample. Further analysis of this particular combination scheme, including hypothesis testing

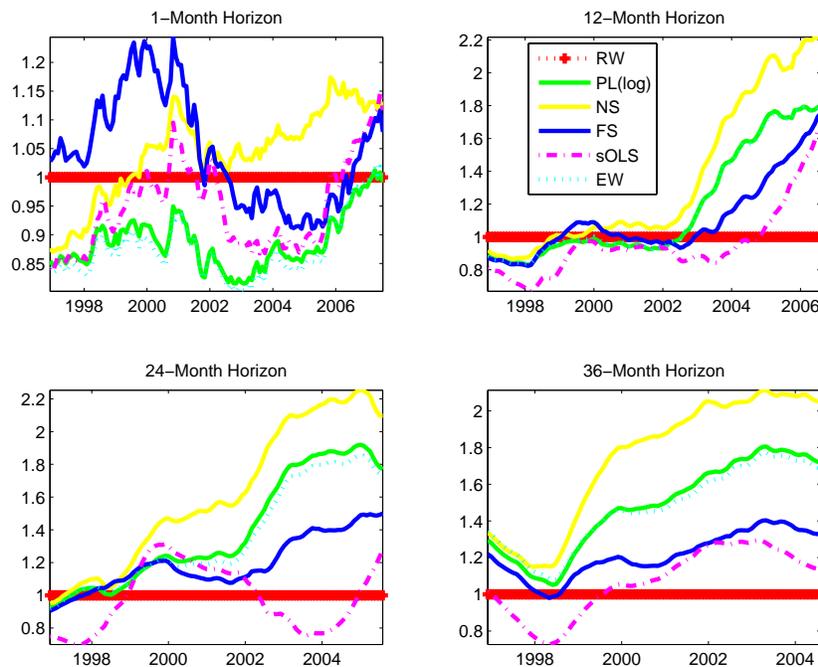


Figure 18: Predictive performance of best individual models and best combinations relative to random walk, static setting.

and forecast error analysis, such as that done in Li and Tkacz (2004), is left for future work.

5 Final Remarks

The main question of this paper is whether or not one can combine multiple interest-rate models to create a single model that outperforms any one individual model. To this end, nine alternative model averaging techniques are considered including choices from the frequentist and Bayesian literature as well as a few new alternatives. These approaches are compared, in the context of both a dynamic and a static forecasting exercise, with more than thirty years of monthly Canadian interest-rate and macroeconomic data. We do not conduct hypothesis tests in this paper, so we do not claim any statistical improvements, but we can still make some observations regarding the predictive performance of the different model combinations.

The principal observation is that we find evidence of model combinations outperforming the best individual forecasts over the evaluation period. The degree of outperformance depends, however, on both the forecasting horizon and the type of model combination. At shorter forecasting horizons, for example, almost all model combinations outperform the best single forecast. As the forecasting horizon increases, however, only the simple OLS averaging scheme consistently outperforms the best single-model forecast. Indeed, the simple OLS approach also outperforms, on a number of occasions, the rather difficult random-walk forecasting benchmark; this is something that none of the individual forecasts achieves on a consistent basis. It is also clear that the simpler model combination approaches tend to outperform their more complex counterparts. Similarly to our results, Ravazzolo, van Dijk and Verbeek (2007) find that uncon-

strained OLS combination scheme (like our simple OLS), and combinations with time-varying weights, outperform more complex schemes. While this is consistent with the evidence in the literature that simpler schemes dominate their more complex counterparts, Stock and Watson (2004) note that it is difficult to explain such findings in the context of combining weights in a stationary environment.

Even though the simple OLS combination scheme generally performs quite well, it does have the disadvantage of demonstrating some instability with respect to the training period selected for the determination of the model-combination parameters. We need to investigate the simple OLS combination further and test its sensitivity to the training period (its length and the time over which the weights are trained). This type of analysis should also be done for other combination schemes, such as Log Predictive Likelihood, that have shown promise in our study. Another interesting direction is to investigate the predictive performance of the combination of the less stable simple OLS and the very stable, and generally well-performing, Equal Weights.

One more possibility for further investigation is to consider combinations that are based on time-varying weights. Ravazzolo, van Dijk and Verbeek (2007) find that time-varying combinations perform well in terms of predictive ability as well as in economic sense, based on the results of an investment exercise. Time-varying weights have the advantage that they may capture structural breaks by assigning varying weights to the combined models at different periods. However, we have to be careful about incorporating time-varying weights in the context of funds management, since we may not be at liberty to update the information set in operational activities.

Appendix

Here we derive the posterior density, the predictive density and the marginal model likelihood for a VAR(L) model. Special thanks to Michiel de Pooter and Francesco Ravazzolo for the results and derivations of posterior and predictive densities for VAR models. For the most part, in this section we follow their notation and derivations.

5.1 The model

A VAR(L) model can be written as

$$Y = X\Pi + \epsilon, \quad \text{vec}(\epsilon) \sim N(0, \Sigma \otimes I_T), \quad (19)$$

where Y is a $T \times N$ matrix of observed data: each row represents an observation of a $1 \times N$ vector \mathbf{y}_t , $t = 1, \dots, T$; X is a $T \times K$ matrix of explanatory variables; Π is a $K \times N$ matrix of parameters; and each of T $1 \times N$ row vectors ϵ_t in the $T \times N$ error matrix ϵ is normally distributed with mean 0 and $N \times N$ variance-covariance matrix Σ . The first column of X is composed of ones, corresponding to the constants in the parameter matrix Π , and the remaining columns are lagged values of Y , so $K = 1 + LN$ for a VAR(L) model.

Note that Y and ϵ are random matrices, and in derivations below we will be using several matrix-variate distributions (their densities and related results are given in Gupta and Nagar (1999) and Poirier (1995)).

5.2 The likelihood

The likelihood function (data-generating process) for the model in (19) is

$$\begin{aligned} p(Y|X, \Pi, \Sigma) &= (2\pi)^{-TN/2} |\Sigma|^{-T/2} |I_T|^{-N/2} \exp\left(-\frac{1}{2} \text{tr}(\Sigma^{-1}(Y - X\Pi)' I_T^{-1} (Y - X\Pi))\right) \\ &= (2\pi)^{-TN/2} |\Sigma|^{-T/2} \exp\left(-\frac{1}{2} \text{tr}(\Sigma^{-1}(Y - X\Pi)' (Y - X\Pi))\right). \end{aligned} \quad (20)$$

This likelihood function is equivalent to the product of T likelihood functions for \mathbf{y}_t (with each \mathbf{y}_t multivariate normal), since we assume that ϵ_t are independent from one period to the next.

5.3 The prior

For the model (19), the conjugate priors for parameters Π and Σ have the following form:

$$\begin{aligned} p(\Pi, \Sigma) &= p(\Pi|\Sigma)p(\Sigma), \\ \text{vec}(\Pi|\Sigma) &\sim N(\text{vec}(P), \Sigma \otimes Q), \\ \Sigma &\sim IW(C, \nu). \end{aligned} \quad (21)$$

IW denotes Inverted Wishart distribution for Σ :

$$p(\Sigma) = c_{IW}^{-1} \cdot |\Sigma|^{-(\nu+N+1)/2} |C|^{\nu/2} \exp\left(-\frac{1}{2}\text{tr}(\Sigma^{-1}C)\right), \quad (22)$$

$$c_{IW} = 2^{\nu N/2} \pi^{N(N-1)/4} \prod_{n=1}^N \Gamma\left(\frac{\nu+1-n}{2}\right).$$

The $N \times N$ symmetric positive definite matrix C is generally referred to as scale matrix and the scalar $\nu \geq N$ as degrees of freedom. The distribution of $\Pi|\Sigma$ is matrix-normal with mean P and symmetric positive definite variance matrices Σ , $N \times N$, and Q , $K \times K$:

$$p(\Pi|\Sigma) = (2\pi)^{-KN/2} |\Sigma|^{-K/2} |Q|^{-N/2} \exp\left(-\frac{1}{2}\text{tr}(\Sigma^{-1}(\Pi - P)'Q^{-1}(\Pi - P))\right). \quad (23)$$

5.4 The posterior

The posterior density of parameter matrices Π and Σ summarizes the information available to us about them from prior belief and observed data. The joint posterior density of Π and Σ is a product of likelihood and prior distribution:

$$\begin{aligned} p(\Pi, \Sigma|Y, X) &\propto p(Y|X, \Pi, \Sigma)p(\Pi, \Sigma) = p(Y|X, \Pi, \Sigma)p(\Pi|\Sigma)p(\Sigma) \\ &\propto \frac{\exp\left(-\frac{1}{2}\text{tr}(\Sigma^{-1}(C + (Y - X\Pi)'(Y - X\Pi) + (\Pi - P)'Q^{-1}(\Pi - P)))\right)}{|\Sigma|^{(T+N+K+\nu+1)/2}}. \end{aligned}$$

For the joint prior distribution $p(\Pi, \Sigma)$, we drop the conditioning on X , assuming that the parameters are independent from the explanatory variables in the matrix X . However, the posterior distribution of the parameters does depend on X . To make draws from the joint posterior, it is convenient to derive the marginal posterior $p(\Sigma|Y, X)$ and the conditional posterior $p(\Pi|\Sigma, Y, X)$, similarly to how we specified the prior distributions.

For derivations that follow, we need the following two results:

Decomposition rule:

$$(Y - X\Pi)'(Y - X\Pi) = (Y - X\hat{B})'(Y - X\hat{B}) + (\Pi - \hat{B})'X'X(\Pi - \hat{B}), \quad (24)$$

where

$$\hat{B} = (X'X)^{-1}X'Y. \quad (25)$$

Inverted Wishart integration step:

$$\int |\Sigma|^{-M/2} \exp\left(-\frac{1}{2}\text{tr}(\Sigma^{-1}A)\right) d\Sigma = k \cdot |A|^{-(M-N-1)/2}, \quad (26)$$

$$k = 2^{N(M-N-1)/2} \pi^{N(N-1)/4} \prod_{n=1}^N \Gamma\left(\frac{M-N-n}{2}\right),$$

for an integer M and $N \times N$ positive definite symmetric matrix Σ .

The first result can be verified by direct multiplication: on the one hand,

$$(Y - X\Pi)'(Y - X\Pi) = Y'Y - Y'X\Pi - \Pi'X'Y + \Pi'X'X\Pi. \quad (27)$$

On the other hand,

$$\begin{aligned} & (Y - X\hat{B})'(Y - X\hat{B}) + (\Pi - \hat{B})'X'X(\Pi - \hat{B}) \\ &= (Y - X(X'X)^{-1}X'Y) + (\Pi - (X'X)^{-1}X'Y)'X'X(\Pi - (X'X)^{-1}X'Y) \\ &= Y'Y - Y'X(X'X)^{-1}X'Y - Y'X(X'X)^{-1}X'Y + Y'X(X'X)^{-1}X'X(X'X)^{-1}X'Y \\ &\quad + \Pi'X'X\Pi - \Pi'X'X(X'X)^{-1}X'Y - Y'X(X'X)^{-1}X'X\Pi + Y'X(X'X)^{-1}X'X(X'X)^{-1}X'Y \\ &= Y'Y + \Pi'X'X\Pi - \Pi'X'Y - Y'X\Pi. \end{aligned} \quad (28)$$

Comparing the last line above to (27), we see that the decomposition rule has been established.

The Inverted Wishart integration step follows from the fact that the integral of the Inverted Wishart density (22) equals 1:

$$\int p(\Sigma)d\Sigma = 1. \quad (29)$$

Taking $M = \nu + N + 1$, dividing inside the integral by the appropriate constant and multiplying 1 by the same constant produces the desired result in (26).

Our goal is to write the joint posterior density in the form

$$p(\Pi, \Sigma|Y, X) = p(\Pi|\Sigma, Y, X)p(\Sigma|Y, X). \quad (30)$$

For this, we rewrite (24) as follows:

$$\begin{aligned} p(\Pi, \Sigma|Y, X) &\propto \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(C + (Y - X\Pi)'(Y - X\Pi) + (\Pi - P)'Q^{-1}(\Pi - P))\right)\right)}{|\Sigma|^{(T+N+K+\nu+1)/2}} \\ &= \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(C + (W - V\Pi)'(W - V\Pi))\right)\right)}{|\Sigma|^{(T+N+K+\nu+1)/2}} \\ &= \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(C + (W - V\hat{\Pi})'(W - V\hat{\Pi}) + (\Pi - \hat{\Pi})'V'V(\Pi - \hat{\Pi}))\right)\right)}{|\Sigma|^{(T+N+K+\nu+1)/2}} \end{aligned} \quad (31)$$

where

$$W = \begin{bmatrix} Y \\ Q^{-1/2}P \end{bmatrix}, \quad V = \begin{bmatrix} X \\ Q^{-1/2} \end{bmatrix}, \quad \hat{\Pi} = (V'V)^{-1}V'W, \quad (32)$$

and $Q^{-1/2}$ is the upper triangular matrix from the Choleski decomposition of the positive definite symmetric matrix $Q = Q^{-1/2'}Q^{-1/2}$.

To verify the first equality in (31), we simplify the products in the exponent in the first and

second line as follows:

$$\begin{aligned}
(W - V\Pi)'(W - V\Pi) &= \left(\begin{bmatrix} Y' & P'Q^{-1/2'} \end{bmatrix} - \Pi' \begin{bmatrix} X' & Q^{-1/2'} \end{bmatrix} \right) \cdot \left(\begin{bmatrix} Y \\ Q^{-1/2}P \end{bmatrix} - \begin{bmatrix} X \\ Q^{-1/2} \end{bmatrix} \Pi \right) \\
&= Y'Y + P'Q^{-1}P - Y'X\Pi - P'Q^{-1}\Pi \\
&\quad - \Pi'X'Y - \Pi'Q^{-1}P + \Pi'X'X\Pi + \Pi'Q^{-1}\Pi,
\end{aligned} \tag{33}$$

and

$$\begin{aligned}
(Y - X\Pi)'(Y - X\Pi) + (\Pi - P)'Q^{-1}(\Pi - P) &= Y'Y - Y'X\Pi - \Pi'X'Y + \Pi'X'X\Pi \\
&\quad + \Pi'Q^{-1}\Pi - \Pi'Q^{-1}P - P'Q^{-1}\Pi + P'Q^{-1}P.
\end{aligned} \tag{34}$$

Comparing (33) and (34), we see that they are the same. To establish the second equality in (31), we use the decomposition rule (24) to verify that

$$(W - V\Pi)'(W - V\Pi) = (W - V\hat{\Pi})'(W - V\hat{\Pi}) + (\Pi - \hat{\Pi})'V'V(\Pi - \hat{\Pi}), \tag{35}$$

with $\hat{\Pi}$ given in (32).

Now we can separate parts of the last equality in (31) to write the joint posterior density in the form

$$\begin{aligned}
p(\Pi, \Sigma | Y, X) &= \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(C + (W - V\hat{\Pi})'(W - V\hat{\Pi}))\right)\right)}{|\Sigma|^{(N+(T+\nu)+1)/2}} \\
&\quad \cdot \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\Pi - \hat{\Pi})'V'V(\Pi - \hat{\Pi})\right)\right)}{|\Sigma|^{K/2}}.
\end{aligned} \tag{36}$$

Comparing this last line with equation (67) and the formulas for the Inverted Wishart and the matric-normal densities ((22) and (23), respectively), we see that the marginal posterior distribution of Σ is Inverted Wishart, and the conditional posterior distribution of Π is matric-normal. That is,

$$\begin{aligned}
(\Sigma | Y, X) &\sim IW(\hat{C}, \hat{\nu}), \\
\hat{C} &= C + (W - V\hat{\Pi})'(W - V\hat{\Pi}), \\
\hat{\nu} &= T + \nu,
\end{aligned} \tag{37}$$

and

$$\begin{aligned}
\text{vec}(\Pi | \Sigma, Y, X) &\sim N(\text{vec}(\hat{\Pi}), \Sigma \otimes \hat{Q}), \\
\hat{Q} &= (V'V)^{-1},
\end{aligned} \tag{38}$$

and W , V and $\hat{\Pi}$ are given in (32).

Note that we could also derive the marginal posterior distribution for Π , $p(\Pi | Y, X)$ using the Inverted Wishart integration step (26), but for our purposes, we generate draws from the

joint posterior $(\Pi, \Sigma|Y, X)$ by drawing from $p(\Sigma|Y, X)$ and then from $p(\Pi|\Sigma, Y, X)$, following (67). Derivations for $p(\Pi|Y, X)$ are given in de Pooter, Ravazzolo and van Dijk (2007).

At this point, we have enough information to describe our model completely: instead of the frequentist point estimates of parameters, we have posterior parameter distributions. The posterior distributions combine information from our prior opinion about parameters and the information contained about them in the observed data. For example, the posterior scale matrix \hat{C} (37) is a function of the prior scale matrix C and data, Y and X (via V and W). The posterior distributions produce explicitly the expected value of the random parameters as well as their variability.

5.5 Predictive density

We are interested in deriving the predictive density for an $h \times N$ matrix \tilde{Y} of h future values of $1 \times N$ vector Y . We assume that the same model that generates the observed data (19) also generates \tilde{Y} :

$$\tilde{Y} = \tilde{X}\Pi + \epsilon, \quad \text{vec}(\epsilon) \sim N(0, \Sigma \otimes I_h), \quad (39)$$

with \tilde{X} an $h \times K$ matrix of explanatory variables, Π the $K \times N$ matrix of parameters, and ϵ an $h \times N$ matrix of errors.

Conditional on \tilde{X} , Π , Σ , as well as Y and X , \tilde{Y} is matrix-normally distributed:

$$\begin{aligned} \text{vec}(\tilde{Y}|\Sigma, \tilde{X}, Y, X) &\sim N(\text{vec}(\tilde{X}\Pi), \Sigma \otimes I_h), \\ p(\tilde{Y}|\Sigma, \tilde{X}, Y, X) &= (2\pi)^{-hN/2} |\Sigma|^{-h/2} |I_T|^{-N/2} \exp\left(-\frac{1}{2} \text{tr}\left(\Sigma^{-1}(\tilde{Y} - \tilde{X}\Pi)' I_T^{-1} (\tilde{Y} - \tilde{X}\Pi)\right)\right) \\ &= (2\pi)^{-hN/2} |\Sigma|^{-h/2} \exp\left(-\frac{1}{2} \text{tr}\left(\Sigma^{-1}(\tilde{Y} - \tilde{X}\Pi)' (\tilde{Y} - \tilde{X}\Pi)\right)\right). \end{aligned} \quad (40)$$

The marginal predictive density is obtained by integrating out the dependence on Π and Σ :

$$\begin{aligned} p(\tilde{Y}|\tilde{X}, Y, X) &= \int \int p(\tilde{Y}, \Pi, \Sigma|\tilde{X}, Y, X) d\Pi d\Sigma \\ &= \int \int p(\tilde{Y}|\Pi, \Sigma, \tilde{X}) p(\Pi, \Sigma|Y, X) d\Pi d\Sigma \\ &= \int \int p(\tilde{Y}|\Pi, \Sigma, \tilde{X}) p(\Pi|\Sigma, Y, X) p(\Sigma|Y, X) d\Pi d\Sigma \\ &\propto \int \int \frac{\exp\left(-\frac{1}{2} \text{tr}\left(\Sigma^{-1}(\hat{C} + (\Pi - \hat{\Pi})' \hat{Q}^{-1} (\Pi - \hat{\Pi}) + (\tilde{Y} - \tilde{X}\Pi)' (\tilde{Y} - \tilde{X}\Pi))\right)\right)}{|\Sigma|^{(K+N+\nu+1+h)/2}} d\Pi d\Sigma. \end{aligned} \quad (41)$$

To perform the integration required in the above formula, we follow the steps similar to those used to simplify the expression for the joint posterior (31). First, note that

$$(\Pi - \hat{\Pi})' \hat{Q}^{-1} (\Pi - \hat{\Pi}) + (\tilde{Y} - \tilde{X}\Pi)' (\tilde{Y} - \tilde{X}\Pi) = (\tilde{W} - \tilde{V}\Pi)' (\tilde{W} - \tilde{V}\Pi), \quad (42)$$

where

$$\tilde{W} = \begin{bmatrix} \tilde{Y} \\ \hat{Q}^{-1/2} \hat{P}i \end{bmatrix}, \quad \tilde{V} = \begin{bmatrix} \tilde{X} \\ \hat{Q}^{-1/2} \end{bmatrix}. \quad (43)$$

This can be verified by direct multiplication, similarly to the derivations in (33) and (34). Here $\hat{Q}^{-1/2}$ is the upper triangular matrix from the Choleski decomposition of the positive definite symmetric matrix $\hat{Q} = \hat{Q}^{-1/2'} \hat{Q}^{-1/2}$.

Second, using the decomposition rule (24), we express $(\tilde{W} - \tilde{V}\Pi)'(\tilde{W} - \tilde{V}\Pi)$ as follows:

$$(\tilde{W} - \tilde{V}\Pi)'(\tilde{W} - \tilde{V}\Pi) = (\tilde{W} - \tilde{V}\bar{\Pi})'(\tilde{W} - \tilde{V}\bar{\Pi}) + (\Pi - \bar{\Pi})'\tilde{V}'\tilde{V}(\Pi - \bar{\Pi}), \quad (44)$$

where

$$\bar{\Pi} = (\tilde{V}'\tilde{V})^{-1}\tilde{V}'\tilde{W}. \quad (45)$$

The two simplification steps above allow us to write the marginal predictive density as

$$\begin{aligned} p(\tilde{Y}|\tilde{X}, Y, X) &\propto \iint \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\hat{C} + (\Pi - \hat{\Pi})'\hat{Q}^{-1}(\Pi - \hat{\Pi}) + (\tilde{Y} - \tilde{X}\Pi)'(\tilde{Y} - \tilde{X}\Pi))\right)\right)}{|\Sigma|^{(K+N+\hat{\nu}+1+h)/2}} d\Pi d\Sigma \\ &= \iint \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\hat{C} + (\tilde{W} - \tilde{V}\Pi)'(\tilde{W} - \tilde{V}\Pi))\right)\right)}{|\Sigma|^{(K+N+\hat{\nu}+1+h)/2}} d\Pi d\Sigma \\ &= \int \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\hat{C} + (\tilde{W} - \tilde{V}\bar{\Pi})'(\tilde{W} - \tilde{V}\bar{\Pi}))\right)\right)}{|\Sigma|^{(N+\hat{\nu}+1+h)/2} |\tilde{V}'\tilde{V}|^{-1} |N/2(2\pi)^{-KN/2}} \\ &\quad \cdot \left(\int \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\Pi - \bar{\Pi})'\tilde{V}'\tilde{V}(\Pi - \bar{\Pi})\right)\right)}{|\Sigma|^{K/2} |\tilde{V}'\tilde{V}|^{-1} |N/2(2\pi)^{KN/2}} d\Pi \right) d\Sigma. \end{aligned} \quad (46)$$

The integral with respect to Π equals 1, since it is the integral of matrix-normal density. Using this fact, and the Inverted Wishart integration step (26) with $M = N + \hat{\nu} + 1 + h$ and $A = \hat{C} + (\tilde{W} - \tilde{V}\bar{\Pi})'(\tilde{W} - \tilde{V}\bar{\Pi})$, the formula for the marginal predictive density simplifies further to

$$\begin{aligned} p(\tilde{Y}|\tilde{X}, Y, X) &\propto \int \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\hat{C} + (\tilde{W} - \tilde{V}\bar{\Pi})'(\tilde{W} - \tilde{V}\bar{\Pi}))\right)\right)}{|\Sigma|^{(N+\hat{\nu}+1+h)/2}} d\Sigma \\ &\propto |\hat{C} + (\tilde{W} - \tilde{V}\bar{\Pi})'(\tilde{W} - \tilde{V}\bar{\Pi})|^{-(\hat{\nu}+h)/2}. \end{aligned} \quad (47)$$

Since \tilde{W} , \tilde{V} and $\bar{\Pi}$ are all functions of \tilde{X} and \tilde{Y} , we need to disentangle this last expression to obtain the formula for the marginal predictive density as a function of \tilde{Y} , \tilde{X} and some constants. We will show that

$$\hat{C} + (\tilde{W} - \tilde{V}\bar{\Pi})'(\tilde{W} - \tilde{V}\bar{\Pi}) = \hat{C} + (\tilde{Y} - \tilde{X}\hat{\Pi})'(I_h - \tilde{X}M^{-1}\tilde{X}')(\tilde{Y} - \tilde{X}\hat{\Pi}), \quad (48)$$

where

$$\tilde{M} = (\tilde{X}'\tilde{X} + X'X + Q^{-1}). \quad (49)$$

This would imply that the marginal predictive density has the form of the matrix- t distribution:

$$\begin{aligned} p(\tilde{Y}|\tilde{X}, Y, X) &= \frac{c_{mt}^{-1} \cdot |I_h - \tilde{X}\tilde{M}^{-1}\tilde{X}'|^{N/2} |\hat{C}|^{\hat{\nu}/2}}{|\hat{C} + (\tilde{Y} - \tilde{X}\hat{\Pi})'(I_h - \tilde{X}\tilde{M}^{-1}\tilde{X}')(\tilde{Y} - \tilde{X}\hat{\Pi})|^{(\hat{\nu}+h)/2}}, \\ c_{mt} &= \frac{\pi^{hN/2} \prod_{i=1}^N \Gamma\left(\frac{\hat{\nu}+1-i}{2}\right)}{\prod_{j=1}^N \Gamma\left(\frac{\hat{\nu}+h+1-j}{2}\right)}. \end{aligned} \quad (50)$$

The symmetric positive definite matrices \hat{C} , $N \times N$, and $I_h - \tilde{X}\tilde{M}^{-1}\tilde{X}'$, $h \times h$, are called scale matrices.

Establishing the relation in (48) is not difficult, only time-consuming. The equality is verified by plugging in all the variables we substituted for convenience (for example, $\bar{\Pi}$), direct multiplication and simplification. We will not show all the details here, just the main steps.

First, plugging in the expressions for \tilde{W} and \tilde{V} (43), we show that

$$(\tilde{W} - \tilde{V}\bar{\Pi})'(\tilde{W} - \tilde{V}\bar{\Pi}) = (\tilde{Y} - \tilde{X}\bar{\Pi})'(\tilde{Y} - \tilde{X}\bar{\Pi}) + (\hat{\Pi} - \bar{\Pi})'Q^{-1}(\hat{\Pi} - \bar{\Pi}). \quad (51)$$

Second, using (49) and (45), we verify that

$$\bar{\Pi} = \tilde{M}^{-1}(\tilde{X}'\tilde{Y} + X'X\hat{\Pi} + Q^{-1}\hat{\Pi}). \quad (52)$$

Third, we plug this expression for $\bar{\Pi}$ into the left-hand side of equation (48) and multiply and simplify the two quadratic forms. The resulting rather lengthy expression involves many terms, which we combine and simplify further. Finally, we multiply and simplify the right-hand side of equation (48) and compare similar terms in the two expressions. We will see that they are the same, establishing the required equality and the fact that the marginal predictive density is matrix- t (as given in (50)).

Similarly to the situation with the posterior distribution of parameters, now we have described fully the distribution of future outcomes \tilde{Y} . To generate these future outcomes, we can draw from the normal conditional predictive density (40), having previously generated posterior draws of (Π, Σ) . Alternatively, we can draw directly from the matrix- t marginal predictive density (50); this way, when running programs, we do not have to use memory to store posterior draws of (Π, Σ) .

5.5.1 Weights based on predictive likelihood

We used two Bayesian model averaging approaches: the first based on the marginal model likelihood, and the second based on the predictive likelihood. We obtain the predictive likelihood as follows. First we generate individual Bayesian forecast distributions for Y for each model, using T data points to estimate posterior parameter distributions. Then we split the data into two parts of length $T - h$ (training sample) and h (evaluation sample), where h is

the forecast horizon. The first sample is used to estimate the posterior distribution and the predictive density, and the second is used to evaluate the predictive likelihood.

More specifically, for each model, we estimate the posterior density using $T - h$ data points. Then, using this posterior distribution, we generate a distribution of forecasted values for time $(T - h) + h = T$. At this point we have D draws of forecasted values Y_T^d , $d = 1, \dots, D$.¹⁴ Next, for each draw, we evaluate the predictive density by plugging the realized value Y_T^o into either the conditional predictive density (40) or the marginal predictive density (50), and take the average to get an estimate of the predictive likelihood. Then we use this value to calculate the model's weight by substituting the predictive likelihood instead of the marginal model likelihood into (12).

We use a single realized value ($1 \times N$ vector Y_T) to evaluate the predictive density, following in Ravazzolo, van Dijk and Verbeek (2007) and de Pooter, Ravazzolo and van Dijk (2007), for the reasons given in these papers: to test the predictive ability of the model – the probability of out-of-sample realized values. Andersson and Karlsson (2007) use more than one observed value of Y from the evaluation sample to obtain predictive likelihood, which results in weights that rely more on model fit.

5.6 Marginal model likelihood

The marginal model likelihood for model \mathcal{M} with parameters Π and Σ is

$$\begin{aligned} p(Y|X, \mathcal{M}) &= \int \int p(Y, \Pi_i, \Sigma_i | X, \mathcal{M}) d\Pi_i d\Sigma_i \\ &= \int \int p(Y | \Pi_i, \Sigma_i, X, \mathcal{M}) p(\Pi_i, \Sigma_i) d\Pi_i d\Sigma_i \\ &= \int \int p(Y | \Pi_i, \Sigma_i, X, \mathcal{M}) p(\Pi_i | \Sigma_i) p(\Sigma_i) d\Pi_i d\Sigma_i. \end{aligned} \quad (53)$$

To derive the formula for the marginal model likelihood, we have to integrate the product of the likelihood (20) and the prior (21) with respect to the model parameters:

$$\begin{aligned} p(Y|X, \mathcal{M}) &= c \cdot \int \int \frac{\exp\left(-\frac{1}{2} \text{tr}\left(\Sigma^{-1}(C + (Y - X\Pi)'(Y - X\Pi) + (\Pi - P)'Q^{-1}(\Pi - P))\right)\right)}{|\Sigma|^{(T+N+K+\nu+1)/2}} d\Pi d\Sigma \\ c &= \frac{(2\pi)^{-TN/2} (2\pi)^{-KN/2} |Q|^{-N/2} |C|^{\nu/2}}{2^{\nu N/2} \pi^{N(N-1)/4} \prod_{n=1}^N \Gamma\left(\frac{\nu+1-n}{2}\right)}. \end{aligned} \quad (54)$$

¹⁴We use $D = 1000$ for the results in this paper. More draws could be used, but we are interested only in the means of forecast distributions, which do not change enough to warrant the longer computational time.

Following the derivations in (31), we have:

$$\begin{aligned}
p(Y|X, \mathcal{M}) &= \frac{c}{(2\pi)^{-KN/2}|(V'V)^{-1}|^{-N/2}} \cdot \int \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(C + (W - V\hat{\Pi})'(W - V\hat{\Pi}))\right)\right) |\Sigma|^{K/2}}{|\Sigma|^{(T+N+K+\nu+1)/2}} \\
&\quad \cdot \left(\int \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\Pi - \hat{\Pi})'V'V(\Pi - \hat{\Pi})\right)\right)}{(2\pi)^{KN/2}|(V'V)^{-1}|^{N/2}} d\Pi \right) d\Sigma \\
&= c \cdot (2\pi)^{KN/2} |\hat{Q}|^{N/2} \int \exp\left(-\frac{1}{2}\text{tr}\left(\Sigma^{-1}\hat{C}\right)\right) |\Sigma|^{-(N+\hat{\nu}+1)/2} d\Sigma. \tag{55}
\end{aligned}$$

Above, we used the decomposition rule (24) and the fact that the integral of the matrix-normal density equals 1. The constants W , V , $\hat{\Pi}$ are given in (32), and \hat{C} , $\hat{\nu}$, and \hat{Q} in (37) and (38).

Now we use the Inverted Wishart integration step (26) with $M = N + \hat{\nu} + 1$ and $A = \hat{C}$ to simplify the expression for $p(Y|X, \mathcal{M})$ further:

$$p(Y|X, \mathcal{M}) = c \cdot (2\pi)^{KN/2} |\hat{Q}|^{N/2} |\hat{C}|^{-\hat{\nu}/2} 2^{\hat{\nu}N/2} \pi^{N(N-1)/4} \prod_{n=1}^N \Gamma\left(\frac{\hat{\nu} + 1 - n}{2}\right). \tag{56}$$

Finally, substituting the constant (from (54)) and simplifying, we obtain the formula for the marginal model likelihood:

$$p(Y|X, \mathcal{M}) = \frac{\prod_{i=1}^N \Gamma\left(\frac{\hat{\nu}+1-i}{2}\right)}{\pi^{TN/2} \prod_{j=1}^N \Gamma\left(\frac{\nu+1-i}{2}\right)} \cdot \frac{|C|^{\nu/2} |\hat{Q}|^{N/2}}{|\hat{C}|^{\hat{\nu}/2} |Q|^{N/2}}. \tag{57}$$

5.6.1 Weights based on marginal model likelihood

To generate a combined forecast at time T for horizon h , we calculate the marginal model likelihood for each model (for its transition VAR equation) using T observations in (57). Then we plug this value into the formula (12) to calculate that model's posterior probability. The resulting number is the model's weight, applied to the mean of the model's forecast distribution to obtain the combination forecast for time $T + h$. The weights based on marginal model likelihood do not depend on the forecasting horizon h .

5.7 Derivation of g-prior for the VAR(L) model

Let us derive the g-prior for our VAR(L) model (19), adapting the results of Zellner (1986) to our multidimensional case: we will obtain explicit formulas for parameters P , Q , C and ν .

Before observing the data Y , suppose that we have a hypothetical (imaginary) sample Y_0 , generated by

$$Y_0 = X_0\Pi + \epsilon_0, \quad \text{vec}(\epsilon_0) \sim N(0, \Sigma_0 \otimes I_\tau). \tag{58}$$

Here Y_0 is $\tau \times N$, X_0 is a $\tau \times K$ matrix of explanatory variables, and Π is a $K \times N$ matrix of parameters.

In this conceptual sample Y_0 , we allow the variance-covariance matrix to differ from that of our observed data Y :

$$\Sigma = g\Sigma_0, \quad 0 < g < \infty. \quad (59)$$

We will need to specify the scaling constant g ; Koop and Potter (2003) recommend setting

$$g = \frac{1}{\tau} \quad \text{or} \quad g = \frac{1}{\ln(\tau)^3}, \quad (60)$$

so for calculations in this paper, we take $g = \frac{1}{\tau}$.

We proceed by deriving the posterior density $p(\Pi, \Sigma_0 | Y_0, X_0)$ for our hypothetical sample. For this, we need to specify the prior $p(\Pi, \Sigma_0)$ and the likelihood $p(Y_0 | \Pi, \Sigma_0)$. We suppose that the prior distribution is diffuse (see Zellner (1971), for instance):

$$p(\Pi, \Sigma_0) \propto |\Sigma_0|^{-(N+1)/2}. \quad (61)$$

The likelihood is

$$p(Y_0 | \Pi, \Sigma_0) \propto |\Sigma_0|^{-\tau/2} \exp\left(\frac{1}{2} \text{tr}(\Sigma_0^{-1}(Y_0 - X_0\Pi)'(Y_0 - X_0\Pi))\right). \quad (62)$$

Using (59), we rewrite the prior and the likelihood in terms of Σ :

$$p(\Pi, \Sigma) \propto |\Sigma|^{-(N+1)/2}, \quad (63)$$

and

$$p(Y_0 | \Pi, \Sigma) \propto |\Sigma|^{-\tau/2} \exp\left(\frac{1}{2} \text{tr}(g\Sigma^{-1}(Y_0 - X_0\Pi)'(Y_0 - X_0\Pi))\right). \quad (64)$$

Using the decomposition rule (24) and the fact that the posterior is proportional to prior times likelihood, we have

$$\begin{aligned} p(\Pi, \Sigma | Y_0, X_0) &\propto |\Sigma|^{-(\tau+N+1)/2} \exp\left(\frac{1}{2} \text{tr}(g\Sigma^{-1}(Y_0 - X_0\Pi)'(Y_0 - X_0\Pi))\right) \\ &= \frac{\exp\left(\frac{1}{2} \text{tr}(g\Sigma^{-1}(Y_0 - X_0\hat{B}_0)'(Y_0 - X_0\hat{B}_0) + g\Sigma^{-1}(\Pi - \hat{B}_0)'X_0'X_0(\Pi - \hat{B}_0))\right)}{|\Sigma|^{(\tau+N+1)/2}}, \end{aligned} \quad (65)$$

where

$$\hat{B}_0 = (X_0'X_0)^{-1}X_0'Y_0. \quad (66)$$

We want to express equation (65) as a product of the conditional posterior density $p(\Pi | \Sigma, Y_0, X_0)$

and the marginal posterior density $p(\Sigma|Y_0, X_0)$, so we group terms as follows:

$$\begin{aligned}
p(\Pi, \Sigma|Y_0, X_0) &= p(\Pi|\Sigma, Y_0, X_0)p(\Sigma|Y_0, X_0) \\
&\propto |\Sigma|^{-K/2} \exp\left(\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\Pi - \hat{B}_0)'(gX_0'X_0)(\Pi - \hat{B}_0)\right)\right) \\
&\quad \cdot |\Sigma|^{-(\tau-K+N+1)/2} \exp\left(\frac{1}{2}\text{tr}\left(\Sigma^{-1}g(Y_0 - X_0\hat{B}_0)'(Y_0 - X_0\hat{B}_0)\right)\right).
\end{aligned} \tag{67}$$

Comparing the two terms with the formulas for the Inverted Wishart density (22) and the matrix-normal density (23), we see that for our hypothetical sample Y_0 ,

$$\begin{aligned}
(\Sigma|Y_0, X_0) &\sim IW(C, \nu), \\
C &= g(Y_0 - X_0\hat{B}_0)'(Y_0 - X_0\hat{B}_0), \\
\nu &= \tau - K,
\end{aligned} \tag{68}$$

and

$$\begin{aligned}
\text{vec}(\Pi|\Sigma, Y_0, X_0) &\sim N(\text{vec}(\hat{B}_0), \Sigma \otimes Q), \\
Q &= (gX_0'X_0)^{-1},
\end{aligned} \tag{69}$$

with \hat{B}_0 given in (66). We take these C , ν , P and Q as the parameters of the prior distribution for our observed sample Y .

The relatively objective manner in which the actual parameter values are derived (starting from an uninformative prior and using the hypothetical sample to deduce C , ν , P and Q) is probably the main advantage of the g-prior. It is also the reason why this prior should be acceptable to those who feel that Bayesian analysis suffers from the subjectivity entering via the researcher's selection of an appropriate prior.

References

- Andersson, M. K., and S. Karlsson (2007). Bayesian Forecast Combination for VAR Models. Sveriges Riksbank Working Paper 216.
- Ang, A., Dong, S., and M. Piazzesi (2007). No-Arbitrage Taylor Rules. National Bureau of Economic Research Working Paper 13448.
- Ang, A., and M. Piazzesi (2003). A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables. *Journal of Monetary Economics*, 50, 745-787.
- Bates, J. M., and C. W. J. Granger (1969). The Combination of Forecasts. *Operational Research Quarterly*, 20(4), 451-468.
- Bolder, D. J. (2007). Modelling Term-Structure Dynamics for Risk Management: A Practitioner's Perspective. Bank of Canada Working Paper 2006-48.
- Bolder, D. J., and S. Gusba (2002). Exponentials, Polynomials, and Fourier Series: More Yield Curve Modelling at the Bank of Canada. Bank of Canada Working Paper 2002-29.
- Bolder, D. J., and S. Liu (2007). Examining Simple Joint Macroeconomic and Term-Structure Models: A Practitioner's Perspective. Bank of Canada Working Paper 2007-49.
- Bolder, D. J., and T. Rubin (2007). Optimization in a Simulation Setting: Use of Function Approximation in Debt Strategy Analysis. Bank of Canada Working Paper 2007-13.
- Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*, 2nd ed. Wiley, Hoboken, New Jersey.
- Cairns, A. J. G. (2004). A Family of Term-Structure Models for Long-Term Risk Management and Derivative Pricing. *Mathematical Finance*, 14(3), 415-444.
- Carriero, A. (2007). Forecasting the Yield Curve Using Priors from No Arbitrage Affine Term Structure Models. Queen Mary, University of London, Working Paper 612.
- Clyde, M., and E. I. George (2004). Model Uncertainty. *Statistical Science*, 19(1), 81-94.
- Dai, Q., and K. J. Singleton (2000). Specification Analysis of Affine Term Structure Models. *Journal of Finance*, 55(5), 1943-1978.
- Diebold, F. X., and C. Li (2003). Forecasting the Term Structure of Government Bond Yields. National Bureau of Economic Research Working Paper 10048.
- Diebold, F. X., and P. Pauly (1987). Structural Change and the Combination of Forecasts. *Journal of Forecasting*, 6, 21-40.
- Diebold, F. X., Rudebusch, G. D., and S. B. Aruoba (2006). The Macroeconomy and the Yield Curve: a Dynamic Latent Factor Approach. *Journal of Econometrics*, 131, 309-338.
- Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), 45-97.

- Duffee, G. R. (2002). Term Premia and Interest Rate Forecasts in Affine Models. *Journal of Finance*, 57(1), 405-443.
- Duffie, D., Filipovic, D., and W. Schachermayer (2003). Affine Processes and Applications in Finance. *Annals of Applied Probability*, 13(3), pp. 984-1053.
- Eklund, J., and S. Karlsson (2007). Forecast Combination and Model Averaging Using Predictive Measures. *Econometric Reviews*, 26(2-4), 329-363.
- Fernandez, C., Ley, E., and M. F. J. Steel (2001). Benchmark Priors for Bayesian Model Averaging. *Journal of Econometrics*, 100, 381-427.
- Geweke, J., and C. Whiteman (2006). Bayesian Forecasting, in *Handbook of Economic Forecasting*, vol. 1, G. Elliott, C. W. J. Granger and A. Timmermann (Eds), North-Holland.
- Gupta, A. K., and D. K. Nagar (1999). *Matrix Variate Distributions*. Chapman and Hall/CRC Press.
- Hall, S. G., and J. Mitchell (2007). Combining Density Forecasts. *International Journal of Forecasting*, 23, 1-13.
- Hendry, D. F., and M. P. Clements (2004). Pooling of Forecasts. *Econometrics Journal*, 7, 1-31.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and C. T. Volinsky (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4), 382-417.
- Kadiyala, K. R., and S. Karlsson (1997). Numerical Methods for Estimation and Inference in Bayesian VAR-Models. *Journal of Applied Econometrics*, 12, 99-132.
- Kaminska, I., and A. Carriero (2007). No-Arbitrage Restrictions and Yield Curve Forecasting. Available at SSRN: <http://ssrn.com/abstract=1086405>.
- Kapetanios, G., Labhard, V., and S. Price (2005). Forecasting Using Bayesian and Information Theoretic Model Averaging: an Application to UK Inflation. Bank of England Working Paper 268.
- Kapetanios, G., Labhard, V., and S. Price (2006). Forecasting Using Predictive Likelihood Model Averaging. *Econometric Letters*, 91, 373-379.
- Kass, R. E., and A. E. Raftery (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773-795.
- Koop, G. (2006). *Bayesian Econometrics*. Wiley, Chichester, West Sussex.
- Koop, G., and S. Potter (2003). Forecasting in Dynamic Factor Models Using Bayesian Model Averaging. *Econometrics Journal*, 7, 550-565.
- Leippold, M., and L. Wu (2000). Quadratic Term Structure Models. Swiss Institute of Banking and Finance Working Paper.
- Li, F., and G. Tkacz (2004). Combining Forecasts with Nonparametric Kernel Regressions. *Studies in Nonlinear Dynamics and Econometrics*, 8(4), article 2.

- Litterman, R. B. (1986). Forecasting with Bayesian Vector Autoregressions - Five Years of Experience. *Journal of Business and Economic Statistics*, 4(1), 25-38.
- Litterman, R., and J. Scheinkman (1991). Common Factors Affecting Bond Returns. *Journal of Fixed Income* 1, 54-61.
- Madigan, D., and A. E. Raftery (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, 89(428), 1535-1546.
- 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.
- Mitchell, J., and S. G. Hall (2005). Evaluating, Comparing and COmbining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR "Fan" Charts of Inflation. *Oxford Bulletin of Economics and Statistics*, 67, 995-1033.
- Moench, E. (2006). Forecasting the Yield Curve in a Data-Rich Environment: A No-Arbitrage Factor-Augmented VAR Approach. Humboldt University, Berlin, Working Paper.
- Nelson, C. R., and A. F. Siegel (1987). Parsimonious Modeling of Yield Curves. *Journal of Business*, 60(4), 473-489.
- Newbold, P., and C. W. J. Granger (1974). Experience with Forecasting Univariate Time Series and the Combination of Forecasts. *Journal of the Royal Statistical Society, Series A*, 137, 131-165.
- Poirier, D. J. (1995). *Intermediate Statistics and Econometrics*. MIT Press, Cambridge, Massachusetts.
- De Pooter, M., Ravazzolo, F., and D. van Dijk (2007). Predicting the Term Structure of Interest Rates: Incorporating Parameter Uncertainty, Model Uncertainty and Macroeconomic Information. Tinbergen Institute Discussion Paper TI 2007-028/4.
- Raftery, A. E., Madigan, D., and J. A. Hoeting (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437), 179-191.
- Ravazzolo, F., van Dijk, H. K., and M. Verbeek (2007). Predictive Gains from Forecast Combinations Using Time-Varying Model Weights. *Econometric Institute Report* 2007-26.
- Sephton, P. Forecasting Recessions: Can We Do Better on MARS? *Federal Reserve Bank of St. Louis Review*, 83(2), 39-49.
- Stock, J. H., and M. W. Watson (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics*, 20(2), 147-162.
- Stock, J. H., and M. W. Watson (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting*, 23, 405-430.
- Swanson, N. R., and T. Zeng (2001). Choosing Among Competing Econometric Forecasts: Regression-based Forecast Combinations Using Model Selection. *Journal of Forecasting*, 20, 425-440.

Timmermann, A. (2006). Forecast Combinations, in Handbook of Economic Forecasting, vol. 1, G. Elliott, C. W. J. Granger and A. Timmermann (Eds), North-Holland.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. Wiley, New York.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions, in Bayesian Inference and Decision Techniques, P. Goel and A. Zellner (Eds), North-Holland.