Infinite Markov Pooling of Predictive Distributions*

Xin Jin† John M. Maheu‡ Qiao Yang§

November 2021

Abstract

This paper introduces novel approaches to forecast pooling methods based on a nonparametric prior for a weight vector combining predictive densities. The first approach places a Dirichlet process prior on the weight vector and generalizes the static linear pool. The second approach uses a hierarchical Dirichlet process prior to allow the weight vector to follow an infinite hidden Markov chain. This generalizes dynamic prediction pools to the nonparametric setting. Efficient posterior simulation based on MCMC methods is also examined. Detailed applications to short-term interest rates, realized covariance matrices and asset pricing models demonstrate that the nonparametric pool forecasts well. The paper concludes with extensions and an application for calibrating and combining predictive densities.

Key words: Prediction pools, Dirichlet process, Beam sampling, Infinite Markov switching, density forecast, short-term interest rates, realized covariance matrices

JEL: C53, C32, C11, Q43

---

*We thank the editor Serena Ng, an associate editor and two anonymous referees for their many helpful comments. We are also grateful for the helpful comments of seminar participants at 2017 RECA Bayesian Econometric workshop, the 2018 SBIIES at Brown University, the 2018 ESOBE at New Orleans, the 2019 CFE at London, the 2021 ISBA online, Hunan University, Tongji University, Shanghai University of Finance and Economics, and ShanghaiTech University. Jin’s research is supported by NSFC through Project 71773069. Maheu thanks SSHRC of Canada for financial support. Yang thanks the start-up fund of ShanghaiTech and Young Scientists Fund of NSFC (Project 72103137) for financial support. A previous version of this manuscript was titled Bayesian Nonparametric Forecast Pooling.

†School of Economics, Shanghai University of Finance and Economics, China. Email: jin.xin@mail.shufe.edu.cn
‡DeGroote School of Business, McMaster University, Canada. Email: maheujm@mcmaster.ca
§Corresponding author: School of Entrepreneurship and Management, ShanghaiTech University, China. Email: yangqiao@shanghaitech.edu.cn
1 Introduction

This research introduces several nonparametric extensions to prediction pools using Dirichlet process based priors to impose structure on the support of model weights and how they change over time. Detailed applications to short-term interest rates, realized covariance matrices, and asset pricing models demonstrate the superior performance of forecasts of the nonparametric pool.

Following the work of Hall & Mitchell (2007) and Geweke & Amisano (2011), considerable interest has been paid toward model pooling methods, presenting a density combination approach that combines several predictive densities to form a forecast. A significant feature of this approach, in contrast to Bayesian model averaging, is that it recognizes that a true model may not be present in a model set. This density combination approach has been subject to many applications and important extensions that allow weights on individual densities to change over time, including Billio et al. (2013), Waggoner & Zha (2012), and Del Negro et al. (2016). Other recent important contributions are Bayesian predictive synthesis (McAlinn & West 2019, McAlinn et al. 2020), nonparametric calibration and combination of predictive distributions (Bassetti et al. 2018), generalized density forecast combination (Kapetanios et al. 2015), and higher-order moment constraints in predictive density combination (Pauwels et al. 2020).

This paper contributes to the forecast pooling literature by proposing an innovative approach, demonstrating how a nonparametric prior can be imposed on the weight vector to flexibly combine models. A Dirichlet process prior can be used to allow for countably infinite support of the weight vector combining models. This is done in Bassetti et al. (2018) combining calibrated predictive densities with random combination weights. Although the Dirichlet process prior can be used to generalize the approach of Geweke & Amisano (2011) to a Bayesian nonparametric setting, our main model adopts the hierarchical Dirichlet process (HDP) of Teh et al. (2006). This results in the dynamic weight vector following an infinite hidden Markov structure. As such, this is a very flexible approach, and due to the infinite support for the weight vector, this class of forecast pooling models is referred to as infinite Markov pooling (IMP).

IMP nests the case of one constant vector of weights as in Geweke & Amisano (2011) but allows for as many states as needed to support the data. By switching between states, IMP captures time variation in model weights through discrete changes. The number of active states can change over time and bypasses the challenge of estimating the dimension of the Markov chain. States are allowed to remain persistent and a prior on persistence is set through the sticky infinite hidden Markov model of Fox et al. (2011). IMP can be considered an extension to the finite state Markov switching model combination of Waggoner & Zha (2012).

Estimation of IMP involves two steps. In the first step, individual models produce a predictive likelihood and any additional quantities of interest, such as a predictive mean. Following this, IMP takes the individual models’ predictive likelihoods as data, applying posterior simulation methods to estimate and combine these individual models assuming the model probability weights are governed by an infinite Markov chain. A new posterior simulation is presented that jointly samples the latent state and the model indicator by extending the beam sampler approach of Van Gael et al. (2008). This makes posterior inference on model weights simple and leads to better mixing of the Markov chain of the posterior distribution. This two-step approach to estimation means that although we use a nonparametric prior, a fairly large number of individual models can be pooled together. The first application pools 20 models.

From the posterior simulation output, a density forecast or other features of the predictive distribution from the combined models can be computed. For instance, simulated values or quantiles

---

can be estimated by standard posterior simulation methods that simulate from the predictive density of individual models and that of the IMP specification. These forecasts consider all past active states as well as the possibility of new future states.

The superiority of our pooling approach is that it allows the pooling weights to change over time in a structured yet suitably flexible way that it can accommodate periods with persistent weights as well as those featuring abrupt weight changes. The nonparametric nature of the prior makes the number of the states (or weight vectors) unbounded and new weight vectors can be added if necessary as it moves out-of-sample and old states can be pruned when they are no longer needed. Additionally, parsimony is achieved with high state persistence that allows the weight vector to remain constant, while introducing many new states allows for smooth rapid changes in the weight vector when needed.

IMP is applied to three empirical applications to assess its strengths and weaknesses. The findings indicate that, overall, IMP can elicit significant improvement in the accuracy of density forecasts, as measured by log-predictive likelihoods. This is not surprising, as the main building block of the model, the infinite Markov chain, is derived from Bayesian dynamic density estimation (Teh et al. 2006). We compare to several state-of-the-art model combination approaches and demonstrate that IMP is a dominant competitor for our applications. Point forecasts, in the form of the predictive mean, show no consistent pattern of improvement over individual models for any model combination approach.

The first application is to short-term interest rates. Over a range of model classes for interest rates, the IMP delivers the best density forecasts among all the pooling methods. Using superior individual forecasting models improves all combination methods and including all reasonable candidates are generally recommended, as pooling of 10 and 20 models results in the best performing IMP. These results are robust to subsamples and various prior settings. There are considerable dynamic changes in weights in contrast to the fixed weights of Geweke & Amisano (2011). Meanwhile, there is posterior support for 3 to 9 different probability weight vectors (states) making a fixed setting of two or another finite number, as in Waggoner & Zha (2012), problematic.

The second application is multivariate and considers forecasts of realized covariance matrices. This application indicates that IMP is useful in higher dimensional settings; in this case, positive definite matrices of dimension 10. Most of the model combination approaches considered improve density forecasts of realized covariance matrices compared to the individual models. The IMP approach has a log-Bayes factor of 48 in its favor against the second best performer of Waggoner & Zha (2012). This is due to the flexibility of the endogenous number of states the nonparametric prior allows for.

The final application predicts monthly returns for 10 industry portfolios. Individual models include popular risk premium specifications with and without GARCH type heteroskedasticity. IMP delivers competitive density forecast improvements compared to other combination methods. Weights from the IMP are more stable compared to the other applications but do display occasional abrupt changes.

Although our focus is on adding the infinite hidden Markov structure to linear pooling methods, the hierarchical prior can be used to extend other model combination approaches, such as Bayesian predictive synthesis or calibration and combination of predictive distributions. For example, we show how to extend the univariate Dirichlet based nonparametric approach of Bassetti et al. (2018) to dynamic weights via the hierarchical prior. An application to interest rates generates some of the best forecasting performance.

The remainder of this paper is organized as follows. The next section reviews existing model combination approaches that are included in the empirical applications for comparison. Section 3 considers a basic Bayesian nonparametric pooling model using a Dirichlet process prior. We discuss
how to extend this to IMP, the main new approach presented in this study. Posterior sampling and computation of forecasts are discussed. Section 4 provides a detailed analysis of the methods applied to short-term interest rate models. Section 5 applies IMP to realized covariance matrix forecasts and Section 6 presents an application to returns from 10 industry portfolios. Section 7 discusses extensions to McAlinn & West (2019) and Bassetti et al. (2018) with the infinite hidden Markov structure and provides an application. Section 8 concludes, while the Online Appendix contains detailed steps of posterior simulation and additional results.

2 Existing Forecast Combination Approaches

In the following, the $p \times 1$ vector of interest is denoted as $y_t$, the past information set as $y_{1:t} = \{y_1, \ldots, y_t\}$ and models as $M_l$, $l = 1, \ldots, L$. Each of the models will produce a predictive density for $y_t$ given $y_{1:t-1}$ but they could also exploit additional data $x_{1:t-1}$, which is suppressed to minimize notation. Next we consider some benchmark density combination approaches, followed by a presentation of our approach.

2.1 Bayesian Model Averaging

Bayesian model averaging (BMA) assumes a complete model space, in which one member of the set $\{M_1, \ldots, M_L\}$ of models is correct. In this setting the predictive density is formed as

$$ p(y_t|y_{1:t-1}) = \sum_{l=1}^L p(y_t|y_{1:t-1}, M_l)p(M_l|y_{1:t-1}), \quad (1) $$

where $p(M_l|y_{1:t-1}) \propto p(y_{t-1}|y_{1:t-2}, M_l)p(M_l|y_{1:t-2})$. $p(y_t|y_{1:t-1}, M_l) = \int p(y_t|\theta_l, M_l)p(\theta_l|y_{1:t-1}, M_l)d\theta_l$ is the predictive density of model $M_l$ and $\theta_l$ is a data density parameter that is integrated out. As pointed out in Geweke & Amisano (2011) and Del Negro et al. (2016), given a stable data generating process the posterior probability for the model that minimizes the Kullback–Leibler distance will tend toward one as the sample size grows. The remaining forecast combination approaches do not suffer from this. Extensions to allow for time-varying weights in the context for BMA have been proposed.\(^2\)

2.2 Static Pooling: Optimal Pooling

Geweke & Amisano (2011) and Hall & Mitchell (2007) propose a model combination setting which is incomplete in that the true model is not assumed to belong to the set of candidate forecasting models. Geweke & Amisano (2011) (GA) optimal prediction pool is obtained by solving for the weights $\omega = \{\omega_1, \ldots, \omega_L\}$, $\omega_l \geq 0$, that $\sum_{l=1}^L \omega_l = 1$ as

$$ \max_{\omega} \sum_{t=1}^T \log \left( \sum_{l=1}^L \omega_l p(y_t|y_{1:t-1}, M_l) \right). \quad (2) $$

Although static pooling can result in significant improvement in density forecasting, as BMA it does not capture time-varying dynamics in weights.

Among other static approaches Kascha & Ravazzolo (2010) assign weights by a rule of thumb, whereas Kapetanios et al. (2015) propose a generalized version by imposing a rolling window as a threshold. Their work clarifies that combination weights need to be time-varying.

\(^2\)Hoogerheide et al. (2010) suggest an extension, in which the weights are allowed to be time-varying by imposing a random walk on weights’ dynamics. Billio et al. (2012) apply BMA to turning point predictions.
2.3 Dynamic Pooling: Autoregressive Weights

Del Negro et al. (2016) use dynamic pooling to combine forecasts from two DSGE models. In their method, weights are \( \omega_t = (\omega_{t,1}, \ldots, \omega_{t,L}) = g(X_t) \) where \( g(\cdot) \) is a function that maps the \( L \times 1 \) stochastic vector \( X_t \) to a discrete probability density of size \( L \), where \( \omega_{t,l} \geq 0 \) and \( \sum_{l=1}^{L} \omega_{t,l} = 1 \). The version we used in this paper is

\[
X_{t,i} = (1 - \rho)\mu + \rho X_{t-1,i} + \sqrt{1 - \rho^2} \epsilon_{t,i}, \quad \epsilon_{t,i} \sim i.i.d. N(0, 1), \quad X_{0,i} \sim N(\mu, \sigma^2), \quad i = 1, \ldots, L, \tag{3a}
\]

\[
p(y_t|y_{1:t-1}, \omega_t) = \sum_{l=1}^{L} \omega_{t,l}p(y_t|y_{1:t-1}, M_l), \quad \omega_{t,i} = \exp(X_{t,i})/\sum_{j=1}^{L} \exp(X_{t,j}). \tag{3b}
\]

In this approach, weights are directed by a set of univariate autoregressions for \( X_{t,i} \) and the logistic transformation maps this to a probability weight vector. Scalar values of \( \rho \) closer to 1 and/or smaller values of \( \sigma^2 \) translate into more persistent weights \( \omega_t \) through time. Since this is a nonlinear state-space model, we follow Del Negro et al. (2016) and sample the weights by using a bootstrap particle filter.

Let (3) be our benchmark, and label it DHS. The same approach as Del Negro et al. (2016) is used to update \( \rho \), whereas \( \mu = 0 \) and \( \sigma = 1 \) are fixed in the estimation.

2.4 Dynamic Pooling: Markov Weights

In contrast to the previously introduced weight dynamics, Waggoner & Zha (2012) assume the weights follow a two-state Markov chain. This approach is labeled WZ2 and takes the form

\[
p(y_t|y_{1:t-1}, \Pi, s_{t-1}, \omega) = \sum_{s_{t-1}} \pi_{s_{t-1},st} \sum_{l=1}^{L} \omega_{st,l}p(y_t|y_{1:t-1}, M_l), \quad s_t | s_{t-1} \sim \Pi_{s_{t-1}}, \quad s_t \in \{1, 2\}, \tag{4}
\]

where \( \Pi_{s_{t-1}} \) denotes the row of the transition matrix \( \Pi \) and governs the probability of the next state after \( s_{t-1} \), while \( \omega_{st} = (\omega_{st,1}, \ldots, \omega_{st,L}) \). This is a Markov switching model of predictive densities and allows the model weights \( \omega_{st} \) to change according to a two-state Markov chain. The prior of \( \omega_{st} \sim Dir(1, 1) \) is assumed. This model can be easily sampled through the forward-filter backward-sampling (FFBS) of Chib (1996).

2.5 Forecast Combination via State-Space Representation

Billio et al. (2013) (BCRV) combine predictors from models allowing for time-varying weights and model the distribution of observables and predictors using a potentially nonlinear state-space model.\(^4\) The predictor is simulated from its predictive distribution given a model instead of, for example, relying on the predictive likelihood. The sampling algorithm is based on bootstrap particle filtering. Billio et al. (2013) introduce an informative prior to the particle generating process.\(^5\) We

---

\(^3\)In Waggoner & Zha (2012), the most general specification they consider jointly estimates the Markov switching process and the parameters of each state specific data density. However, their joint sampling approach is limited to models without path dependence. For instance, Markov switching and GARCH models induce path dependence that increases model complexity and computational cost of estimation (Bauwens et al. 2014).

\(^4\)Aastveit et al. (2018) use the same method for Macroeconomic nowcasting.

\(^5\)Weight dynamics follow an exponentially weighted learning strategy.
consider the following version

\[ X_{t,i} = \rho X_{t-1,i} + \sigma_{\epsilon_{t,i}} \epsilon_{t,i} \sim N(0,1), \quad \omega_t = g(X_t), \quad (5a) \]

\[ p(y_t|\hat{y}_t, \omega_t) = (2\pi \sigma^2)^{-1/2} \exp \left( -\frac{1}{2\sigma^2} \left( y_t - \sum_{l=1}^{L} \omega_{t,l}\hat{y}_{t,l} \right) \right), \quad (5b) \]

\[ \hat{\gamma}_{t,l} \sim p(\hat{y}_t|y_{1:t-1}, M_l) \quad \text{for} \quad l = 1, \ldots, L, \quad (5c) \]

where \( \hat{y}_{t,l} \) are simulated values from a model’s predictive density and combined in (5b). The predictive density is formed as

\[ p(y_t|y_{1:t-1}) = \int p(y_t|\hat{y}_t)p(\hat{y}_t|y_{1:t-1})d\hat{y}_t, \quad (6) \]

where \( p(y_t|\hat{y}_t) = \int p(y_t|\hat{y}_t, \omega_t)p(\omega_t|y_{1:t-1})d\omega_t \). The last two integrals can be approximated using Monte Carlo methods and will naturally generate a mixture of distributions for the predictive density; however, in contrast to the other methods, the mixing occurs through the mean of (5b) and not the variance. As a result, this model may have difficulty in fully capturing fat tails and heteroskedasticity over time. \( \omega_t \) and \( g(X_t) \) are defined in the same way as DHS in Section 2.3. We set \( \rho = 0.8 \) and \( \sigma_{\rho} = \sigma = 1 \) and sample the model weights using the bootstrap particle filter with pre-simulated \( \hat{y} \) from candidate models.

### 2.6 Bayesian Predictive Synthesis

McAlinn & West (2019) propose a general version of Billio et al. (2013) called Bayesian predictive synthesis (BPS). Rather than taking the distribution of latent predictors as given, McAlinn & West (2019) sample them conditioned on weights and a synthesis function that can capture important model dependencies, allowing for the time-varying change of weights via a random walk scheme (autoregressive change of weights). Due to the synthesis feature, the weights in this method are unrestricted. The following BPS model, labeled MW, is considered in the following as

\[ y_t = \omega_{t,0} + \sum_{j=1}^{L} \omega_{t,j}\hat{y}_{t,j} + e_t, \quad e_t \sim N(0, \nu_t), \quad (7a) \]

\[ \omega_t = \omega_{t-1} + w_t, \quad w_t \sim N(0, \nu_t W_t), \quad (7b) \]

where \( \nu_t \) is defined via a standard beta-gamma random walk volatility model and \( W_t \) follows a standard single discount factor specification. Let \( \hat{y}_t = \{\hat{y}_{t,1}, \ldots, \hat{y}_{t,L}\} \) be sampled latent predictors, then the conditional posterior is

\[ p(\hat{y}_t|y_{1:t}, \omega_{t}, M_{1:L}) \propto \alpha(y_t|\omega_t, \hat{y}_t) \prod_{j=1}^{L} p(\hat{y}_{t,j}|y_{1:t-1}, M_j), \quad (8) \]

where \( \alpha(y_t|\omega_t, \hat{y}_t) \) denotes the synthesis function based on (7a). \( \hat{y}_t \) is sampled from (8), while in the Billio et al. (2013) approach \( \alpha(y_t|\omega_t, \hat{y}_t) \) is excluded. The dynamic weights are sampled through a FFBS and the predictors \( \hat{y}_t \) are sampled via conjugacy (see McAlinn & West (2019) and their Online Appendix for details).
3 Bayesian Nonparametric Prediction Pooling

To begin consider a simple Bayesian pooling approach for $L$ models

$$p(y_t|y_{1:t-1}, \omega) = \sum_{l=1}^{L} \omega_l p(y_t|y_{1:t-1}, M_l), \quad \omega \sim H, \tag{9}$$

where $H$ is the prior distribution for the weight vector, $\omega = (\omega_1, \ldots, \omega_L)$. This is a Bayesian analogue of Geweke & Amisano (2011) with one vector parameter $\omega$ to be estimated. The predictive density is formed in the usual way by integrating out parameter uncertainty from the posterior of $\omega$.

There are several Bayesian nonparametric extensions possible. The simplest is to replace the prior $H$ with a Dirichlet process (DP) prior. This model, in hierarchical form, is

$$\omega_t \text{iid} \sim G, \quad G|\alpha \sim DP(\alpha, G_0), \tag{10a}$$

$$p(y_t|y_{1:t-1}, \omega_t) = \sum_{l=1}^{L} \omega_{t,l} p(y_t|y_{1:t-1}, M_l), \tag{10b}$$

where $DP(\alpha, G_0)$ denotes the Dirichlet process prior with precision parameter $\alpha > 0$ and base distribution $G_0$. $G$ is an almost surely discrete probability distribution from which the weight vector $\omega_t$ is drawn at each time period.\(^{6}\) This model is a Dirichlet process mixture model (Antoniak 1974). In large samples parameter uncertainty from the weight vector for these last two models will be small and the posterior will peak around the mode resulting in a predictive density very similar to Geweke & Amisano (2011) which selects the mode and has fixed probability weights.

3.1 Infinite Markov Pooling

The main nonparametric specification we focus in this study replaces the DP prior proposed above with the hierarchical Dirichlet process (HDP) of Teh et al. (2006). The resulting model can be thought of as extending Waggoner & Zha (2012), from a two-state, to an infinite state Markov chain. Although this framework continues to combine $L$ models, the possible ways of combining models become unbounded. As such, it accommodates persistent changes in weights like Del Negro et al. (2016), in addition to abrupt structural changes. For the latter effect new weight vectors can be introduced through time as new combinations of models are preferred compared to past combinations. This makes the approach very flexible.

We refer to this approach as infinite Markov pooling, which is denoted as IMP. Here, infinite refers to the unbounded potential number of weight vectors (states) used to combine the finite $L$

\(^{6}\)A draw of $G$ can be represented as

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}, \quad \omega_i \text{iid} \sim G_0, \quad i = 1, \ldots, \infty$$

$$\pi_i = v_i \prod_{j=1}^{i-1} (1 - v_j), \quad v_i \text{iid} \sim Beta(1, \alpha), \quad j = 1, 2, \ldots.$$
models. The model is
\begin{equation}
\Gamma \sim Stick(\eta), \quad \Pi_i \overset{i.i.d.}{\sim} DP \left( \alpha + \kappa, \frac{\alpha \Gamma + \kappa \delta_i}{\alpha + \kappa} \right), \quad i = 1, 2, \ldots, \quad (12a)
\end{equation}

\begin{equation}
f(y_t|I_{t-1}, s_t) = \sum_{l=1}^{L} \omega_{s_t,l} p(y_t|I_{t-1}, M_l), \quad s_t|s_{t-1} \sim \Pi_{s_{t-1}}, \quad (12b)
\end{equation}

\begin{equation}
\omega_i \sim Dir \left( \frac{\alpha_\omega}{L}, \ldots, \frac{\alpha_\omega}{L} \right), \quad i = 1, 2, \ldots, \quad (12c)
\end{equation}

where \( \omega_i = (\omega_{i,1}, \omega_{i,2}, \ldots, \omega_{i,L}) \) is an \( L \)-dimensional vector of model weights corresponding to each state \( s_t = i \). \( \omega_{s,t} \) changes over time according to the first-order Markov chain with infinite dimension. \( Dir(\cdot) \) stands for the Dirichlet distribution with dimension \( L \). State variable \( s_t \) follows an infinite Markov transition matrix, \( \Pi_t \), with priors governed by an HDP and \( \Pi_{s_{t-1}} \) denotes a row of the transition matrix given the previous state \( s_{t-1} \). This version of the HDP is the sticky version of Fox et al. (2011) and allows for estimation of state persistence. The term \( \kappa \delta_i \) indicates that to element \( \alpha \Gamma_i \) (\( i^{th} \) element) is added \( \kappa \geq 0 \). Through \( \kappa \), state persistence can be reinforced.

\( \eta > 0 \) and \( \alpha > 0 \) are two layers of precision parameters that govern the likelihood of introducing new states in the HDP. Small values of \( \eta \) and \( \alpha \) promote parsimony of states, while large values are consistent with a higher likelihood of new states being introduced.

It can be helpful to view this model as a stick breaking process (Sethuraman 1994). Let \( \Gamma = \{\gamma_1, \ldots, \gamma_\infty\} \) and \( \pi_{ij} \) be the \( i^{th} \) row and \( j^{th} \) column of \( \Pi \). The distributional draw \( \Gamma \sim Stick(\eta) \) can be represented as \( \Gamma = \sum_{i=1}^{\infty} \gamma_i \delta_i \), where \( \delta_i \) is a point mass at \( i \) and \( \gamma_i \) is the associated probability mass that is generated as

\begin{equation}
\gamma_i = v_i \prod_{l=1}^{i-1} (1 - v_l), \quad v_j \overset{i.i.d.}{\sim} Beta(1, \eta), \quad j = 1, 2, \ldots, \quad (13)
\end{equation}

Similarly, \( \Pi_i \sim DP(\alpha, \Gamma) \) can be represented as \( \Pi_i = \sum_{j=1}^{\infty} \pi_{ij} \delta_j \), where the probability weights \( \pi_{ij} \) are generated as

\begin{equation}
\pi_{ij} = \hat{\pi}_{ij} \prod_{l=1}^{j-1} (1 - \hat{\pi}_{il}), \quad \hat{\pi}_{ij} \overset{i.i.d.}{\sim} Beta \left( \alpha \gamma_j + \kappa \delta_i, \alpha + \kappa - j \sum_{l=1}^{j} (\alpha \gamma_l + \kappa \delta_l) \right), \quad j = 1, 2, \ldots, \quad (14)
\end{equation}

Each row \( \Pi_i \) of the transition matrix is centered on \( E(\pi_{ij}) = \alpha \Gamma_i/(\alpha + \kappa) \) for \( i \neq j \) and \( E(\pi_{ii}) = (\alpha \Gamma_i + \kappa)/(\alpha + \kappa) \) for \( i = j \). Larger values of \( \kappa \) favor self transition of states, while \( \kappa = 0 \) gives the standard HDP for infinite hidden Markov models. As the magnitudes of \( \eta \) and \( \alpha \) increase, the probability mass is dispersed among a greater number of states. Due to the importance of these parameters, we place the following priors on them and learn their values from the data, \( \eta \sim Gamma(a_1, b_1), \alpha + \kappa \sim Gamma(a_2, b_2), \rho \sim Beta(a_3, b_3) \) with \( \rho = \alpha/(\alpha + \kappa) \), which are easier to sample. A hyper-prior is set on \( \alpha_\omega \) as \( \alpha_\omega \sim Gamma(a_4, b_4) \). Setting a hyper-prior on \( \alpha_\omega \) makes sampling the posterior of \( \omega_k \) within each state group more flexible than a fixed \( \alpha_\omega \). A larger value of \( \alpha_\omega \) is more informative and favors equal weights. Empirically the forecasting results from the model are found to be better with smaller \( \alpha_\omega \), which is one reason that estimating this parameter is favored, rather than using a preset value.

After marginalizing over \( s_t \), the predictive density of \( y_t \) given \( s_{t-1} \) is determined by

\begin{equation}
p(y_t|y_{1:t-1}, s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} p(y_t|y_{1:t-1}, s_t) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \sum_{l=1}^{L} \omega_{s_t,l} p(y_t|y_{1:t-1}, M_l), \quad (15a)
\end{equation}

where \( p(y_t|y_{1:t-1}, M_l) \) is the predictive density at time \( t \) of the individual model \( M_l \).

\footnote{In practice, posterior simulation will explore a finite number of weight vectors for a finite dataset.}
3.2 Posterior Sampling

To facilitate sampling, an indicator variable, \( z_t \in \{1, 2, \ldots, L \} \), is introduced for indexing the model assigned to observation \( t \). The state variable \( s_t \) and model variable \( z_t \) are jointly sampled. Our mechanism for sampling \( z_{1:T} \) is an important contribution, as jointly sampling \((s_t, z_t)\) makes posterior inference on model weights simple and leads to better mixing of the Markov chain of the posterior distribution.

We extend the beam sampler approach of Van Gael et al. (2008) to sample the state space \((s_{1:T}, z_{1:T})\). This introduces auxiliary variables that stochastically truncate the infinite state space to a finite space, after which FFBS can be used to sample states. Marginalizing over the auxiliary variables gives the desired posterior distribution.

Define the auxiliary latent variable \( u_t > 0 \) (slices) with the following uniform density

\[
p(u_t|s_t, s_{t-1}, z_t, \Pi, \omega) = \frac{1(u_t < \pi_{s_{t-1}, s_t} \omega_{s_t, z_t})}{\pi_{s_{t-1}, s_t} \omega_{s_t, z_t}}.
\]

Define the natural number \( K \) such that the set \( \{ s_t, s_t < K \} \) contains all instances of \( u_t < \pi_{s_{t-1}, s_t} \omega_{s_t, z_t} \) for each \( t \). Inclusivity is guaranteed if \( K \) satisfies \( \max_{i \in \{1, \ldots, K\}} \{1 - \sum_{j=1}^{K} \pi_{i,j}\} < \min_{t \in \{1, \ldots, T\}} \{u_t\} \).

With this, the infinite outer summation in (15a) is reduced to, at most, \( K \) non-zero terms and variables \( s_t \) and \( z_t \) can be sampled jointly in the following way. Define \( \omega = (\omega_1, \ldots, \omega_K) \), and each of its elements as \( \omega_j = (\omega_{j,1}, \ldots, \omega_{j,L}) \) for \( j = 1, \ldots, K \), then iterate over the following steps. From \( t = 1, \ldots, T \), repeat the following forward filter steps:

Prediction step: for \( k = 1, \ldots, K, l = 1, \ldots, L \) calculate

\[
p(s_t = k, z_t = q|u_{1:T}, \Pi, \omega, y_{1:t-1}) \propto \sum_{j=1}^{K} \sum_{l=1}^{L} \mathbf{1}(u_t < \pi_{j,k} \omega_{j,l}) p(s_{t-1} = j, z_{t-1} = l|u_{1:T}, \Pi, \omega, y_{1:t-1}).
\]

Update step: for \( k = 1, \ldots, K, l = 1, \ldots, L \) calculate

\[
p(s_t = k, z_t = q|u_{1:T}, \Pi, \omega, y_{1:t}) \propto p(s_t = k, z_t = q|u_{1:T}, \Pi, \omega, y_{1:t-1}) p(y_t|y_{1:t-1}, M_q).
\]

Followed by the backward sampling steps.

1. Sample \((s_T, z_T)\) from \(p(s_T, z_T|u_{1:T}, \Pi, \omega, y_{1:T})\).
2. Sample \((s_t, z_t)\) from \(p(s_t, z_t|u_{1:T}, \Pi, \omega, y_{1:t}) \mathbf{1}(u_{t+1} < \pi_{s_t, s_{t+1}} \omega_{s_{t+1}, z_{t+1}})\) for \( t = T - 1, \ldots, 1 \).

After states (indexed by \( s_t \)) are sampled we track the number of active states (visited at least once) and order them as the initial \( K \) states, and sort \( \omega_{1:K} \) and \( \Pi_{1:K+1,1:K+1} \) accordingly. Each sweep of the sampler updates the value of \( K \). \{\( u_{1:T}, s_{1:T}, z_{1:T}, \eta, \alpha, \omega, \Gamma_{1:K}, \Pi_{1:K+1,1:K+1}, \omega_{1:K}, K \}\) is the parameter set. Posterior sampling is sequentially repeated from the following conditional posterior distributions:

\[
p(u_{1:T}|s_{1:T}, z_{1:T}, \Pi_{1:K+1,1:K+1}, \omega_{1:K})
p(\omega_{1:K}|s_{1:T}, z_{1:T}, \alpha)
p(\Gamma_{1:K}|s_{1:T}, \eta, \alpha)
p(\Pi_{1:K+1,1:K+1}|s_{1:T}, \Gamma_{1:K}, \alpha)
p(\eta, \alpha, \rho, \kappa|s_{1:T}, \Gamma_{1:K})
p(\alpha|\omega_{1:K}).
\]
Each of these steps is further detailed in the Online Appendix.

Iterating the above steps produces posterior draws for each parameter of interest. With 20,000 burn-in draws, the posterior average of each parameter and predictive density are computed from 40,000 draws following burn-in. A parameter of interest is \( \omega_{st} \) as it indicates the model pooling dynamics at time \( t \). The posterior mean is estimated as

\[
E(\omega_{st}|y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} \omega_{st(i)}^{(i)}, \quad \text{for } t = 1, \ldots, T,
\]

where \( i \) indicates the \( i \)th MCMC draw of the associated parameter and \( \omega_{st} = \{\omega_{st,1}, \ldots, \omega_{st,L}\} \). Any posterior statistic of interest can be computed in a similar way.

Finally, no computational problems were encountered in the empirical applications; however, for larger datasets, more active states and combining more individual models will present computational challenges. Parallel programming can partly offset these challenges, but we leave this to future work.

### 3.3 Predictive Density

Given data \( y_{1:T} \) and \( N \) MCMC draws of parameters, the predictive density can be estimated directly or simulated from. The following steps are used to compute the predictive density and predictive likelihood.

1. For each MCMC draw of \( s_T \), simulate the future state \( s_{T+1} \) according to the Markov transition probability \( \Pi_{st} \).

2. If \( s_{T+1} \leq K \), set \( \omega_{s_{T+1}} \) from the existing draws of \( \omega_{1:K} \). Otherwise, set \( s_{T+1} \) as a new state generated from the prior \( \omega_{s_{T+1}} \sim \text{Dir}(\alpha_{s_{T+1}}, \ldots, \alpha_{s_{T+1}}) \).

The predictive density can be estimated as

\[
p(y_{T+1}|y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} \omega_{s_{T+1}^{(i)}}^{(i)} p(y_{T+1}|y_{1:T}, M_l),
\]

which integrates out parameter and distributional uncertainty. A predictive likelihood value is obtained by evaluating (19) at the realized data \( y_{T+1} \). Similarly, predictive moments can be estimated such as the predictive mean

\[
E(y_{T+1}|y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} \omega_{s_{T+1}^{(i)}}^{(i)} E(y_{T+1}|y_{1:T}, M_l),
\]

where \( E(y_{T+1}|y_{1:T}, M_l) \) is the predictive mean from model \( M_l \).

To evaluate predictive accuracy over the out-of-sample period \( t = \tau_1, \ldots, \tau_2, \tau_1 \leq \tau_2 \), log-predictive likelihood (LPL) and root mean squared forecast errors (RMSFE) are computed as follows

\[
\text{LPL} = \sum_{t=\tau_1}^{\tau_2} \log p(y_t|y_{1:t-1}), \quad \text{RMSFE} = \sqrt{\frac{\sum_{t=\tau_1}^{\tau_2} (E(y_t|y_{1:t-1}) - y_t)^2}{\tau_2 - \tau_1 + 1}}.
\]

In addition to the LPL, the continuous rank probability score (CRPS) defined on out-of-sample data is computed as

\[
\text{CRPS}_M = \frac{1}{\tau_2 - \tau_1 + 1} \sum_{t=\tau_1}^{\tau_2} \int_{-\infty}^{\infty} (F_{t-1}(u) - I\{u \geq y_t\})^2 du,
\]

where \( F_{t-1}(u) \) is the cumulative predictive density for the model \( M \) and is computed numerically.\(^8\)

---

\(^8\) \( F_{t-1}(y_t) \) is computed on an equidistant grid of 4,000 points with minimum and maximum values of -20 and 20.
Calculating these measures involves recursively estimating the model for each time period in the out-of-sample period.

3.4 Training Sample

The training sample is a bit more complicated than in a conventional setting. There are two layers of training sample and out-of-sample periods which must be clarified. Each individual model requires a history of data for estimation, as do the pooling models. Each individual model is assumed to use all data from period \(1\) to \(t-1\) to compute a predictive likelihood for \(y_t\). The pooling models will use data from \(\tau_0 \geq 1\) to \(t-1\) to compute a predictive density for \(y_t\). In general, a \(\tau_0\) of 1 could be used, but larger values are desirable as initial predictive likelihoods will be dominated by parameter uncertainty and the pooling model using these data to learn may degrade its forecasts. Therefore, in the applications, a \(\tau_0 > 1\) is generally used and robustness to this parameter is presented. In summary, as above, forecasts for all models are compared over a common out-of-sample period \(y_{r_1}, \ldots, y_{r_2}\). Individual models use data from \(t = 1\) and onward to compute a forecast while the pooling models use data from \(t = \tau_0\) onward.

4 Application to U.S. Short-term Interest Rate Models

There is an extensive literature devoted to the time-series dynamics of short-term T-bill rates. This section will investigate different model combination approaches using four groups of models to forecast interest rates.

4.1 Models

4.1.1 Basic Group

The following general form summarizes the models in the Basic group

\[
 r_t - r_{t-1} = \lambda + \beta r_{t-1} + e_t \quad e_t \sim N(0, \sigma^2 r_{t-1}^x),
\]

where \(r_t\) is the short-term interest rate, and the conditional variance is \(\sigma^2 r_{t-1}^x\). The first model is Vasicek (1977) and specified as \(x = 0\) (VSK). The second model is Cox et al. (1985), wherein \(x = 1\) (CIR). The next model is from Black & Scholes (1973) with restrictions of \(\lambda = 0\) and \(x = 2\) (GBM). The fourth model is denoted as MER, referring to Merton (1973), which sets \(\beta = 0\) and \(x = 0\). The final model was introduced by Brennan & Schwartz (1977), Brennan & Schwartz (1979), and Brennan & Schwartz (1980). It is denoted as BSZ and restricted by \(x = 2\). These five models are defined as the Basic group.

4.1.2 MS2 Group

Markov switching models have been a popular specification for interest rates (Ang & Bekaert 2002, Durham 2003, Pesaran et al. 2006, Guidolin & Timmermann 2009). The MS2 group contains five models that are direct extensions to the previous group with \(\lambda, \beta\) and \(\sigma\) becoming state dependent. These are

\[
 r_t - r_{t-1} = \lambda_{s_t} + \beta_{s_t} r_{t-1} + e_t \quad e_t \sim N(0, \sigma^2_{s_t} r_{t-1}^x) \quad (22a)
\]

\[
 s_t \in \{1, 2\} \quad s_t | s_{t-1} \sim \Pi. \quad (22b)
\]
VSK-MS2, CIR-MS2, and BSZ-MS2 correspond to \(x = 0\), \(x = 1\), and \(x = 2\), respectively. GBM-MS2 imposes \(\lambda_{1:2} = 0\) and \(x = 2\) on the above equation. MER-MS2 restricts \(\beta_{1:2} = 0\) and \(x = 0\). These five models form the **MS2 group**.

### 4.1.3 IHMM Group

The IHMM group contains five models, like the MS2, extending the benchmark set of models to be governed by an unobserved discrete state. But in this case, the state follows an infinite Markov chain. These models are motivated from Maheu & Yang (2016), who use them to nonparametrically model interest rate dynamics. The five models of the IHMM group are defined by

\[
\Gamma \sim \text{Stick}(\eta), \quad \Pi_i \overset{i.i.d.}{\sim} DP(\alpha, \Gamma), i = 1, 2, ..., \quad (23a)
\]

\[
r_t - r_{t-1} = \lambda s_t + \beta s_t r_{t-1} + e_t \quad e_t \sim N(0, \sigma^2_{s_t} r_{t-1}^x) \quad (23b)
\]

\[
s_t | s_{t-1} \sim \Pi_{s_{t-1}}, \quad s_t \in \{1, 2, ..., \} \quad (23c)
\]

The models of the IHMM group allow \(\lambda\), \(\beta\) and \(\sigma\) to be state dependent.

VSK-IHMM, CIR-IHMM, and BSZ-IHMM are denoted by letting \(x = 0\), \(x = 1\), and \(x = 2\) in (23). GBM-IHMM imposes \(\lambda_{1:2} = 0\) and \(x = 2\) on the above equation. MER-IHMM takes the constraints of \(\beta_{1:2} = 0\) and \(x = 0\). These five models constitute the **IHMM group**.

### 4.1.4 GARCHt Group

Motivated by the importance of volatility dynamics (Durham 2003) in interest rates, this group extends the basic models with a GARCH-type conditional variance with Student-t innovations. The five models in the GARCHt group are summarized through

\[
r_t - r_{t-1} = \lambda + \beta r_{t-1} + e_t \quad e_t \sim St(0, \sigma^2_t r_{t-1}^x, \nu), \quad (24a)
\]

\[
\sigma_t = \alpha_0 + \alpha_1 e^2_{t-1} + \alpha_2 \sigma^2_{t-1}, \quad (24b)
\]

where \(St(0, \sigma^2_t r_{t-1}^x, \nu)\) denotes a Student-t density with mean 0 and variance \(\sigma^2_t r_{t-1}^x/\nu(\nu - 2)\) for \(\nu > 2\). VSK-GARCHt, CIR-GARCHt, and BSZ-GARCHt correspond to \(x = 0\), \(x = 1\), and \(x = 2\) of (24). GBM-GARCHt imposes \(\lambda = 0\) and \(x = 2\) while MER-GARCHt has \(\beta_1 = 0\) and \(x = 0\). These five models are denoted as the **GARCHt group**.

### 4.1.5 Priors

For individual models of Basic, MS2, IHMM and GARCHt groups, the prior for \(\lambda\) and \(\beta\) are independent standard normal, and \(\sigma \sim Gamma(5, 1)\). Additionally, each row of \(\Pi\) in the MS2 group follows a Dirichlet distribution with a vector of one as the concentration parameters. Independent \(N(0, 100)\) are used for \(\alpha_0, \alpha_1\) and \(\alpha_2\) with the restrictions \(\alpha_0 > 0, \alpha_1 \geq 0, \alpha_2 \geq 0\), while \(\nu \sim U(2, 100)\) in the GARCHt models.

In terms of pooling, the following hyper-prior is set on IMP

\[
\eta \sim Gamma(3, 1), \quad \alpha + \kappa \sim Gamma(2, 1), \quad \rho \sim Beta(3, 1).
\]

\(\alpha\) and \(\kappa\) are sampled together, and recall that \(\rho = \frac{\kappa}{\alpha + \kappa}\). Finally, \(\alpha_\omega \sim Gamma(4, 1)\). These prior settings for IMP are used in the three empirical examples unless otherwise noted. The priors for alternative forecast combination methods are referred to in Section 2.
4.2 Data

Data include monthly three-month treasury bill rates from the secondary market (T-Bill rate) and are downloaded from the FRED database maintained by the Research Department at the Federal Reserve Bank of St. Louis.\textsuperscript{9} The 882 observations span June 1947 to November 2020.

4.3 Posterior Analysis

Figure 1 presents the posterior average of weight allocations, \( E[\omega_s | y_{1:T}] \), in pooling the different groups of models, Basic, MS2, IHMM and GARCHt. There is strong evidence of changing weights across most model groups except the GARCHt group. Several models jointly play a significant role within each group.

For the Basic group, the CIR and MER are the top models based on weights across periods and this applies to the MS2 models as well. However, in the IHMM class, the MER-IHMM captures more weight. In the GARCHt group the MER-GARCHt has the dominant weight over time.

The IMP allows for as many weight vectors as needed. As previously discussed, if there is one weight vector it corresponds to the Geweke & Amisano (2011) optimal prediction pool. Comparing the posterior density intervals for weights of the CIR-IHMM individual model in the GA and IMP pooling approaches reveals notable differences between the two.\textsuperscript{10}

The posterior density of unique states or weight vectors indicates most of the mass is from 1 to 8 regimes. It is switches between these different weight vectors that result in a time-varying weight for a specific model, as seen in Figure 1. In addition, Figure 2 shows that the number of active states grows over time for all groups when pooled by the IMP, with the Basic group adding many more states along the sample whereas the the GARCHt specifications finishing with only three states at the end of the sample.

4.4 Out-of-Sample Forecasts

Forecasts are computed from \( \tau_1 = 31 \) to \( \tau_2 = T = 882 \) with \( \tau_0 = 20 \), for a total of 852 out-of-sample periods from December 1949 to November 2020. Forecasts are computed recursively over the out-of-sample period and as each new observation arrives each prediction pooling model is fully re-estimated to compute the next forecast. LPL, RMSFE and CRPS are reported in Table 1. Panel 1 shows forecast results of individual models from all four groups, while within-group pooling results of various model combination approaches are presented in Panel 2.

Among individual models the CIR specification achieves the largest LPL value in each of the groups. Sometimes model pooling approaches beat the CIR specification and other times not. Only the IMP approach consistently produces the largest LPL value in each group and overall. Among the different groups the largest LPL value is from the IMP over the GARCHt specifications with a value of 233.6. The breakdown of these LPL values into cumulative log-Bayes factors of the IMP against alternatives is reported in the Online Appendix.

In addition to pooling models within each group, we also consider pooling across different groups. The results on the IMP are presented in the final panel of Table 1. IMP-20 pools all 20 individual models (all four groups), while IMP-10 pools the 10 models in the groups of IHMM and GARCHt. Both of these larger pools increase the LPL making them strongly favored over the other pools. The IMP-10 has the largest value and a log-Bayes factor of 8.5 against the IMP of the GARCHt group. These results indicate that pooling over more models is desirable but which class is pooled over can influence results.

\textsuperscript{9} Federal Reserve Economic Data (FRED): https://fred.stlouisfed.org/series/TB3MS.

\textsuperscript{10} See the Online Appendix for this and additional results pertaining to this Section.
The IMP approach produces the smallest CRPS, but the improvement over other models and pooling methods is small. The smallest RMSFE in each group is produced by an individual model, but the margin is extremely small against IMP.

Some insight into the flexibility of the IMP approach is seen in Figure 3. The sample period 1999-2020 includes dramatic changes in both the levels and volatility dynamics of T-bill rates making it difficult for an individual model to capture. The IMP-10 makes significant gains as measured by Log-Bayes factors against the individual models particularly for 2008-2016 when short rates were forced to be artificially low. The nonparametric nature of the weight vector in the IMP allows for abrupt changes to a new value to accommodate this unique period.

Figure 4 plots the aggregate weights associated with the Basic, MS2, IHMM and GARCHt groups used in the IMP-20 model. For instance, the aggregate weight of the MS2 group is the sum of the weights assigned by IMP to VSK-MS2, CIR-MS2, BSZ-MS2, GBM-MS2 and MER-MS2 models. The IHMM and GARCHt groups are the major contributors to the pooling model but their relative contributions change over time.

4.5 Robustness

Forecast results for various subsamples are investigated for all forecast combination methods using the IHMM and GARCHt model groups. IMP generally dominates in these subsamples and when it does not it is very close to the best model in terms of LPL. As before there are only minor differences in RMSFE and CRPS results.

Next, the sensitivity of results to the value of $\tau_0$ with $\tau_1$ and $\tau_2$ fixed is considered. Recall that the conditioning data or training sample before the first forecast is $\tau_1 - 1 - (\tau_0 - 1) = \tau_1 - \tau_0$. Focusing on the IMP pooling over the IHMM and GARCHt groups with training sample sizes ranging from 10 to 110 shows little to no impact on the LPL values.

Finally, we assess the sensitivity of forecast results for the IMP-20 pooling specification to different prior settings for $\alpha$, $\eta$, $\rho$ and $\alpha_\omega$. Overall, the results are shown to be insensitive.

These and additional results are found in the Online Appendix.

4.6 Dynamic States of Waggoner & Zha (2012)

The key advantage of IMP versus the fixed states of Waggoner & Zha (2012) is that the number of states is estimated along with other model parameters rather than fixing it. In this section we report forecast results for Waggoner & Zha (2012) with two, three, four, and five states (denoted as WZ2, WZ3, WZ4, and WZ5) in addition to several versions that pick the number of states based on a performance metric. We consider three different performance metrics: LPL, RMSFE, and CRPS. For instance, for the forecast of $t + 1$, we select the best Waggoner & Zha (2012) model amongst the two to five states according to the best time $t$ performance of LPL (WZ-LPL), RMSFE (WZ-RMSFE), and CRPS (WZ-CRPS). Table 2 reports forecast performance for these models. The IMP shows robust performance against these alternatives in the T-Bill application. This is confirmed by Figure 2 that, instead of being fixed, the optimal state number inferred from the data is clearly time-varying.

5 Application to Realized Covariance Models

The next application is multivariate, in this case, forecasting 10-dimensional realized covariance (RCOV) matrices. This means there are 55 unique elements to forecast at each time period.
5.1 Models

Let $\Sigma_t$, $t = 1, \ldots, T$ denote a RCOV matrix of dimension $k$ and $\Sigma_{1:T-1} = \{\Sigma_1, \ldots, \Sigma_{t-1}\}$. Although the object of interest is a matrix, forecast pooling methods can be applied to any model that produces a predictive density of a quantity of interest. In this application, we will pool five popular RCOV models that have been introduced in the last decade.

The first model is from Jin & Maheu (2013), the additive component Wishart (W) specification

$$
\Sigma_t|\Sigma_{1:t-1} \sim \text{Wishart}_k(\nu, V_t/\nu),
$$

with three ($M = 3$) components. $\text{Wishart}_k(\nu, \frac{1}{\nu}V_t)$ denotes a Wishart distribution over positive definite matrices of dimension $k$ with $\nu$ degrees of freedom and scale matrix $\frac{1}{\nu}V_t$. $\odot$ denotes the element-by-element (Hadamard) product of two matrices. $B_0$ is a $k \times k$ symmetric positive-definite matrix, and $B_j = b_j b_j'$ where $b_j$s are $k \times 1$ vectors making each $B_j$ rank 1. $\Gamma_{t-1,\ell_j}$ is the $j^{th}$ (additive) component defined as the average of past $\Sigma_t$ over $\ell_j$ observations. The first component has one lag, $\ell_1 = 1$, while $\ell_2$ and $\ell_3$ are estimated which can lead to significantly better forecasts than assigned preset values. In our Bayesian inference, the priors on the elements of component inverse-Wishart model (IW) is from Jin & Maheu (2016) and follows Maheu (2013) for full details on estimation).

The next specification replaces the Wishart with an inverse-Wishart distribution. This additive component inverse-Wishart model (IW) is from Jin & Maheu (2016) and follows

$$
\Sigma_t|\Sigma_{1:t-1} \sim \text{invWishart}_k(\nu, (\nu - k - 1)V_t)
$$

where $\text{invWishart}_k(\nu, (\nu - k - 1)V_t)$ denotes an inverse-Wishart distribution over positive definite matrices of dimension $k$ with $\nu$ degrees of freedom and scale matrix $(\nu - k - 1)V_t$. The rest of the specification is the same as the previous model and the parameters are given the same priors.

The third model is the generalized conditional autoregressive Wishart model (GCAW) of Yu et al. (2017).

$$
\Sigma_t|\Sigma_{1:t-1} \sim \text{NCW}_k(\nu, V_t/\nu, \Lambda_t), \quad \Lambda_t = \sum_{i=1}^r M_t \Sigma_{t-i} M_t'
$$

where $\text{NCW}_k(\nu, V_t/\nu, \Lambda_t)$ is a noncentral Wishart distribution over positive definite matrices of dimension $k$, $\nu$ is the degree of freedom parameter, and $\nu > k - 1$. $V_t/\nu$ and $\Lambda_t$ are the scale matrix and the noncentrality matrix, respectively, both of which are symmetric positive definite. $C$ is a $k \times k$ lower triangular matrix and $A_i$, $B_i$, and $M_i$ are $k \times k$. Following the results in GCAW the best model is used with $p = 2, q = 2, r = 1$. For inference, independent $N(0, 100)$ are assigned as priors to all elements of $C, A_i, B_i, M_i$ except the $(1,1)^{th}$ element of each matrix, which uses positively
truncated $N(0, 100)$ for identification and $\nu \sim exp(100)I_{\nu > k-1}$. Posterior simulation for this model and the next is conducted with a Metropolis–Hastings step that jointly samples the full parameter vector using a random walk proposal.

The fourth model is the conditional autoregressive Wishart (CAW) model of Golosnoy et al. (2012) and specifies

$$\Sigma_t | \Sigma_{1:t-1} \sim \text{Wishart}_k(\nu, V_t/\nu)$$

$$V_t = CC' + \sum_{i=1}^p B_i V_{t-i} B_i' + \sum_{i=1}^q A_i \Sigma_{t-i} A_i',$$

(28a)

(28b)

where $\nu$ is the degree of freedom and $\nu > k-1$. $V_t/\nu$ is the scale matrix, which is symmetric positive definite. $C$ is a $k \times k$ lower triangular matrix and $A_i, B_i,$ and $M_i$ are $k \times k$. Since CAW is nested within GCAW, the same prior distributions are used for common parameters of the two models in the application of $p = q = 2$.

The last model (Dis), first introduced by Jin et al. (2019), adapts the matrix discounting model of West & Harrison (1997) for RCOV matrices.

$$\Sigma_t | \Sigma_{1:t-1} \sim \text{invWishart}_k(\beta n_t + k - 1, \beta n_t S_t),$$

(29a)

$$n_t = \beta n_{t-1} + 1, \quad S_t = \frac{1}{n_t}(\beta n_{t-1} S_{t-1} + \Sigma_{t-1}).$$

(29b)

RCOV is assumed to follow a fully specified inverse-Wishart distribution with $\beta = 0.95$, the discount factor reflecting the decay of information from $t-1$ to $t$.

5.2 Data

The 10-asset daily RCOV data is from Noureldin et al. (2011).\(^{11}\) The list of stocks include Alcoa (AA), American Express (AXP), Bank of America (BAC), Coca Cola (KO), Du Pont (DD), General Electric (GE), International Business Machines (IBM), JP Morgan (JPM), Microsoft (MSFT), and Exxon Mobil (XOM). The data ranges from 2001/02/01 to 2009/12/31 (2092 obs). Here $T=2092$, $\tau_0 = 1400$, $\tau_1 = 1550$ and $\tau_2 = 2092$. As a result, an out-of-sample of 693 periods of predictive likelihoods are produced from candidate models, and 543 (2007/11/06 to 2009/12/31) periods of predictive likelihoods are computed for pooling models.

5.3 Forecast Performance and Weight Dynamics

The computation of the LPL on each of the forecast combination methods and individual RCOV models is the same as in the interest rate application. The root mean squared forecast error is computed as

$$RMSFE_A = \sqrt{\frac{\sum_{t=\tau_1}^{\tau_2} \sum_{i=1}^k \sum_{j=1}^k (\Sigma_{t,ij} - E[\Sigma_{t,ij}|\Sigma_{1:t-1}, A])^2}{\tau_2 - \tau_1 + 1}}$$

where $\Sigma_{t,ij}$ is element $(i, j)$ of $\Sigma_t$, with associated predictive mean $E[\Sigma_{t,ij}|\Sigma_{1:t-1}, A]$ from model $A$, and $k = 10$. Due to computation cost, CRPS for RCOV model forecasts is omitted.

Table 3 presents the forecast performance of each RCOV model as well as all combination approaches.\(^{12}\) According to the LPL values, and among individual models, the IW model performs

\(^{11}\)http://realized.oxford-man.ox.ac.uk/data/download.

\(^{12}\)The BCRV and MW approaches were excluded as their present form was not designed for combining matrix forecasts without further assumptions.
the best and has a log-Bayes factor of more than 8600 over the second best model, the CAW model. The Dis model is very poor relative to the other alternatives. The last row reports results for the model combinations. Except for BMA all pooling methods improve upon the IW model with the IMP having the largest value. The IMP has a log-Bayes factor of 52 over WZ2, the second best pooling model. Turning to point forecasts, the CAW performs the best and consistently beats all forecast combination approaches.

Consistent with individual model performance the IW receives a large weight in the IMP as shown in Figure 5. Although the discounting model is uniformly the worst model based on forecast statistics it captures a substantial share of weight in many periods. The discounting model weight increases exactly when the weight on the IW model declines. This illustrates the importance of including even poor forecasting models in pooling approaches. The GCAW has a small weight over the out-of-sample and appears to be dominated by the CAW.

A full sample analysis reveals a posterior mean of active states in the IMP specification of 7.65, with a 0.95 density interval of (5, 11).

In summary, the IMP approach produces superior density forecasts but is not as competitive for point forecasts as several individual models.

6 Application to Fama-French and Q-factor Models

The Fama–French 5-factor model (FF) of Fama & French (2015) and the Q-factor model of Hou et al. (2015) are two prominent models used to explain the pricing of risky assets. In this application, the two asset pricing models are used to forecast returns of the 10 industry portfolios from the Kenneth R. French data library, which includes the FF factors. The Q-factor data are obtained from the Hou-Xue-Zhang data library. The data are monthly value-weighted returns, from January 1967 to December 2019 (636 observations). In the following we set \( \tau_0 = 10, \tau_1 = 61 \) and \( \tau_2 = T = 636 \), leaving 576 out-of-sample observations. Although homoskedastic and heteroskedastic versions are considered, the main feature distinguishing one model from the another is the factors included in the conditional mean.

6.1 Models and Data

The following is the 5-factor (FF) model of Fama & French (2015), which postulates that portfolio excess returns follow

\[
 r_t - r_{ft} = \alpha + \beta_1 f_1 t + \beta_2 f_2 t + \beta_3 f_3 t + \beta_4 f_4 t + \beta_5 f_5 t + e_t, \quad e_t \sim N(0, \sigma^2),
\]

where \( r_{ft} \) denotes the risk free rate. The Q-factor (Qf) model of Hou et al. (2015) and Hou et al. (2021) is

\[
 r_t - r_{ft} = \alpha + \beta_1 q_1 t + \beta_2 q_2 t + \beta_3 q_3 t + \beta_4 q_4 t + \beta_5 q_5 t + e_t, \quad e_t \sim N(0, \sigma^2).
\]

The factors \( f_1 t, \ldots, f_5 t \) are excess market returns, return difference between diversified small and big stocks, robust and weak profitability firms, low and high investment firms, and high and low Book to Market (B/M) firms, respectively. \( q_1 t, \ldots, q_5 t \) denote market excess returns, size factor returns, investment factor returns, equity factor returns, and expected growth returns.

---

13 [https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html].
14 [http://global-q.org/factors.html].
The original models of Fama & French (2015) and Hou et al. (2015) are homoskedastic models that did not consider volatility dynamics. To improve density forecasts we introduce heteroskedastic versions of (30) and (31) by replacing the $\sigma^2$ with $\sigma^2_t$ in the following GARCH model

$$\sigma^2_t = \omega_0 + \omega_1 \varepsilon^2_{t-1} + \omega_2 \sigma^2_{t-1}. \quad (32)$$

The GARCH version of the models are denoted as FF-GH and Qf-GH.

The priors of $\alpha; \beta_1; \ldots; \beta_5$ are set to follow independent standard normal distributions and we assume $\sigma^{-2} \sim \text{Gamma}(3, 1)$. Priors for $\omega_0, \omega_1$ and $\omega_2$ are independent $N(0, 100)$, with $\omega_0 > 0$, $\omega_1 \geq 0$, $\omega_2 \geq 0$ and $\omega_1 + \omega_2 < 1$.

6.2 Results

Table 4 presents the out-of-sample forecast performance for the four individual models as well as combination approaches. In two cases (Manuf and HiTec), the IMP method has the largest LPL. In four other cases, the IMP produces essentially the same value as the top performing model (Durbl, Telcm, Healh, and Utils). The other model combination methods are more competitive in this application with MW performing with the best consistency overall.

In the IMP, GARCH specifications generally receive the largest weight, but the homoskedastic versions also receive significant weights. With the exception of the Durable portfolio, the time-varying nature of the weights is clear (see the Online Appendix for more details).

7 Extensions

This section demonstrates how the HDP prior can be used to allow dynamic weights in other model combination approaches.

7.1 Infinite Markov Predictive Synthesis

Bayesian predictive synthesis of McAlinn & West (2019) can be extended to a Bayesian nonparametric setting. This combination approach is labeled as infinite Markov predictive synthesis (IMPS). The key difference is that here the weight vector has countably infinite support and is governed via an infinite Markov chain, whereas the weights in McAlinn & West (2019) follow an autoregressive process with support on $\mathbb{R}$. The IMPS specification is

$$\Gamma \sim \text{Stick}(\eta), \quad \Pi_i \overset{iid}{\sim} DP\left(\alpha + \kappa, \frac{\alpha \Gamma + \kappa \delta_i}{\alpha + \kappa}\right), \quad i = 1, 2, \ldots, \quad (33a)$$

$$\alpha(y_t|I_{t-1}, s_t) = \frac{1}{\sqrt{2\pi \sigma^2_{s_t}}} \exp\left(-\frac{(y_t - \omega_{s_t,0} - \sum_{j=1}^L \omega_{s_t,j} \hat{y}_{jt})^2}{2\sigma^2_{s_t}}\right), \quad s_t|s_{t-1} \sim \Pi_{s_{t-1}} \quad (33b)$$

$$\omega_i \sim MN(a, A), \quad \sigma^{-2}_i \sim \text{Gamma}(\chi, \nu), \quad i = 1, 2, \ldots. \quad (33c)$$

As previously, a sticky version of the HDP is used to allow for estimation of state persistence. $\Gamma$ and $\Pi$ construct the Markov transition matrix with infinite dimension. The precision parameters, $\eta$, $\alpha$ and $\kappa$ have the same interpretation as in the IMP model. $\alpha(y_t|I_{t-1}, s_t)$ is the new synthesis function. It is a function of the latent state $s_t$ and allows for heteroskedasticity through $\sigma^2_{s_t}$. $\omega_{s_t,0}$ can potentially play a role in capturing model errors. For instance, a positive number implies that the aggregation of the pool of models is biased downward. $MN(a, A)$ denotes a multivariate normal distribution with mean vector $a$ and covariance $A$. 

18
A hierarchical prior can be used in this model,

\[ a \sim MN(h_0, H_0), \quad A^{-1} \sim Wishart(a_0, A_0), \quad \chi \sim exp(c_0), \quad \nu \sim Gamma(c_1, d_1), \]  \hspace{1cm} (34)

and hyper-priors set for \( \eta, \alpha + \kappa \) and \( \rho = \frac{\alpha}{\alpha + \kappa} \) as before.

Posterior sampling for the IMPS model is essentially the same as the IMP model. The key difference is sampling \( \hat{y} \) from each individual model. A Metropolis–Hasting sampler for \( \hat{y} \) can be designed to simulate from the target density

\[ p(\hat{y}_t|y_{1:t}, M_{1:L}, s_t) \propto \alpha(y_t|s_t, \hat{y}) \prod_{j=1}^L p(\hat{y}_j|y_{1:t-1}, M_j). \]  \hspace{1cm} (35)

This coupled with sampling steps from IMP complete the posterior simulation. The predictive density for \( y_{t+1} \) is now an infinite Markov mixture of (33b) which integrates over all possible future states, latent predictors and model parameters.

### 7.2 Infinite Markov Calibration

This section demonstrates that the HDP prior can also be used to extend Bassetti et al. (2018) to a dynamic setting.

To combine a set of univariate predictive cumulative distribution functions (cdf) from different forecasting models, Bassetti et al. (2018) extend Ranjan & Gneiting (2010)’s Beta-transformed linear pool and propose to use Beta mixtures to achieve distribution calibration and model combination at the same time. This approach is denoted as calibration and combination.

The Bayesian Infinite Beta Mixture Model (or \( BM_\infty \)) by Bassetti et al. (2018) utilizes a Dirichlet process prior for the mixing distribution and the component parameters and has the following form

\[ F(y_t|I_{t-1}, \Theta) = \sum_{i=1}^{\infty} w_i B_{\mu_i, \nu_i}^* \left( \sum_{l=1}^L \omega_{i,l} F(y_t|I_{t-1}, M_l) \right), \]  \hspace{1cm} (36a)

\[ W \sim Stick(\psi), \]  \hspace{1cm} (36b)

\[ \mu_i \overset{i.d.}{\sim} Beta(\xi_\mu, \xi_\mu), \quad \nu_i \overset{i.d.}{\sim} Gamma(\nu_{\mu}, \nu_{\mu}), \quad \omega_i \overset{i.d.}{\sim} Dir(\xi_\omega, \ldots, \xi_\omega), \quad i = 1, 2, \ldots, \]  \hspace{1cm} (36c)

where \( F(.|I_{t-1}, M_l) \) is the predictive cdf of model \( M_l \) and \( F(.|I_{t-1}, \Theta) \) is the aggregated predictive cdf. The parameter set \( \Theta \) includes mixture weights \( W \equiv \{w_i\}_{i=1}^{\infty} \) and component parameters \( \{\mu_i, \nu_i, \omega_i\}_{i=1}^{\infty} \), where \( \omega_i = (\omega_{i,1}, \ldots, \omega_{i,L}) \) are the combination weights for the \( i \)-th component. \( B_{\mu, \nu}^*(\cdot) \) denotes the cdf of a beta distribution with mean \( \mu \) and precision \( \nu \) and with its pdf, \( b_{\mu, \nu}^*(x) \propto x^{\mu-1}(1-x)^{(1-\mu)\nu-1} \) supported on \([0,1]\). If \( w_1 = 1 \) and \( w_i = 0 \) for \( i > 1 \), (36) degenerates to the original beta-transformed linear pool of Ranjan & Gneiting (2010).

The associated density of \( F(.|I_{t-1}, \Theta) \) is obtained by taking its derivative. Under \( BM_\infty \), the predictive density of \( y_t \) is

\[ p(y_t|I_{t-1}, \Theta) = \sum_{i=1}^{\infty} \left[ w_i b_{\mu_i, \nu_i}^* \left( \sum_{l=1}^L \omega_{i,l} F(y_t|I_{t-1}, M_l) \right) \right] \left( \sum_{l=1}^L \omega_{i,l} p(y_t|I_{t-1}, M_l) \right). \]  \hspace{1cm} (37)

Bayesian inference can be conducted straightforwardly and Bassetti et al. (2018) provide details for prior specification and sampling steps as well as predictive distribution approximation. It is worth noting that extensions to higher dimension are difficult due to the need to evaluate a multivariate cdf.

---

\(^{15}\)Applications of the calibration method to combine densities have been confined to the univariate settings. Extensions to higher dimension are difficult due to the need to evaluate a multivariate cdf.
noting, however, that although the predictive density in (37) has a closed form, the predictive mean in general does not. This makes estimating the predictive mean a non-trivial task, since it requires directly sampling from (37), which usually does not come from any “simulation friendly” distribution family.

The BM$_\infty$ can be nonparametrically extended to accommodate time-varying calibration and combination by replacing the DP prior with the HDP prior of Teh et al. (2006). This new combination method, called infinite Markov calibration and denoted as BM$_{ihmm}$, allows distribution calibration and model combination to be state-dependent in a Markovian structure where the number of states is unbounded. The specification is as follows

\begin{equation}
\Gamma \sim \text{Stick}(\eta), \quad \Pi_i \sim \text{DP} \left(\alpha + \kappa, \frac{\alpha \Gamma + \kappa \delta_i}{\alpha + \kappa} \right), \quad i = 1, 2, \ldots,
\end{equation}

\begin{equation}
F(y_t|I_{t-1}, \Theta, s_t) = B_{\mu_{s_t}, \nu_{s_t}}^* \left( \sum_{l=1}^L \omega_{s_t,l} F(y_t|I_{t-1}, M_l) \right), \quad s_t|s_{t-1} \sim \Pi_{s_{t-1}},
\end{equation}

\begin{equation}
\mu_i \sim \text{Beta}(\xi, \xi), \quad \nu_i \sim \text{Gamma}(\nu, \nu), \quad \omega_i \sim \text{Dir}(\xi, \ldots, \xi), \quad i = 1, 2, \ldots.
\end{equation}

Under BM$_{ihmm}$, the calibration function and combination weights both depend on state $s_t$. $s_t$ follows an infinite Markov chain with the transition matrix $\Pi$ drawn from an HDP prior. After marginalizing over $s_t$, the aggregated predictive cdf of $y_t$ given $s_{t-1}$ is

\begin{equation}
F(y_t|I_{t-1}, \Theta, s_{t-1}) = \sum_{s_t=1}^\infty \pi_{s_{t-1}, s_t} B_{\mu_{s_t}, \nu_{s_t}} \left( \sum_{l=1}^L \omega_{s_t,l} F(y_t|I_{t-1}, M_l) \right),
\end{equation}

with density

\begin{equation}
p(y_t|I_{t-1}, \Theta, s_{t-1}) = \sum_{s_t=1}^\infty \pi_{s_{t-1}, s_t} B_{\mu_{s_t}, \nu_{s_t}} \left( \sum_{l=1}^L \omega_{s_t,l} F(y_t|I_{t-1}, M_l) \right) \times \left( \sum_{l=1}^L \omega_{s_t,l} p(y_t|I_{t-1}, M_l) \right).
\end{equation}

For posterior sampling, Van Gael et al. (2008)’s beam sampler is applied, coupled with FFBS to draw the state sequence $s_1:T$. For the parameters $\Gamma, \Pi, \eta, \alpha, \text{and} \kappa$, sampling can be performed in a similar fashion as in the IMP model. Lastly, the same steps in BM$_\infty$ are used to sample $\mu_i, \nu_i$, and $\omega_i$. Given a sufficient amount of MCMC draws from the posterior distribution, Bayesian inference proceeds in the usual way, such as estimating the predictive cdf and pdf with parameter uncertainty integrated out. Sampling steps are detailed in the Online Appendix.

7.3 Application to U.S. Short-term Interest Rate Models

Table 5 presents forecast results for the U.S. short-term interest rate application using identical data to that of Table 1. The BM$_\infty$, BM$_{ihmm}$, and IMP are included. As mentioned, simulation from the predictive density is not a trivial task and therefore RMSFE is omitted.

Although the IMP performs better than the BM$_\infty$ for pooling over IHMM and GARCHt individual models, the BM$_{ihmm}$ dominates all the other alternatives. This indicates that the dynamic beta weights governed by the infinite Markov chain provide very large gains compared to the static weights from the DP prior used in Bassetti et al. (2018). For instance, the best LPL value is 260.3, whereas the largest value not from a BM$_{ihmm}$ specification is 233.6 for the IMP pooling of GARCHt. The IMP-10, the best model from Table 1, is also soundly beaten. CRSPs are uniformly better for the BM$_{ihmm}$.
It is also interesting to note that the $BM_{ihmm}$ specifications are all competitive no matter what individual models are pooled over. For example, the $BM_{ihmm}$ results using Basic and MS2 groups of models is impressive. For the other approaches we only see significantly better forecast results when better models are pooled. The $BM_{ihmm}$ seems to overcome this somewhat, and it appears due to the addition of the HDP prior for the beta weights. This makes the model very promising, in that pooling over mediocre forecast models can still result in a top performing forecasting model.

8 Conclusion

This paper introduces a novel approach to forecast pooling methods based on a nonparametric prior for the weight vector combining predictive densities. An HDP prior allows the weight vector on a set of models to follow an infinite hidden Markov chain. This generalizes dynamic prediction pools to the nonparametric setting. We discuss efficient posterior simulation based on MCMC methods. Detailed applications to short-term interest rates, RCOV matrices, and asset pricing models show the nonparametric pool forecasts well.

The superiority of our pooling approach is based on its allowance for pooling weights to change over time in a structured, yet flexible enough, manner that can accommodate periods with persistent weights in addition to those featuring abrupt weight changes. The nonparametric nature of the prior makes the number of the states (hence weight vectors) unbounded and new weight vectors can be added if necessary when moving out-of-sample and pruning old states when they are no longer needed. Furthermore, parsimony is achieved with high state persistence that allows the weight vector to remain constant, while introducing many new states allows for smooth rapid changes in the weight vector when needed.
References


Table 1: T-Bill Rate Application: Forecast Performance of Individual Models and Pooling Methods

<table>
<thead>
<tr>
<th>Panel 1: Individual Interest Rate Models</th>
<th>VSK</th>
<th>CIR</th>
<th>GBM</th>
<th>MER</th>
<th>BSZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Group</td>
<td>-441.2</td>
<td>-38.4</td>
<td>-955.4</td>
<td>-439.0</td>
<td>-1049.0</td>
</tr>
<tr>
<td></td>
<td>(0.4027)</td>
<td>(0.4012)</td>
<td>(0.4061)</td>
<td>(0.3991)</td>
<td>(0.4035)</td>
</tr>
<tr>
<td>MS2 Group</td>
<td>-241.6</td>
<td>17.8</td>
<td>-163.3</td>
<td>-291.6</td>
<td>-168.9</td>
</tr>
<tr>
<td></td>
<td>(0.3962)</td>
<td>(0.4012)</td>
<td>(0.4003)</td>
<td>(0.4009)</td>
<td>(0.4012)</td>
</tr>
<tr>
<td>IHMM Group</td>
<td>55.8</td>
<td>114.2</td>
<td>-46.2</td>
<td>71.7</td>
<td>-54.9</td>
</tr>
<tr>
<td></td>
<td>(0.3954)</td>
<td>(0.4017)</td>
<td>(0.4026)</td>
<td>(0.4023)</td>
<td>(0.4001)</td>
</tr>
<tr>
<td>GARCHt Group</td>
<td>173.1</td>
<td>199.9</td>
<td>66.9</td>
<td>195.0</td>
<td>-76.5</td>
</tr>
<tr>
<td></td>
<td>(0.4093)</td>
<td>(0.4035)</td>
<td>(0.4090)</td>
<td>(0.3994)</td>
<td>(0.3997)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel 2: Forecast Pooling Methods</th>
<th>BMA</th>
<th>GA</th>
<th>DHS</th>
<th>WZ2</th>
<th>BCRV</th>
<th>MW</th>
<th>IMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Group</td>
<td>-40.1</td>
<td>4.7</td>
<td>-14.2</td>
<td>31.2</td>
<td>-32.6</td>
<td>-23.5</td>
<td>68.7</td>
</tr>
<tr>
<td></td>
<td>(0.4003)</td>
<td>(0.3998)</td>
<td>(0.4003)</td>
<td>(0.3997)</td>
<td>(0.4009)</td>
<td>(0.4053)</td>
<td>(0.3997)</td>
</tr>
<tr>
<td>MS2 Group</td>
<td>16.1</td>
<td>13.1</td>
<td>2.5</td>
<td>30.8</td>
<td>-30.5</td>
<td>-6.3</td>
<td>69.9</td>
</tr>
<tr>
<td></td>
<td>(0.3961)</td>
<td>(0.3982)</td>
<td>(0.3991)</td>
<td>(0.3982)</td>
<td>(0.3986)</td>
<td>(0.4049)</td>
<td>(0.3983)</td>
</tr>
<tr>
<td>IHMM Group</td>
<td>112.5</td>
<td>136.2</td>
<td>127.8</td>
<td>165.0</td>
<td>117.0</td>
<td>49.1</td>
<td>188.0</td>
</tr>
<tr>
<td></td>
<td>(0.4022)</td>
<td>(0.3991)</td>
<td>(0.3982)</td>
<td>(0.3976)</td>
<td>(0.3990)</td>
<td>(0.4042)</td>
<td>(0.3976)</td>
</tr>
<tr>
<td>GARCHt Group</td>
<td>198.5</td>
<td>232.9</td>
<td>224.0</td>
<td>231.3</td>
<td>208.0</td>
<td>149.3</td>
<td>233.6</td>
</tr>
<tr>
<td></td>
<td>(0.3995)</td>
<td>(0.4013)</td>
<td>(0.4030)</td>
<td>(0.4024)</td>
<td>(0.4035)</td>
<td>(0.4069)</td>
<td>(0.4013)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel 3: IMP Forecast Pooling using Multiple Groups</th>
<th>IMP-10</th>
<th>IMP-20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>242.1</td>
<td>237.2</td>
</tr>
<tr>
<td></td>
<td>(0.3983)</td>
<td>(0.3978)</td>
</tr>
<tr>
<td></td>
<td>{0.1522}</td>
<td>{0.1522}</td>
</tr>
</tbody>
</table>
### Table 2: T-Bill Rate Application: Forecast Performance of Pooling Methods by Waggoner & Zha (2012) with 3 to 5 States and Dynamic Versions

<table>
<thead>
<tr>
<th></th>
<th>Basic Group</th>
<th></th>
<th>MS2 Group</th>
<th></th>
<th>IHMM Group</th>
<th></th>
<th>GARCHt Group</th>
<th></th>
<th>IMP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WZ2</td>
<td>31.2 (0.3997)</td>
<td>30.8 (0.3982)</td>
<td>165.0 (0.3976)</td>
<td>231.3 (0.4024)</td>
<td>68.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WZ3</td>
<td>27.1 (0.3997)</td>
<td>26.1 (0.3977)</td>
<td>152.4 (0.3978)</td>
<td>230.3 (0.4026)</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WZ4</td>
<td>15.2 (0.4002)</td>
<td>20.1 (0.3980)</td>
<td>153.1 (0.3977)</td>
<td>229.4 (0.4026)</td>
<td>188.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WZ5</td>
<td>8.1 (0.4002)</td>
<td>7.9 (0.3980)</td>
<td>155.5 (0.3977)</td>
<td>229.3 (0.4028)</td>
<td>169.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WZ-LPL</td>
<td>32.6 (0.3998)</td>
<td>36.6 (0.3976)</td>
<td>169.5 (0.3977)</td>
<td>231.0 (0.4028)</td>
<td>161.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WZ-RMSFE</td>
<td>23.9 (0.3987)</td>
<td>21.7 (0.3976)</td>
<td>161.4 (0.3977)</td>
<td>230.8 (0.4023)</td>
<td>160.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WZ-CRPS</td>
<td>33.5 (0.3998)</td>
<td>36.5 (0.3976)</td>
<td>169.3 (0.3976)</td>
<td>230.8 (0.4023)</td>
<td>160.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP</td>
<td>68.7 (0.3998)</td>
<td>69.9 (0.3983)</td>
<td>188.0 (0.3976)</td>
<td>233.6 (0.4013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See the note to Table 1 for table entries. WZ2 to WZ5 correspond to Waggoner & Zha (2012) with fixed number of states from 2 to 5. WZ-LPL, WZ-RMSFE and WZ-CRPS correspond to the dynamic versions that select the number of states based on the best past performance using the metrics of LPL, RMSFE and CRPS, respectively.

### Table 3: RCOV Application: Forecast Performance of Individual Models and Pooling Methods

<table>
<thead>
<tr>
<th></th>
<th>Panel 1: Individual RCOV Models</th>
<th></th>
<th>Panel 2: Forecast Pooling Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W</td>
<td>IW</td>
<td>CAW</td>
</tr>
<tr>
<td></td>
<td>-45941 (74.8323)</td>
<td>-36962 (76.4418)</td>
<td>-45600 (72.4369)</td>
</tr>
</tbody>
</table>

This table reports LPL and RMSFE in () for individual models and model pooling methods on 543 (November 6, 2007 to December 31, 2009) out-of-sample observations. Each pooling method in the bottom panel combines all the models in the top panel.
<table>
<thead>
<tr>
<th>NoDur</th>
<th>Individual Asset Pricing Models</th>
<th>Forecast Pooling Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FF</td>
<td>Qf</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table reports LPL, RMSFE in (), and CRPS in {} for the out-of-sample period from January 1972 to December 2019 (576 observations). The 2nd to 5th column represent individual models (FF, Qf, FF-GH and Qf-GH). The 6th to 11th column show results of various pooling methods combing FF, Qf, FF-GH and Qf-GH.
Table 5: T-Bill Rate Application: Forecast Comparison of Bayesian Nonparametric Pooling Methods

<table>
<thead>
<tr>
<th>Group</th>
<th>IMP</th>
<th>BM∞</th>
<th>BM_{ihmm}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>68.7 (0.1625)</td>
<td>100.4 (0.1597)</td>
<td><strong>243.2 (0.1473)</strong></td>
</tr>
<tr>
<td>MS2</td>
<td>69.9 (0.1621)</td>
<td>125.1 (0.1584)</td>
<td><strong>245.5 (0.1486)</strong></td>
</tr>
<tr>
<td>IHMM</td>
<td>188.0 (0.1534)</td>
<td>176.5 (0.1527)</td>
<td><strong>260.3 (0.1459)</strong></td>
</tr>
<tr>
<td>GARCHt</td>
<td>233.6 (0.1537)</td>
<td>220.7 (0.1544)</td>
<td><strong>256.3 (0.1497)</strong></td>
</tr>
</tbody>
</table>

LPL is the top entry followed by CRPS in {}.
This figure shows the posterior average of $\omega_{1:T}$ from IMP for each group. Each panel represents a group and each color represents the associated weights assigned to the indicated model.
Figure 2: T-Bill Rate Application: Posterior Average of Active States Out-of-sample

This figure displays the posterior average of the number of active states for pooling various groups by IMP method from recursive estimation at each date in the out-of-sample period.

Figure 3: T-Bill Rate Application: Cumulative Log-Bayes Factors for IMP-10 Against Alternatives
This figure shows the posterior average aggregate of $\omega_{1:T}$ over each group (weight summation of five models) entering the IMP-20 model.

This figure displays the posterior average of $\omega_{1:T}$ from pooling the five RCOV models.