创业与管理学院

School of Entrepreneurship and Management

SHANGHAITECH SEM WORKING PAPER SERIES

No. 2020-008

Bayesian Nonparametric Forecast Pooling

Xin Jin

Shanghai University of Finance and Economics

John M. Maheu

McMaster University

Qiao Yang

ShanghaiTech University

July 2020

https://ssrn.com/abstract=3649934

School of Entrepreneurship and Management
ShanghaiTech University
http://sem.shanghaitech.edu.cn

Bayesian Nonparametric Forecast Pooling[*]

Xin Jin[†] John M. Maheu[‡] Qiao Yang[§]

July 2020

Abstract

This paper introduces a new approach to forecast pooling methods based on a nonparametric prior for the 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. We discuss efficient posterior simulation based on MCMC methods. Detailed applications to short-term interest rates, realized covariance matrices and asset pricing models show the nonparametric pool forecasts well.

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

[*] This paper is previously circulated as *”A Bayesian Nonparametric Approach on Modeling Combination for Short-term Rates”*.

[†] School of Economics, Shanghai University of Finance and Economics. Email:jin.xin@mail.shufe.edu.cn

[‡] DeGroote School of Business, McMaster University. Email:maheujm@mcmaster.ca

[§] School of Entrepreneurship and Management, ShanghaiTech University. Email:yangqiao@shanghaitech.edu.cn

1 Introduction

This paper introduces several nonparametric extensions to prediction pools. We use 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 shows the forecasts for the nonparametric pools to perform well.

Since Hall & Mitchell (2007) and Geweke & Amisano (2011) there has been considerable interest in model pooling methods. This is a density combination approach that combines several predictive densities to form a forecast. A significant feature of this approach is that unlike Bayesian model averaging it recognizes that the true model may not be in the model set. This density combination approach has had many applications and important extensions that allow weights on individual densities to change over time include Billio et al. (2013), Waggoner & Zha (2012) and Del Negro et al. (2016).

This paper contributes to the forecast pooling literature by proposing a new approach. We show how a nonparametric prior can be imposed on the weights vector to flexibly combine models. A Dirichlet process prior can be used to allow for countably infinite support of the weights vectors combining models. Although this generalizes 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 weight vector following an infinite hidden Markov structure. As such this is very flexible and due to the infinite support for the weight vector we call this class of forecast pooling models 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 sidesteps the problem of estimating the dimension of the Markov chain. States are allowed to be persistent and a prior on persistence is set through the sticky infinite hidden Markov model of Fox et al. (2011). IMP can be thought of as 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 take the individual models' predictive likelihoods as *data* and uses posterior simulation methods to estimate and combine these individual models assuming the model probability weights are governed by an infinite Markov chain. We design a new posterior simulation that jointly samples the latent state and the model indicator by extending the beam sampler approach 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 even though 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 model can be computed. For instance, simulated values or quantiles can be estimated by standard posterior simulation methods that simulate both from the predictive density of individual models and the predictive density of the IMP specification. These forecasts take into account all past active states as well as the possibility of new future states.

Infinite Markov pooling is applied to three empirical applications to assess its strengths and weaknesses. Overall IMP can result in large improvements in the accuracy of density forecasts as measured by log-predictive likelihoods. We compare to several state-of-the-art model combination approaches and show that IMP is a dominate 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 applications to different model classes for interest rates the IMP provides the best density forecasts. Using better individual forecasting models improves all combination methods and we generally recommend including all reasonable candidates as pooling of 10 and 20 models give the best performing IMP. These results are robust to subsamples and various prior settings. There is considerable dynamic changes in weights in contrast to the fixed weights of Geweke & Amisano (2011). On the other hand there is posterior support for 3–9 different probability weight vectors making a fixed setting of 2 or another finite number as in Waggoner & Zha (2012) problematic.

The second application is multivariate and forecasts realized covariance matrices. This application shows that IMP is useful in higher dimensional settings in this case positive definite matrices of dimension 10. Most of the model combination approaches we consider improve density forecasts of realized covariance matrices compared to the individual models. The IMP approach has a log-Bayes factor of 48 in its favour 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 is to predicting monthly returns for ten industry portfolios. Individual models include popular risk premium specifications with and without GARCH type heteroskedasticity. IMP delivers robust density forecast improvements compared to other combination methods. Weights from the IMP are more stable compared to the other applications but do display abrupt changes from time to time.

The paper is organized as follows. The next section reviews existing model combination approaches which we include 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 infinite Markov pooling the main new approach we focus on. Posterior sampling and computation of forecasts are discussed. Section 4 provides a detailed analysis of the methods to short-term interest rate models. Section 5 applies IMP to realized covariance matrix forecasts and Section 6 is an application to returns from 10 industry portfolios. Section 7 concludes while an Appendix contains detailed steps of posterior simulation.

2 Existing Forecast Combination Approaches

In the following we denote the $p \times 1$ vector of interest 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 we suppress to minimize notation. Next we consider some benchmark density combination approaches followed by our approach.

2.1 Bayesian Model Averaging (BMA)

Bayesian model averaging (BMA)[1] assumes a *complete* model space, in that one of the set $\{M_1, \ldots, M_L\}$ of models is correct. In this setting the predictive density is formed as

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

where $p(M_l|y_{1:t-1}) \propto p(y_{t-1}|y_{1:t-2}, M_l)p(M_l)$. $p(y_t|y_{1:t-1}, M_l) = \int p(y_t|\theta_l, M_j)p(\theta_l|y_{1:t-1}, M_l)d\theta_l$ is the predictive likelihood of model M_l and θ_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 (DGP) the posterior probability for the model that minimizes the Kullback-Leibler distance will tend to 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].

[1] Wright (2008)

[2] Hoogerheide et al. (2010) suggests an extension, where the weights are allowed to be time-varying by imposing a random walk on weights dynamics. Billio et al. (2012) uses BMA and apply it to turning point predictions.

2.2 Static Pooling: Optimal Pooling

Geweke & Amisano (2011) and Hall & Mitchell (2007) proposes 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

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

Although static pooling can result in significant improvements in density forecasts, 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 while 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.

2.3 Dynamic Pooling: Autoregressive Weights

Del Negro et al. (2016) introduces density pooling with time-varying weights. Now weights are $\omega_t = (\omega_{t,1}, \ldots, \omega_{t,L}) = g(X_t)$ where $g(\cdot)$ is 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 $\sum_{l=1}^{L} \omega_{t,l} = 1$. A version we consider in this paper is

$$X_{t,i} = (1-\rho)\mu + \rho X_{t-1,i} + \sqrt{1-\rho^2}\sigma\epsilon_{t,i},\ \epsilon_{t,i} \overset{iid}{\sim} N(0,1),\ x_{0,i} \sim N(\mu, \sigma^2),\ i = 1, \ldots, L, \tag{3a}$$

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

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 ρ closer to 1 and/or smaller values of σ^2 translate into more persistent weights ω_t through time. Since this model is a nonlinear state-space model we follow Del Negro et al. (2016) and sample the weights by a bootstrap particle filter.

We take (3) as our benchmark and label this model DHS. Updating ρ in the particle filter is challenging due to expensive computational cost. The magnitude of ρ plays an important role in determining ω_t. A ρ too large restricts the dynamics of ω_t and make any structural change in the weights unlikely. For these reasons we set $\rho = 0.8$, $\mu = 0$ and $\sigma = 1.67$ in estimation[3].

2.4 Dynamic Pooling: Markov Weights:

Instead of the previous weight dynamics, Waggoner & Zha (2012) assume the weights follow a two-state Markov chain. We label this approach WZ and it takes the form,[4]

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

where $\Pi_{s_{t-1}}$ denotes the row of the transition matrix Π and governs the probability of the next state after s_{t-1}, while $\omega_{s_t} = (\omega_{s_t,1}, \ldots, \omega_{s_t,L})$. This is a Markov switching model of predictive densities and allows the model weights ω_{s_t} to change according to a two state Markov chain. We assume the prior of $\omega_{s_t} \sim Dir(1,1)$. This model can be easily sampled through 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.[5] 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[6] to the particle generating process. We consider the following version,

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

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

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

[3] $\sigma = 1.67$ results in $\sigma\sqrt{1-\rho^2} = 1$

[4] 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.

[5] Aastveit et al. (2018) uses the same method for Macroeconomic nowcasting.

[6] It is an exponentially weighted learning strategy into the weight dynamics for estimating the density of X_t

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

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

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 with 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 can have difficulty in capturing fat tails and heteroskedasticity over time. The ω_t and $g(X_t)$ are defined the same way as DHS in the previous subsection. 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.

3 Bayesian Nonparametric Prediction Pooling

To begin consider a simple Bayesian pooling approach for L models

$$p(y_t|y_{1:t-1}, \omega) = \sum_{l=1}^{L} \omega_l p(y_t|y_{1:t-1}, M_l), \;\; \omega \sim H, \tag{7}$$

where H is the prior distribution for the weights vector. This is a Bayesian analogue of Geweke & Amisano (2011) with one vector parameter ω to be estimated.[7] The predictive density is formed in the usual way by integrating out parameter uncertainty from the posterior of ω.

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

$$G|\alpha \sim DP(\alpha, G_0), \tag{8a}$$

$$\omega_t \overset{iid}{\sim} G, \tag{8b}$$

$$p(y_t|y_{1:t-1}, \omega_t) = \sum_{l=1}^{L} \omega_{t,l} p(y_t|y_{1:t-1}, M_l). \tag{8c}$$

$DP(\alpha, G_0)$ 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 each time period the weights

[7] $\omega = (\omega_1, \ldots, \omega_L)$

vector ω_t is drawn[8]. This model is a Dirichlet process mixture (DPM) model (Antoniak 1974). In large samples parameter uncertainty from the weights vector for these last two models will be small and 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 on replaces the DP prior 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 infinite state Markov chain. Although this framework continues to combine L models, the possible ways of combining models remain 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 denoted as IMP. The infinite here refers to the unbounded potential number of weight vectors used to combine the finite L models.[9] The model is

$$\Gamma \sim Stick(\eta), \tag{10a}$$

$$\Pi_i \overset{iid}{\sim} DP\left(\alpha+\kappa, \frac{\alpha\Gamma+\kappa\delta_i}{\alpha+\kappa}\right), \quad i=1,2,..., \tag{10b}$$

$$s_t|s_{t-1} \sim \Pi_{s_{t-1}}, \tag{10c}$$

$$f(y_t|I_{t-1},s_t) = \sum_{l=1}^{L} \omega_{s_t,l}\, p(y_t|I_{t-1},M_l) \tag{10d}$$

$$\omega_i \sim Dir\left(\frac{\alpha_\omega}{L},\ldots,\frac{\alpha_\omega}{L}\right), \quad i=1,2,...., \tag{10e}$$

where $\omega_i = (\omega_{i,1},\omega_{i,2},\ldots,\omega_{i,L})$ is an L vector of model weights corresponding to each state $s_t = i$.

[8] A draw of G can be represented as

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

$$\pi_i = v_i \sum_{j=1}^{i-1}(1-v_j), \quad v_j \overset{iid}{\sim} Beta(1,\alpha),\ j=1,2,\ldots.$$

[9] In practice posterior simulation will explore a finite number of weight vectors for a finite dataset.

ω_{s_t} changes over time according to the first-order Markov chain with infinite dimension. $Dir(\cdot)$ stands for Dirichlet distribution with dimension L. State variable s_t follows an infinite Markov transition matrix, Π, with priors governed by a hierarchical Dirichlet process. $\Pi_{s_{t-1}}$ 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 $\kappa\delta_i$ means that to element $\alpha\Gamma_i$ (*ith* element) is added $\kappa \geq 0$. Through κ state persistence can be reinforced. Larger values of κ favour self transition of states while $\kappa = 0$ gives the standard HDP for infinite hidden Markov models.

$\eta > 0$ and $\alpha > 0$ are two layers of precision parameters that govern the likelihood of introducing new states in the HDP. Small values of η and α 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 processes (Sethuraman 1994). Let $\Gamma = \{\gamma_1 \ldots, \gamma_\infty\}$ and π_{ij} be the ith row and jth column of Π. The distributional draw $\Gamma \sim Stick(\eta)$ can be represented as $\Gamma = \sum_{i=1}^{\infty} \gamma_i \delta_i$, where δ_i is a point mass at i and γ_i is the associated probability mass that is generated as

$$\gamma_i = v_i \prod_{l=1}^{i-1} (1 - v_l), \quad v_j \overset{iid}{\sim} \text{Beta}(1, \eta), \ \ j = 1, 2, ... \tag{11}$$

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 π_{ij} are generated as

$$\pi_{ij} = \hat{\pi}_{ij} \prod_{l=1}^{j-1} (1 - \hat{\pi}_{il}), \quad \hat{\pi}_{ij} \overset{iid}{\sim} \text{Beta}\left(\alpha\gamma_j + \kappa\delta_i, \alpha + \kappa - \sum_{l=1}^{j} (\alpha\gamma_l + \kappa\delta_i)\right), \quad j = 1, 2, ... \tag{12}$$

Each row Π_i of the transition matrix is centered on Γ in that $E(\pi_{ij}) = \Gamma_i$ and $\text{Var}(\pi_{ij}) = \Gamma_i(1 - \Gamma_i)/(1+\alpha)$. As the magnitude of η and α 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, $\eta \sim Gamma(c_0, c_1), \alpha + \kappa \sim Gamma(c_2, c_3), \rho \sim Beta(c_4, c_5)$ with $\rho = \alpha/(\alpha + \kappa)$ which is easier to sample. We set a hyper-prior on α_ω as $\alpha_\omega \sim Gamma(c_6, c_7)$. Setting a hyper-prior on α_ω makes sampling the posterior of ω_k within each state group more flexible than a fixed α_ω.

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

$$p(y_t|y_{1:t-1}, s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1}s_t} p(y_t|y_{1:t-1}, s_t) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1}s_t} \sum_{l=1}^{L} \omega_{s_t,l} p(y_t|y_{1:t-1}, M_l) \tag{13a}$$

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

3.2 Posterior Sampling

To facilitate sampling, we introduce an indicator variable $z_t \in \{1, 2, \ldots, L\}$, indexing the model assigned to observation t. We jointly sample the state variable s_t and model variable z_t. Our mechanism for sampling $z_{1:T}$ is an important contribution and 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 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 a 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:

$$p(u_t|s_t, s_{t-1}, z_t, \Pi, \omega) = \frac{\mathbf{1}(u_t < \pi_{s_{t-1},s_t}\omega_{s_t,z_t})}{\pi_{s_{t-1},s_t}\omega_{s_t,z_t}}. \tag{14}$$

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 (13a) 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 $\omega = (\omega_1, \ldots\ldots, \omega_K)$, and each of its element 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

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

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

$$\begin{aligned} & p(s_t = k, z_t = q | u_{1:T}, \Pi, \omega, y_{1:t}) \\ & \propto p(s_t = k, z_t = q | u_{1:T}, \Pi, \omega, y_{1:t-1}) p(y_t | y_{1:t-1}, M_q). \end{aligned}$$

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 accordingly sort $\omega_{1:K}$ and $\Pi_{1:K+1,1:K+1}$. Each sweep of the sampler updates the value of K. $\{u_{1:T}, s_{1:T}, z_{1:T}, \eta, \alpha, \alpha_\omega, \Gamma_{1:K}, \Pi_{1:K+1,1:K+1}, \omega_{1:K}, K\}$ is the parameter set. Posterior sampling is sequentially repeated from the following conditional posterior distributions:

$$\begin{array}{ll} p(u_{1:T} | s_{1:T}, z_{1:T}, \Pi_{1:K+1,1:K+1}, \omega_{1:K}) & p(s_{1:T}, z_{1:T} | \Pi_{1:K+1,1:K+1}, \omega_{1:K}, u_{1:T}, y_{1:T}) \\ p(\omega_{1:K} | z_{1:T}, s_{1:T}, \alpha_\omega) & p(\Pi_{1:K+1,1:K+1} | s_{1:T}, \Gamma_{1:K}, \alpha) \\ p(\Gamma_{1:K} | s_{1:T}, \eta, \alpha) & p(\eta, \alpha, \rho, \kappa | s_{1:T}, \Gamma_{1:K}) \\ p(\alpha_\omega | \omega_{1:K}). & \end{array}$$

Each of these steps are detailed in the 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 interested is ω_{s_t} as it indicates the model pooling dynamics at time t. We estimate the posterior mean as

$$E(\omega_{s_t} | y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} \omega^{(i)}_{s_t^{(i)}}, \text{ for } t = 1, \ldots, T, \tag{16}$$

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

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. To compute the predictive density and predictive likelihood we do the following steps.

1. For each MCMC draw of s_T, simulate the future state s_{T+1} according to the Markov transition probability Π_{s_T}.

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 $\omega_{s_{T+1}} \sim Dir(\frac{\alpha_\omega}{L}, \ldots, \frac{\alpha_\omega}{L})$.

The predictive density can be estimated as

$$p(y_{T+1}|y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} \omega^{(i)}_{s^{(i)}_{T+1},l} p(y_{T+1}|y_{1:T}, M_l), \tag{17}$$

which integrates out parameter and distributional uncertainty. A predictive likelihood value is obtained by evaluating (17) at the realized data y_{T+1}. Similarly, predictive moments can be estimated such as the predictive mean,

$$E(y_{T+1}|y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} \omega^{(i)}_{s^{(i)}_{T+1},l} E(y_{T+1}|y_{1:T}, M_l), \tag{18}$$

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

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

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

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 need to be clarified. Each individual model requires a history of data for estimation as does the pooling models. Each individual model is assumed to use all data from $t = 1$ 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 τ_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 we compare forecasts for all models over a common out-of-sample period $y_{\tau_1}, \ldots, y_{\tau_2}$. Individual models use data from $t = 1$ and onward to compute a forecast while 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 four groups of models to forecast interest rates using different model combination approaches.

4.1 Models

4.1.1 Basic Models

The following general form summarizes the basic group:

$$r_t - r_{t-1} = \lambda + \beta r_{t-1} + e_t \quad e_t \sim N(0, \sigma^2 r_{t-1}^x), \tag{19}$$

where r_t is the short-term interest rate. The conditional variance is $\sigma^2 r_{t-1}^{2x}$. 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 the only restriction of $x = 2$ (GBM). The fourth model is denoted as MER introduced by Merton (1973), which sets $\beta = 0$ and $x = 0$. The last model is introduced by Brennan & Schwartz (1977), Brennan & Schwartz (1979) and Brennan & Schwartz (1980). It is denoted as GBM and it is restricted by $\lambda = 0$ and $x = 2$. These five models are defined as the basic group.

4.1.2 MS2 Models

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 and are direct extensions to the previous group with λ, β and σ becoming state dependent.

These are

$$r_t - r_{t-1} = \lambda_{s_t} + \beta_{s_t} r_{t-1} + e_t \quad e_t \sim N(0, \sigma^2_{s_t} r^x_{t-1}) \tag{20a}$$

$$s_t \in (1,2) \quad s_t | s_{t-1} \sim \Pi. \tag{20b}$$

VSK-MS2, CIR-MS2 and BSZ-MS2 correspond to $x = 0$, $x = 1$ and $x = 2$. GBM-MS2 which imposes $\lambda_{1:2} = 0$ and $x = 2$ of above equation. MER-MS2 takes the constraints of $\beta_{1:2} = 0$ and $x = 0$. These five models are denoted as the five models in MS2 group.

4.1.3 IHMM Models

The IHMM group contains five models which like the MS2 extend the benchmark set of models to be governed by a 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 IHMM group are defined by,

$$\Gamma \sim Stick(\eta), \quad \Pi_i \sim DP\left(\alpha, \Gamma\right), i = 1, 2, ..., \tag{21a}$$

$$r_t - r_{t-1} = \lambda_{s_t} + \beta_{s_t} r_{t-1} + e_t \quad e_t \sim N(0, \sigma^2_{s_t} r^x_{t-1}) \tag{21b}$$

$$s_t | s_{t-1} \sim \Pi_{s_{t-1}}, \quad s_t \in \{1, 2, \ldots, \} \tag{21c}$$

The models of IHMM group allows the λ, β and σ to change through a infinite Markov transition probability.

We denote VSK-IHMM, CIR-IHMM and BSZ-IHMM by letting $x = 0$, $x = 1$ and $x = 2$ in (21). GBM-IHMM which imposes $\lambda_{1:\infty} = 0$ and $x = 2$ of above equation. MER-IHMM takes the constraints of $\beta_{1:\infty} = 0$ and $x = 0$.

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 conditional variance with the Student-t error component. The five models in the GARCHt group are summarized through,

$$r_t - r_{t-1} = \lambda + \beta r_{t-1} + e_t \quad e_t \sim St(0, \sigma^2_t r^x_{t-1}, \nu) \tag{22a}$$

$$\sigma_t = \alpha_0 + \alpha_1 \mathrm{e}^2_{t-1} + \alpha_2 \sigma^2_{t-1} \tag{22b}$$

where $St(0, \sigma_t^2 r_{t-1}^x, \nu)$ denotes a Student-t density with mean 0 and variance $\sigma_t^2 r_{t-1}^x \nu/(\nu - 2)$ for $\nu > 2$. VSK-GARCHt, CIR-GARCHt and BSZ-GARCHt correspond to $x = 0$, $x = 1$ and $x = 2$ of (22). GBM-GARCHt imposes $\lambda = 0$ and $x = 2$ while MER-GARCHt has $\beta_1 = 0$ and $x = 0$.

4.1.5 Priors

For individual models of Basic, MS2, IHMM and GARCHt groups, the prior for λ and β are independent standard normal. Let $\sigma \sim Gamma(5, 1)$. Additionally, each row of Π in the MS2 group follows a Dirichlet distribution with a vector of one. An independent $N(0, 100)$ applies to α_0, α_1 and α_2 with the restrictions $\alpha_0 > 0, \alpha_1 \geq 0, \alpha_2 \geq 0$, $\nu \sim U(2, 100)$ in the GARCHt models.

In terms of pooling, we set the following hyper-prior on infinite Markov pooling (IMP),

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

We sample α and κ together and let $\rho = \frac{\kappa}{\alpha+\kappa}$. Finally, $\alpha_\omega \sim Gamma(4, 1)$. These prior settings for IMP are used in the three empirical examples unless otherwise stated. The priors for alternative forecast combination methods are referred to in Section 2.

4.2 Data

Data are monthly three-month treasury bill rates from the secondary market (T-Bill rate) and downloaded from Federal Reserves at St.Louis[10]. The 1,033 observations span January 1934 to January 2020. Figure 1 shows the time-series of T-Bill rates. The upper panel illustrates that T-Bill rates peaked during the 1980s and remained near zero for almost 10 years after 2008 making this a challenging dataset for any model to capture. The bottom panel shows the first difference of T-Bill rates.

4.3 Posterior Analysis

Figure 2 displays the posterior average of weight allocations, $E[\omega_{s_t}|y_{1:T}]$, in pooling the different groups of models: Basic, MS2, IHMM and GARCHt. There is strong evidence of changing weights across all model groups. Several models jointly a play significant role within each group.

For the Basic group, the CIR is the top model based on weight across periods and this tends to extend to the MS2 models as well. However, in the IHMM class the the IHMM-MER captures more weight. In the GARCHt group the MER-GARCHt has the dominate weight over time.

[10] https://fred.stlouisfed.org/series/TB3MS

IMP allows for as many weight vectors as needed. As discussed before if there was one weight vector this corresponds to the Geweke & Amisano (2011) optimal prediction pool. Figure 3 shows the posterior of unique states or weight vectors. Most of the mass is from 3 to 9 regimes. It is switches between these different weight vectors that result in a time-varying weight for a specific model as seen in Figure 2.

4.4 Out-of-Sample Forecasts

Forecasts are computed from $\tau_1 = 21$ to $\tau_2 = T = 1033$ with $\tau_0 = 10$ giving a total of 1013 out-of-sample periods from September-1935 to January-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. Cumulative log-predictive likelihood (LPL) and RMSFE are reported in Table 1. Each section of the table reports forecast results for each individual model of a group in addition to various model combination approaches. The final panel of the table pools over several groups of models using the IMP approach.

With the exception of the GARCHt group the CIR specification achieves the largest LPL value in each of the other groups. Sometimes model pooling approaches beat the CIR specification and other time not. Only the IMP approach consistently produces the largest LPL value in each group and overall. Figure 4 displays cumulative log-Bayes factors of the individual models in a group against the IMP of those models. In some periods the IMP makes large gains while in others they are minor indicating similar forecast quality. Among the different groups the largest LPL value is from the IMP over the GARCHt specifications with a value of 587.1.

The final panel of Table 1 considers pooling over the different groups: basic, MS2, IHMM and GARCHt. Two additional entries are reported. IMP-20 pools over all 20 individual models while IMP-10 pools over the 10 models in the groups IHMM and GARCHt. Both of these larger pools increase the LPL by over 10 making them strongly favoured over the other pools. The IMP-10 has the largest value and a log-Bayes factor of 17.9 against the IMP of the GARCHt models. These results seem to indicate that pooling over more models is desirable but which class to pool over can influence results. The IMP approach produces the smallest RMSFE but the improvement over other models and pooling methods is small.

Figure 5 plots the aggregated 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 group of models are the major contributors to the pooling model but their relative contributions changes over time.

4.5 Robustness

Table 2 displays forecast results for various subsamples 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 results.

Next we consider the sensitivity of results to the value of τ_0 with τ_1 and τ_2 fixed. Recall that the conditioning data or training sample before the first forecast is $\tau_1 - 1 - (\tau_0 - 1) = \tau_1 - \tau_0$. Table 3 report forecast results for the IMP approach for IHMM and GARCHt group models. Each column shows forecast results for different training samples. For pooling over the IHMM models a larger training sample improves LPL values as more data aids learning about the nonparametric structure of the IHMM. The IMP of GARCHt models shows the training sample size has no impact on forecast results.

Finally, Table 4 displays forecast results for the IMP-20 pooling specification with different prior settings for α, η and ρ which govern states dynamics and the active state dimension. The base prior used in estimation is $a_1 = 2, b = 1, a_2 = 3, b_2 = 1, a_3 = 3, b_3 = 1$ with a LPL value of 598.2 and RMSFE of 0.3638 In general the results are insensitive to different prior settings for the IMP-20 specification.

5 Realized Covariance (RCOV) Models Application

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

5.1 Models

Let Σ_t, $t = 1, \ldots, T$ denote a realized covariance matrix of dimension k and $\Sigma_{1:t-1} = \{\Sigma_1, \ldots, \Sigma_{t-1}\}$. Even though 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 which are introduced in last decade.

The first model is from Jin & Maheu (2013) and named as additive component Wishart model

(W) in the following way,

$$\begin{aligned}\Sigma_t|\Sigma_{1:t-1} &\sim \text{Wishart}_k(\nu, V_t/\nu),\\ V_t = B_0{+}\sum_{j=1}^{M} B_j \odot \Gamma_{t-1,\ell_j} &\quad \Gamma_{t-1,\ell} = \frac{1}{\ell}\sum_{i=1}^{\ell}\Sigma_{t-i},\end{aligned}$$

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 ν 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. Γ_{t-1,ℓ_j} is the j^{th} (additive) component defined as the average of past Σ_t over ℓ_j observations. The first component has one lag, $\ell_1 = 1$, while ℓ_2 and ℓ_3 are estimated which can lead to significantly better forecasts than assigning preset values. In our Bayesian inference, the priors on the elements of b_j's are all $N(0, 100)$, except the first element of each b_j is truncated to be positive for identification purposes, and $\nu \sim exp(100)I_{\nu>k-1}$, an exponential distribution with support truncated to be greater than $k - 1$. The prior for ℓ_2, ℓ_3 are uniform discrete over $\{2, \ldots, 200\}$ with $\ell_2 < \ell_3$. See Jin & 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,

$$\begin{aligned}\Sigma_t|\Sigma_{1:t-1} &\sim \text{invWishart}_k(\nu, (\nu - k - 1)V_t)\\ V_t = &B_0 + \sum_{j=1}^{M} B_j \odot \Gamma_{t-1,\ell_j} \quad \Gamma_{t-1,\ell} = \frac{1}{\ell}\sum_{i=1}^{\ell}\Sigma_{t-i}.\end{aligned}$$

$\text{invWishart}_k(\nu, (\nu-k-1)V_t)$ denotes an inverse-Wishart distribution over positive definite matrices of dimension k with ν 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. See Jin & Maheu (2016) for estimation details.

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

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

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

$\text{NCW}_k(\nu, V_t/\nu, \Lambda_t)$ is a noncentral Wishart distribution over positive definite matrices of dimension k. ν is the real-valued degree of freedom and $\nu > k-1$. V_t/ν and Λ_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, M_i are $k \times k$. Following the results in GCAW we use their best model 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 use 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 from a random walk proposal.

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

$$\begin{aligned} \Sigma_t|\Sigma_{1:t-1} &\sim \text{Wishart}_k(\nu, V_t/\nu) \\ V_t &= CC' + \sum_{i=1}^{p} B_i V_{t-i} B_i' + \sum_{i=1}^{q} A_i \Sigma_{t-i} A_i'. \end{aligned}$$

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

The last model is based on the matrix discounting model (West & Harrison 1997) adapted by Jin et al. (2019) to model RCOV matrices.

$$\begin{aligned} \Sigma_t|\Sigma_{1:t-1} &\sim \text{invWishart}_k(\beta n_t + k - 1, \beta n_t S_t) \\ n_t &= \beta n_{t-1} + 1 \quad S_t = \frac{1}{n_t}(\beta n_{t-1} S_{t-1} + \Sigma_{t-1}) \end{aligned}$$

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. (2012).[11] The list of stocks used are: 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

[11] http://realized.oxford-man.ox.ac.uk/data/download

(MSFT), and Exxon Mobil (XOM). The data ranges from 2001/02/01 to 2009/12/31 (2092 obs). Here $\tau_0 = 1400$, $\tau_1 = 1550$ and $\tau_2 = T = 2092$. As a result, a total 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 log-predictive likelihood (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 Σ_t and the associated predictive mean, $E[\Sigma_{t,ij}|\Sigma_{1:t-1}, A]$ for model A, and $k = 10$.

Table 5 shows the forecast performance of each RCOV model as well as all combination approaches.[12] According to the LPL values and among individual models the additive component inverse-Wishart (IW) performs the best and has log-Bayes factor of more than 8600 over the second best model, the conditional autoregressive Wishart (CAW) specification. The discounting 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 58 over WZ, 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 6. Although the discounting model is uniformly the worse 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 whole out-of-sample time and seems to be dominated by the CAW.

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

[12] The BCRV approach was excluded as its present form was not designed for combining matrix forecasts without further assumptions.

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

6 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, we use the two asset pricing models to forecast returns using the 10 industry portfolios from the Kenneth R. French data library[13] which include the FF factors while the Q-factor data is obtained separately.[14] The data are monthly value-weighted returns, from January 1967 to Dec 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 the models is the factors included in the conditional mean.

6.1 Models and Data

The following model is the 5-factor (FF) version by Fama & French (2015) which postulates excess returns follow:

$$r_t - r_{ft} = \alpha + \beta_1 f_{1t} + \beta_2 f_{2t} + \beta_3 f_{3t} + \beta_4 f_{4t} + \beta_5 f_{5t} + e_t \qquad e_t \sim N(0, \sigma^2). \tag{28}$$

The Q-factor (Qf) model of Hou et al. (2015), Hou et al. (forthcoming) is the following,

$$r_t - r_{ft} = \alpha + \beta_1 q_{1t} + \beta_2 q_{2t} + \beta_3 q_{3t} + \beta_4 q_{4t} + \beta_5 q_{5t} + e_t \qquad e_t \sim N(0, \sigma^2). \tag{29}$$

Here r_t denotes the portfolio return and r_{ft} is the risk free rate. The factors $f_{1t}, \ldots, f_{5t}$ are excess market returns, return difference between diversified small & big stocks, robust & weak profitability firm, low & high investment firms, and high & low Book to Market (B/M) firms, respectively. Let $q_{1t}, \ldots, q_{5t}$ denote market excess returns, size factor returns, investment factor returns, equity factor returns and expected growth returns, respectively.

The original models of Fama & French (2015) and Hou et al. (2015) are homoskedastic models which did not consider volatility dynamics. To improve density forecasts we introduce het-

[13] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[14] http://global-q.org/factors.html

eroskedastic versions of (28) and (29) by replacing the σ^2 with σ_t^2 in the following GARCH model,

$$\sigma_t^2 = \omega_0 + \omega_1 e_{t-1}^2 + \omega_2 \sigma_{t-1}^2. \tag{30}$$

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

We set the priors of $\alpha, \beta_1, \ldots, \beta_5$ to follow independent standard normal distributions and assume $\sigma^{-2} \sim Gamma(3, 1)$. Priors for ω_0, ω_1 and ω_2 are independent $N(0, 100)$ with $\omega_0 > 0$, $\omega_1 \leq 0$, $\omega_2 \leq 0$ and $\omega_1 + \omega_2 < 1$.

6.2 Results

The posterior mean of the weights from the IMP model are displayed in Figures 7 and 8. A model with GARCH generally receives the largest weight but the homoskedastic versions receive significant weighs also. With the exception of the Durable portfolio the time-varying nature of the weights is clear.

Table 6 reports the out-of-sample forecast performance for the four individual models as well as combination approaches. In five cases the IMP method has the largest LPL value and in two other cases is has essentially the same value as the top performing model (Durbl and Healt). Only in the case of Other portfolio is the evidence strongly against the IMP with a log-Bayes factor of 5.5 in favour of the FF-GARCH. The picture is different for point forecasts. Only in one portfolio does the IMP method have the lowest RMSFE. There is no dominate model for point forecasts. The best is the BCRV but the differences amongst models is relatively small.

Overall in terms of density forecasts the IMP approach is very competitive compared to individual models and other combination methods. Point forecast accuracy is very similar among models and no one specification consistently provides the lowest values.

7 Conclusion

This paper introduces a new approach to forecast pooling methods based on a nonparametric prior for the weight vector combining predictive densities. A hierarchical Dirichlet process 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, realized covariance matrices and asset pricing models show the nonparametric pool forecasts well.

References

Aastveit, K., Ravazzolo, F. & van Dijk, H. K. (2018), 'Combined density nowcsting in an uncertain economic environment', *Journal of Business & Economic Statistics* **136**(1), 131–145.

Ang, A. & Bekaert, G. (2002), 'Short rate nonlinearities and regime switches', *Journal of Economic Dynamics and Control* **26**(7–8), 1243 – 1274.

Antoniak, C. E. (1974), 'Mixtures of dirichlet processes with applications to bayesian nonparametric problems', *The Annals of Statistics* **2**(6), 1152–1174.

Billio, M., Casarin, R., Ravazzolo, F. & van Dijk, H. K. (2012), 'Combination schemes fro turning point predictions', *The Quarterly Review of Economics and Finance* **52**, 402–412.

Billio, M., Casarin, R., Ravazzolo, F. & Van Dijk, H. K. (2013), 'Time-varying combinations of predictive densities using nonlinear filtering', *Journal of Econometrics* pp. 213–232.

Black, F. & Scholes, M. (1973), 'The pricing of options and corporate liabilities', *Journal of Political Economy* pp. 637–654.

Brennan, M. & Schwartz, E. (1977), 'Saving bonds, retractable bonds, and callable bonds', *Journal of Financial Economics* pp. 67–88.

Brennan, M. & Schwartz, E. (1979), 'A continuous time approach to the priceing of bonds', *Journal of Banking and Finance* pp. 133–155.

Brennan, M. & Schwartz, E. (1980), 'Analyzing convertible bonds', *Journal of Financial and Quantitative Analysis* pp. 907–929.

Chib, S. (1996), 'Calculating posterior distributions and modal estimates in Markov mixture models', *Journal of Econometrics* **75**, 79–97.

Cox, J., Ingersoll, J. & Ross, R. (1985), 'A theory of the term structure of interest rates', *Econometrica* **53**, 385–407.

Del Negro, M., Hasegawa, R. B. & Schorfheide, F. (2016), 'Dynamic prediction pools: an investigation of financial frictions and forecasting performance', *Journal of Econometrics* pp. 391–405.

Durham, G. B. (2003), 'Likelihood-based specification analysis of continuous-time models of the short-term interest rate', *Journal of Financial Economics* **70**(3), 463 – 487.

Fama, E. F. & French, K. R. (2015), 'A five-factor asset pricing model', *Journal of Financial Economics* pp. 1–22.

Fox, E., Sudderth, E., Jordan, M. & Willsky, A. (2011), 'A sticky hdp-hmm with application to speaker diarization', *Annals of Applied Statistics* **5**, 1020–1056.

Geweke, J. & Amisano, G. (2011), 'Optimal prediction pools', *Journal of Econometrics* pp. 130–141.

Golosnoy, V., Gribisch, B. & Liesenfeld, R. (2012), 'The conditional autoregressive wishart model for multivariate stock market volatility', *Journal of Econometrics* **167**, 211–223.

Guidolin, M. & Timmermann, A. (2009), 'Forecasts of us short-term interest rates: a flexible forecast combination approach', *Journal of Econometrics* pp. 297–311.

Hall, S. G. & Mitchell, J. (2007), 'Combining density forecasts', *International Journal of Forecasting* pp. 1–13.

Hoogerheide, L., Kleijn, R., Ravazzolo, F., Van Dijk, H. K. & Verbeek, M. (2010), 'Combining density forecasts', *Journal of Forecasting* pp. 251–269.

Hou, K., Chen, X. & Zhang, L. (forthcoming), 'An augmented q-factor model with expected growth', *Review of Finance* .

Hou, K., Xue, C. & Zhang, L. (2015), 'Digesting anomalies: an investment approach', *The Review of Financial Studies* pp. 650–705.

Jin, X. & Maheu, J. (2013), 'Modellling realized covariance and returns', *Jounral of Financial Econometrics* **11**, 335–369.

Jin, X. & Maheu, J. (2016), 'Bayesian semiparametric modeling of realized covariance matrices', *Jounral of Econometrics* **192**, 19–39.

Jin, X., Maheu, J. & Yang, Q. (2019), 'Bayesian parametric and semiparametric factor models for large realized covariance matrices', *Journal of Applied Econometrics* **34**, 641–660.

Kapetanios, G., Mitchell, J., Price, S. & Fawcett, N. (2015), 'Generalised density forecast combinations', *Journal of Econometrics* pp. 150–165.

Kascha, C. & Ravazzolo, F. (2010), 'Combining inflation density forecast', *Journal of Forecasting* pp. 231–250.

Maheu, J. & Yang, Q. (2016), 'An infinite hidden markov model for short-term interest rates', *Journal of Empirical Finance* pp. 202–220.

Merton, R. (1973), 'Theory of rational option pricing', *Bell Journal of Economics and Management Science* pp. 141–183.

Noureldin, D., Shephard, N. & Sheppard, K. (2012), 'Multivariate high-frequency-based volatility (heavy) models', *Journal of Applied Econometrics* **27**, 907–933.

Pesaran, M. H., Pettenuzzo, D. & Timmermann, A. (2006), 'Forecasting time series subject to multiple structural breaks', *Review of Economic Studies* **73**(4), 1057 – 1084.

Sethuraman, J. (1994), 'A constructive definition of dirichlet priors', *Statistica Sinica* **4**, 639–650.

Teh, Y., Jordan, M., Beal, M. & Blei, D. (2006), 'Hierarchical dirichlet processes', *Journal of the American Statistical Association* **101**, 1566–1581.

Van Gael, J., Saatci, Y., Teh, Y. & Ghahramani, Z. (2008), Beam sampling for the infinite hidden markov model, *in* 'Proceedings of the 25th International Conference on Machine Learning:', pp. 1088–1095.

Vasicek, O. (1977), 'An equilibrium characterization of the term structure', *Journal of Financial Economics* **5**, 177–188.

Waggoner, D. F. & Zha, T. (2012), 'Confronting model misspecification in macroeconomics', *Journal of Econometrics* pp. 167–184.

West, M. & Harrison, J. (1997), *in* 'Bayesian forecasting and Dynamic Models', Springer Series in Statistics, New York.

Wright, J. H. (2008), 'Bayesian model averaging and exchange rate forecasts', *Journal of Econometrics* pp. 329–411.

Yu, P. L. H., Li, W. K. & Ng, F. C. (2017), 'The generalized conditional autoregressive wishart model for multivariate realzied volatility', *Journal of Business & Economic Statistics* **35**, 513–527.

Table 1: Out-of-sample Forecast Performance of Individual Models and Pooling Approaches

Basic Group					
VSK	CIR	GBM	MER	BSZ	
-426.1 (0.3668)	204.2 (0.3660)	-708.3 (0.3829)	-423.3 (0.3642)	-653.6 (0.3644)	
Model Pooling on Basic Group					
BMA	GA	DHS	WZ	BCRV	IMP
196.1 (0.3660)	251.1 (0.3660)	202.5 (0.3662)	276.0 (0.3671)	134.9 (0.3665)	**330.4** (**0.3660**)
MS2 Group					
VSK-MS2	CIR-MS2	GBM-MS2	MER-MS2	BSZ-MS2	
-43.2 (0.3600)	280.8 (0.3660)	-330.8 (0.3711)	-121.6 (0.3664)	-325.0 (0.3676)	
Model Pooling on MS2 Group					
BMA	GA	DHS	WZ	BCRV	IMP
272.6 (0.3659)	263.5 (0.3643)	201.6 (0.3634)	286.1 (0.3623)	131.4 (0.3641)	**330.8** (**0.3628**)
IHMM Group					
VSK-IHMM	CIR-IHMM	GBM-IHMM	MER-IHMM	BSZ-IHMM	
252.6 (0.3599)	381.8 (0.3632)	257.6 (0.3696)	268.4 (0.3648)	243.4 (0.3694)	
Model Pooling on IHMM Group					
BMA	GA	DHS	WZ	BCRV	IMP
373.6 (0.3633)	403.3 (0.3625)	431.9 (0.3629)	464.4 (0.3618)	401.6 (0.3631)	**481.7** (**0.3612**)
GARCHt Group					
VSK-GARCHt	CIR-GARCHt	GBM-GARCHt	MER-GARCHt	BSZ-GARCHt	
488.6 (0.3729)	526.7 (0.3698)	269.5 (0.3759)	528.3 (0.3641)	215.1 (0.3664)	
Model Pooling on GARCHt Group					
BMA	GA	DHS	WZ	BCRV	IMP
524.6 (0.3658)	569.1 (0.3659)	567.5 (0.3686)	570.4 (0.3662)	522.2 (0.3691)	**587.1** (**0.3658**)
Pooling Over Many Groups					
	IMP-20	IMP-10	IMP-IHMM	IMP-GARCHt	
	598.2 (0.3638)	**605.0** (0.3642)	481.7 (**0.3612**)	587.1 0.3658	

Log-predictive likelihood (LPL) and root mean squared forecast errors (RMSFE) in parentheses for individual models and combination methods for the out-of-sample period 1935-09 to 2020-01 (1013 periods). Bold values denote the maximum LPL value and minimum RMSFE in a panel.

Table 2: Subsample Analysis on IHMM and GARCHt Group

	1935-Sep to 1962-June					
	BMA	GA	DHS	WZ	BCRV	IMP
IHMM	264.0	278.3	286.7	288.8	274.2	**306.0**
	(0.1542)	(0.1508)	(0.1496)	(**0.1474**)	(0.1497)	(0.1490)
GARCHt	380.3	392.1	399.3	394.5	365.2	**408.3**
	(0.1515)	(**0.1510**)	(0.1516)	(0.1511)	(0.1519)	(0.1515)
	1962-July to 1990-Jan					
	BMA	GA	DHS	WZ	BCRV	IMP
IHMM	-163.3	-150.2	-136.0	-126.0	-143.3	**-125.1**
	(**0.5880**)	(0.5908)	(0.5915)	(0.5917)	(0.5915)	(0.5907)
GARCHt	-126.5	**-116.5**	-121.9	**-116.5**	-127.8	-116.7
	(0.5935)	(0.5935)	(0.5976)	(0.5938)	(0.5982)	(**0.5934**)
	1990-Feb to 2020-Jan					
	BMA	GA	DHS	WZ	BCRV	IMP
IHMM	273.0	275.2	281.2	**301.6**	270.7	300.8
	(0.1799)	(0.1691)	(0.1700)	(0.1640)	(0.1708)	(**0.1631**)
GARCHt	270.7	293.5	290.0	292.4	284.7	**295.5**
	(**0.1786**)	(0.1805)	(0.1828)	(0.1807)	(0.1836)	(**0.1786**)

This table illustrates out-of-sample log-predictive likelihood (LPL) and root mean squared forecast errors (RMSFE) (within bracket) in subsample cases for pooling five-benchmark.

Table 3: Training Sample Sensitivity for IMP Forecasts

	Training sample size ($\tau_1 - \tau_0$)				
	10	20	50	80	110
IHMM	232.9	234.0	236.3	236.9	237.1
	(0.3938)	(0.3934)	(0.3941)	(0.3939)	(0.3937)
GARCHt	244.2	244.2	244.4	244.0	244.4
	(0.3996)	(0.3994)	(0.3991)	(0.3986)	(0.3985)

This table illustrates out-of-sample log-predictive likelihood (LPL) and root mean squared forecast errors (RMSFE) (within bracket) for IMP with IHMM and GARCHt groups for various training sample sizes $\tau_1 - \tau_0$ by varying τ_0 and fixing $\tau_1 = 83$ and $\tau_2 = T$. The out-of-sample period $t = \tau_1, \ldots, T$ is common to each column and spans 1949-April to 2020-January (850 Observations).

Table 4: Hyper-Prior Sensitivity Testing on IMP-20

	$a_1 = 1, b_1 = 1$ $a_2 = 1, b_2 = 1$	$a_1 = 3, b_1 = 1$ $a_2 = 3, b_2 = 1$	$a_1 = 6, b_1 = 1$ $a_2 = 6, b_2 = 1$
$a_3 = 5, b_3 = 1$	596.8 (0.3640)	598.6 (0.3638)	597.9.9 (0.3638)
$a_3 = 1, b_3 = 1$	598.9 (0.3640)	601.0 (0.3638)	598.0 (0.3640)
$a_3 = 1, b_3 = 2$	594.8 (0.3639)	598.0 (0.3640)	597.7 (0.3638)

This table illustrates out-of-sample log-predictive likelihood (LPL) and root mean squared forecast errors (RMSFE) (within bracket) with respect to various hyper-priors combinations. $\alpha + \kappa \sim G(a_1, b_1)$, $\eta \sim G(a_2, b_2)$ and $\rho \sim B(a_3, b_3)$

Table 5: RCOV: Model Forecasts

Individiual RCOV Models

W	IW	CAW	GCAW	Dis
-45941 (74.8323)	-36962 (76.4418)	-45600 (72.4369)	-45612 **(72.1216)**	-60712 (83.8508)

Model Pooling

BMA	GA	DHS	WZ	IMP
-36962 (76.4418)	-35037 (73.7681)	-35070 (73.1618)	-34964 (76.2534)	**-34916** (74.8056)

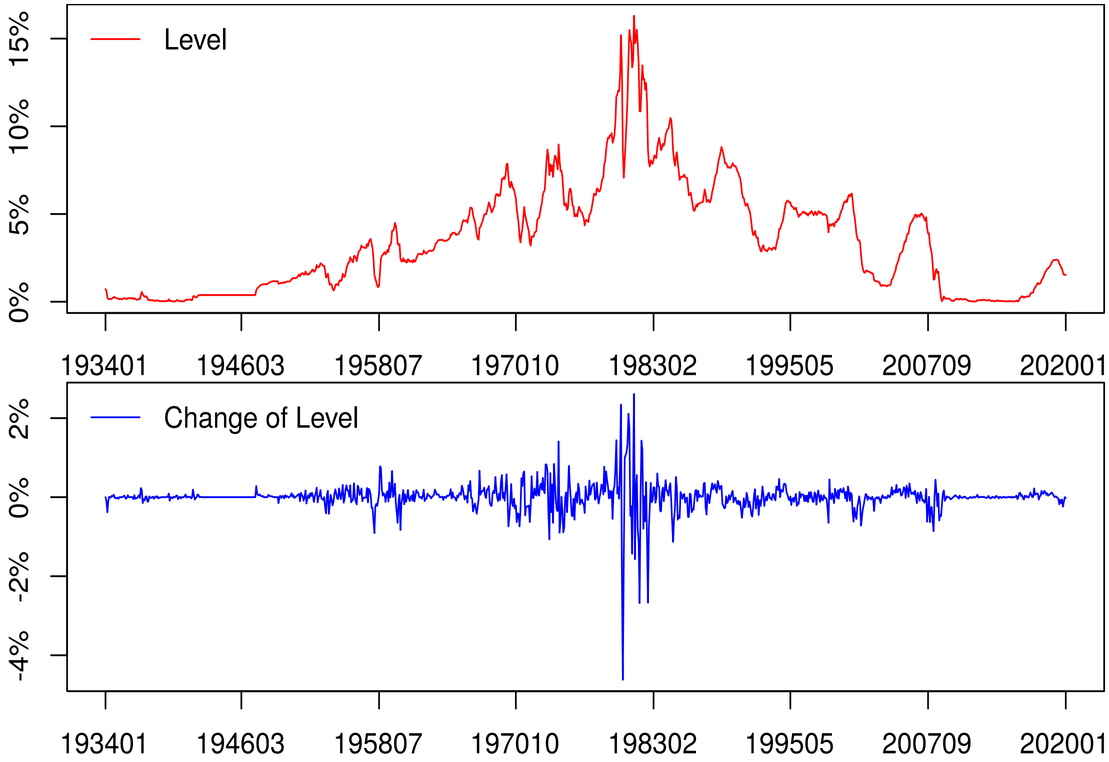
This table reports the log-predictive likelihood and root mean squared forecast errors in parentheses () for individual models and model pooling methods for 543 (2007/11/06 to 2009/12/31) out-of-sample observations.

Table 6: Forecast Performance for 10 Portfolios

	Individual Models				Combinations					
	FF	Qf	FF-GARCH	Qf-GARCH	BMA	GA	DHS	WZ	BCRV	IMP
NoDur 1	-1277.4 (2.1488)	-1342.5 (2.3916)	-1261.4 **(2.1146)**	-1308.8 (2.4185)	-1262.0 (2.1451)	-1258.2 (2.1622)	-1260.6 (2.2021)	-1259.0 (2.1793)	-1260.3 (2.2009)	**-1258.0** (2.1609)
Durbl 2	-1557.7 (3.5593)	-1582.1 (3.7332)	-1538.0 (3.5735)	-1546.8 (3.7242)	-1539.6 (3.5733)	-1522.4 (3.5461)	**-1521.3** (3.5378)	-1521.8 (3.5461)	**-1521.3** **(3.5294)**	-1521.9 (3.5481)
Manuf 3	-1107.9 (1.6343)	-1143.6 (1.7423)	-1089.3 **(1.6400)**	-1101.1 (1.7808)	-1090.7 (1.6547)	-1085.9 (1.6659)	-1085.9 (1.6610)	-1086.5 (1.6660)	-1085.8 (1.6565)	**-1085.5** (1.6632)
Energy 4	-1664.4 (4.2575)	-1646.9 (4.1590)	-1635.9 (4.2564)	**-1624.3** **(4.1339)**	-1628.3 (4.1878)	-1627.4 (4.1690)	-1629.9 (4.1689)	-1628.3 (4.1702)	-1629.5 (4.1671)	-1626.8 (4.1607)
HiTec 5	-1394.2 (2.7120)	-1414.7 (2.8037)	-1378.1 (2.7208)	-1417.0 (2.9045)	-1381.2 (2.7112)	-1373.3 (2.6807)	-1374.7 (2.6819)	-1375.2 (2.6874)	-1374.7 (2.6808)	**-1362.3** **(2.6563)**
Telcm 6	-1452.0 (3.0071)	-1445.0 (2.9741)	-1428.6 (2.9892)	-1424.7 (2.9548)	-1427.1 (2.9848)	-1424.8 (2.9521)	-1426.6 **(2.9481)**	-1425.7 (2.9506)	-1426.8 (2.9483)	**-1424.4** (2.9506)
Shops 7	-1349.9 (2.5087)	-1395.3 (2.7011)	**-1327.2** (2.4875)	-1364.7 (2.7152)	-1329.7 **(2.4846)**	-1330.7 (2.5105)	-1332.6 (2.5211)	-1332.9 (2.5232)	-1332.7 (2.5182)	-1330.0 (2.5075)
Healh 8	-1455.0 (3.0218)	-1479.2 (3.1253)	**-1422.2** (3.0565)	-1450.5 (3.1278)	-1423.5 (3.0593)	-1423.1 (3.0299)	-1425.7 **(3.0140)**	-1424.8 (3.0217)	-1425.4 (3.0172)	-1422.8 (3.0300)
Utils 9	-1481.6 (3.1745)	-1489.2 (3.2112)	-1463.7 (3.1920)	-1466.4 (3.2020)	-1464.3 (3.1940)	-1457.1 (3.1591)	-1455.1 (3.1364)	-1457.2 (3.1533)	-1454.7 **(3.1342)**	**-1451.9** (3.1353)
Other 10	-1095.2 (1.6385)	-1193.7 (1.9198)	**-1043.1** (1.7641)	-1117.9 (2.0083)	-1046.7 (1.7758)	-1050.0 (1.7592)	-1059.2 (1.7737)	-1054.2 **(1.7561)**	-1059.7 (1.7757)	-1048.6 (1.7564)

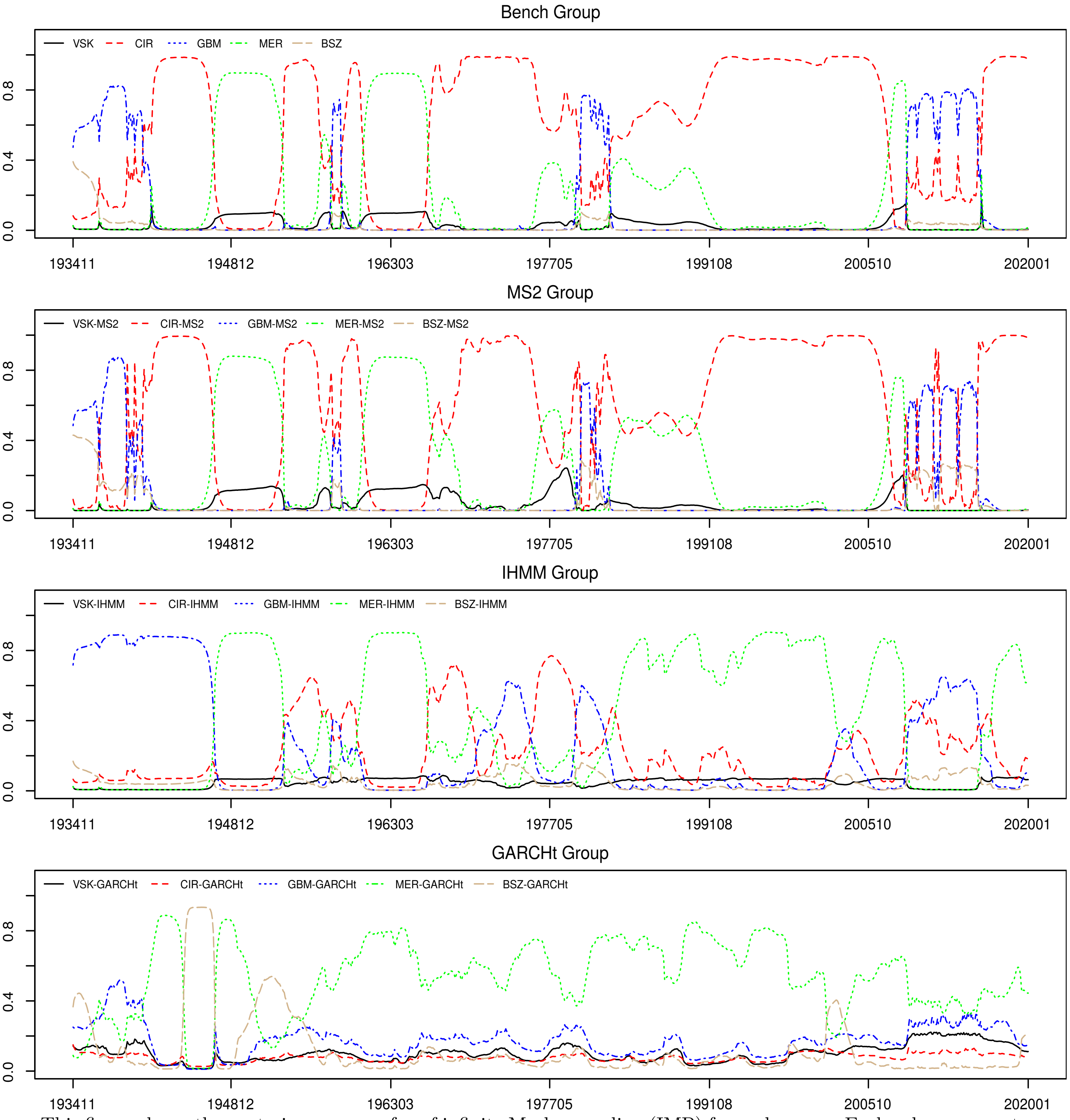
This table reports out-of-sample log-predictive likelihood (LPL) values and root mean squared forecast errors (RMSFE) (within bracket) for the out-of-sample period January 1972 – December 2019 (576 observations). The first column lists the portfolio from the 10 portfolio obtained from the Kenneth R. French Data Library. The second to fifth columns represent individual models. The sixth to eleventh columns indicate various model combinations methods.

Figure 1: T-Bill Interest Rates (Top) and Changes in Rates (Bottom)



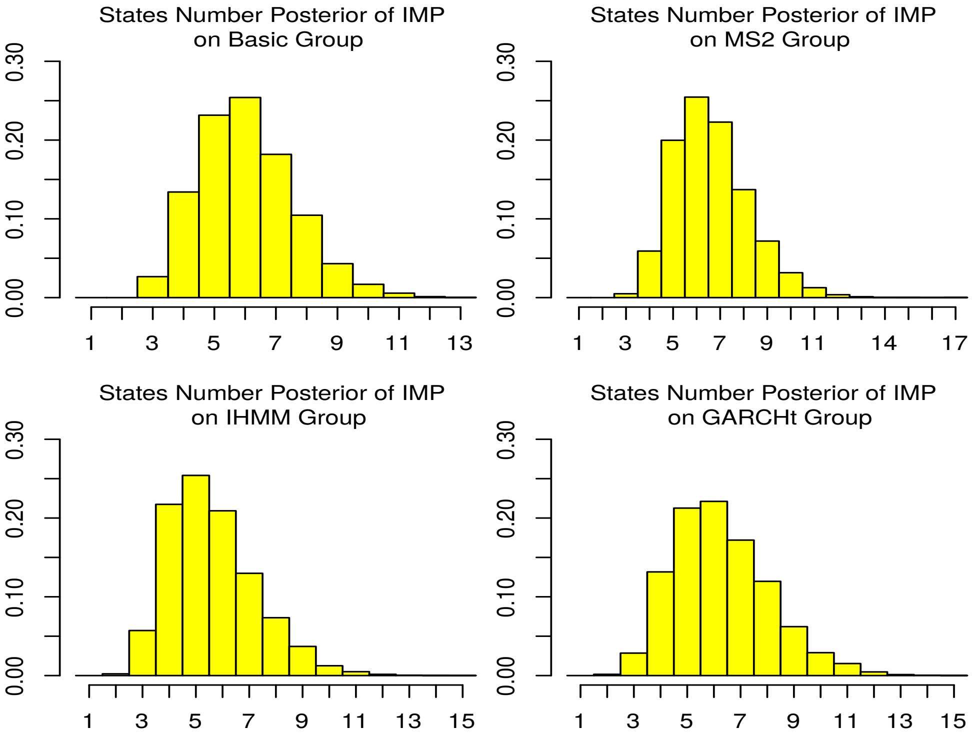
The top panel is r_t and the bottom panel is the change $(r_t - r_{t-1})$.

Figure 2: Posterior Average of Infinite Markov Pooling Weights



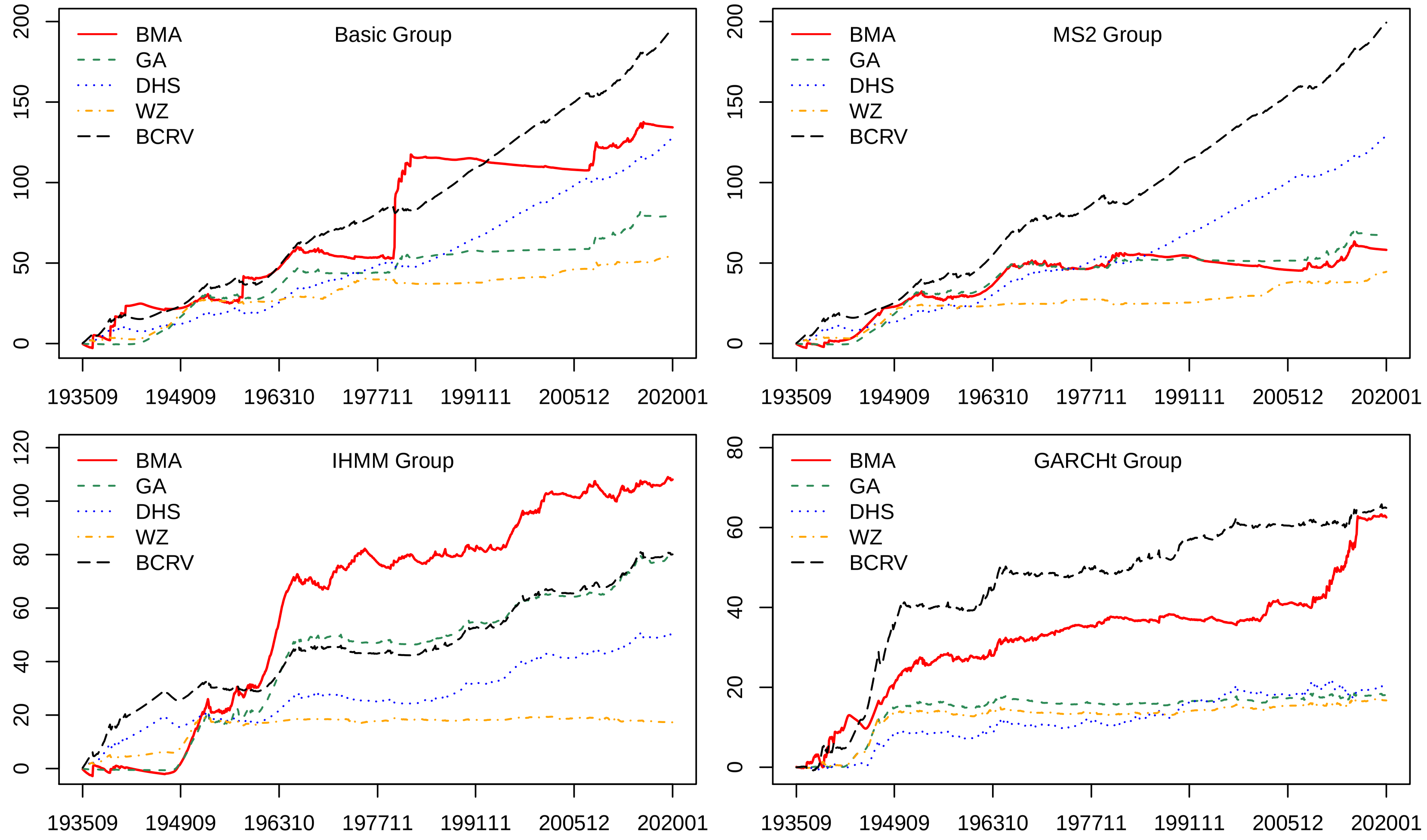
This figure shows the posterior average of ω of infinite Markov pooling (IMP) for each group. Each color represents the indicated type of model.

Figure 3: Posterior Number of Active States



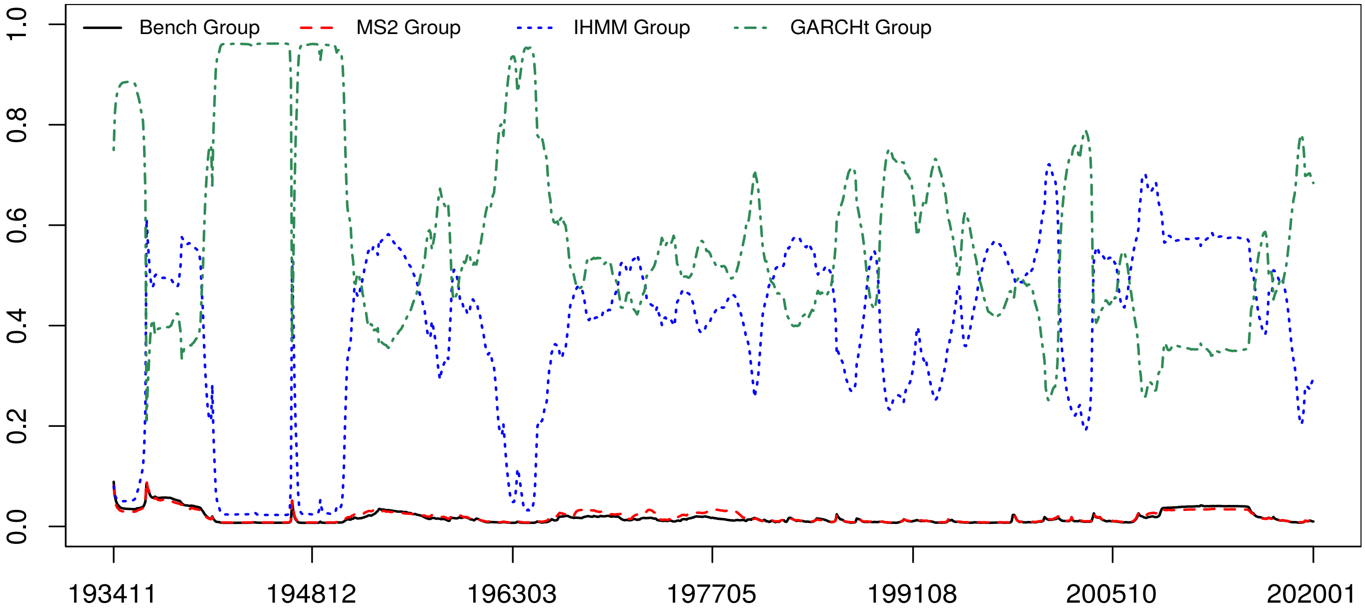
These histograms show the number of active states for Infinite Markov Pooling model.

Figure 4: Cumulative log-Bayes Factors of IMP v.s. Alternative Forecast Combination Approaches



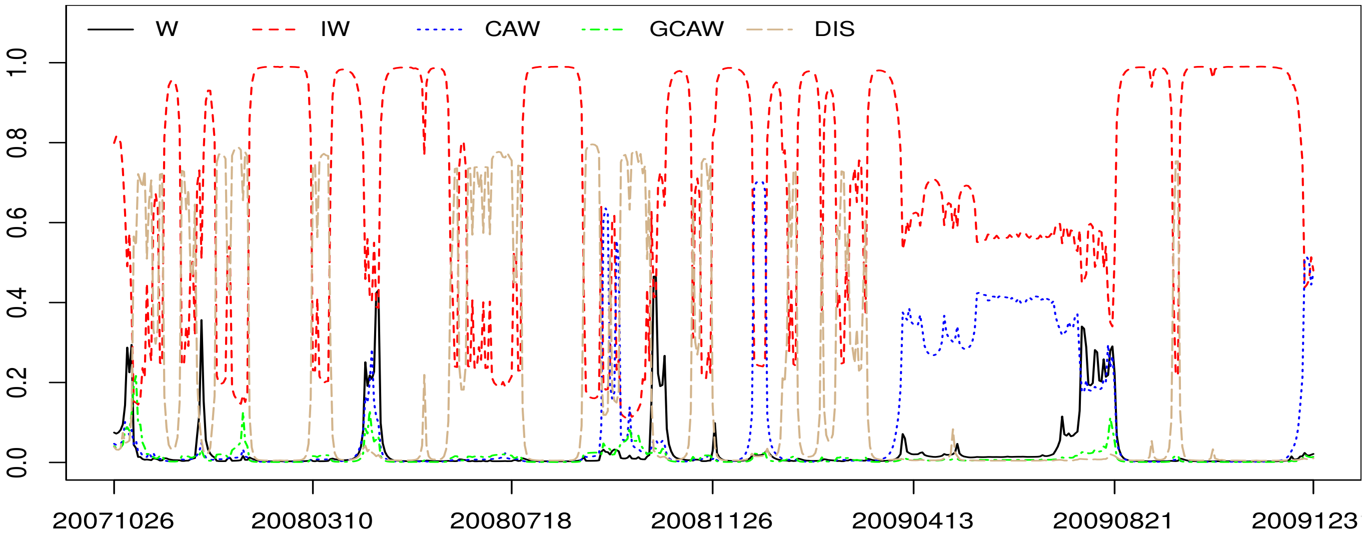
Cumulative log-Bayes factors of above combination methods over the out-of-sample period 1935-09 to 2020-01 (1013 periods).

Figure 5: Posterior Average Aggregate Weights for Model Groups



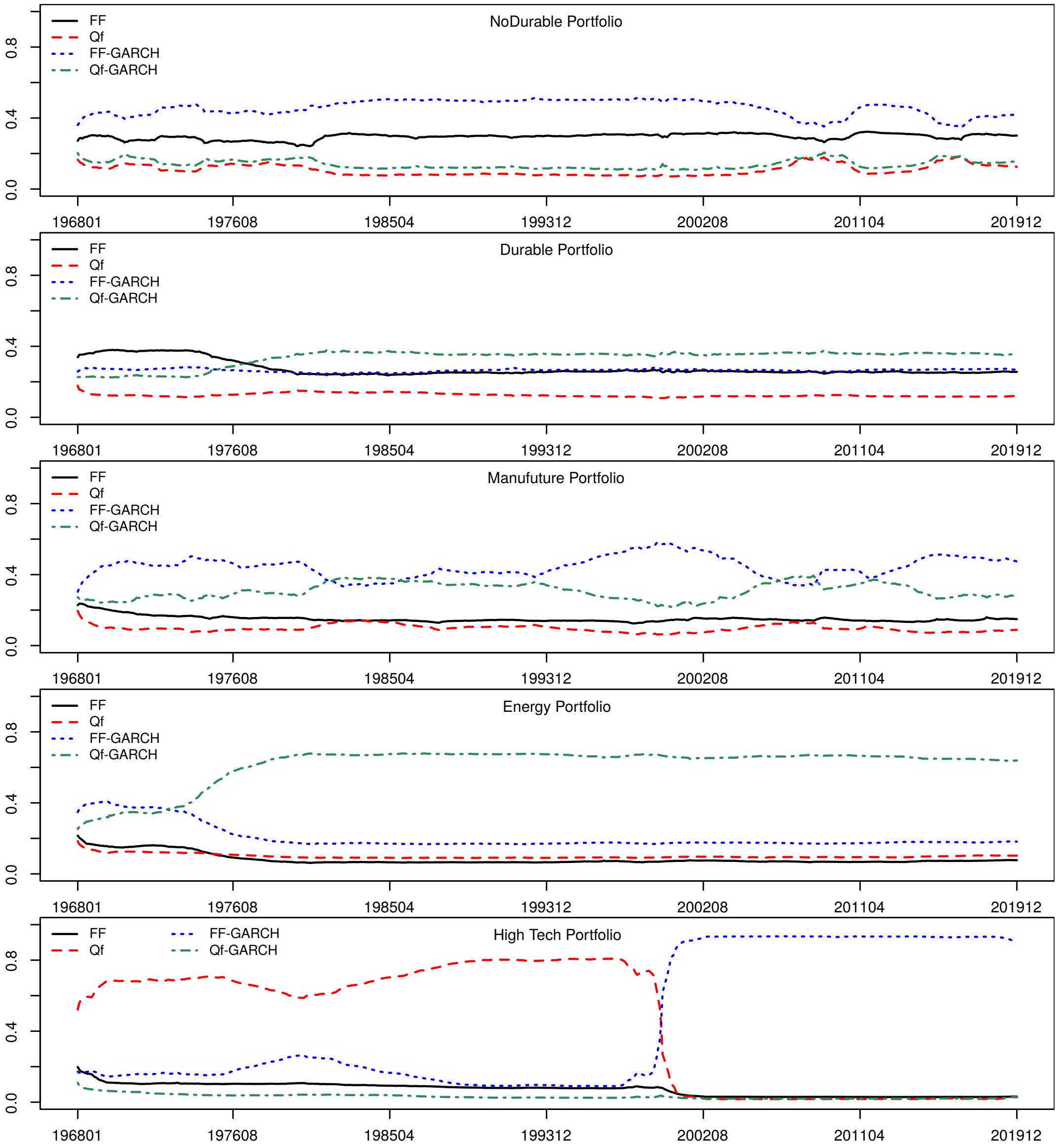
This figure shows the posterior average aggregate of ω over each group entering the infinite Markov pooling (IMP) for all four groups (20 total models) altogether.

Figure 6: RCOV: Posterior Mean of Weights from Infinite Markov Pooling



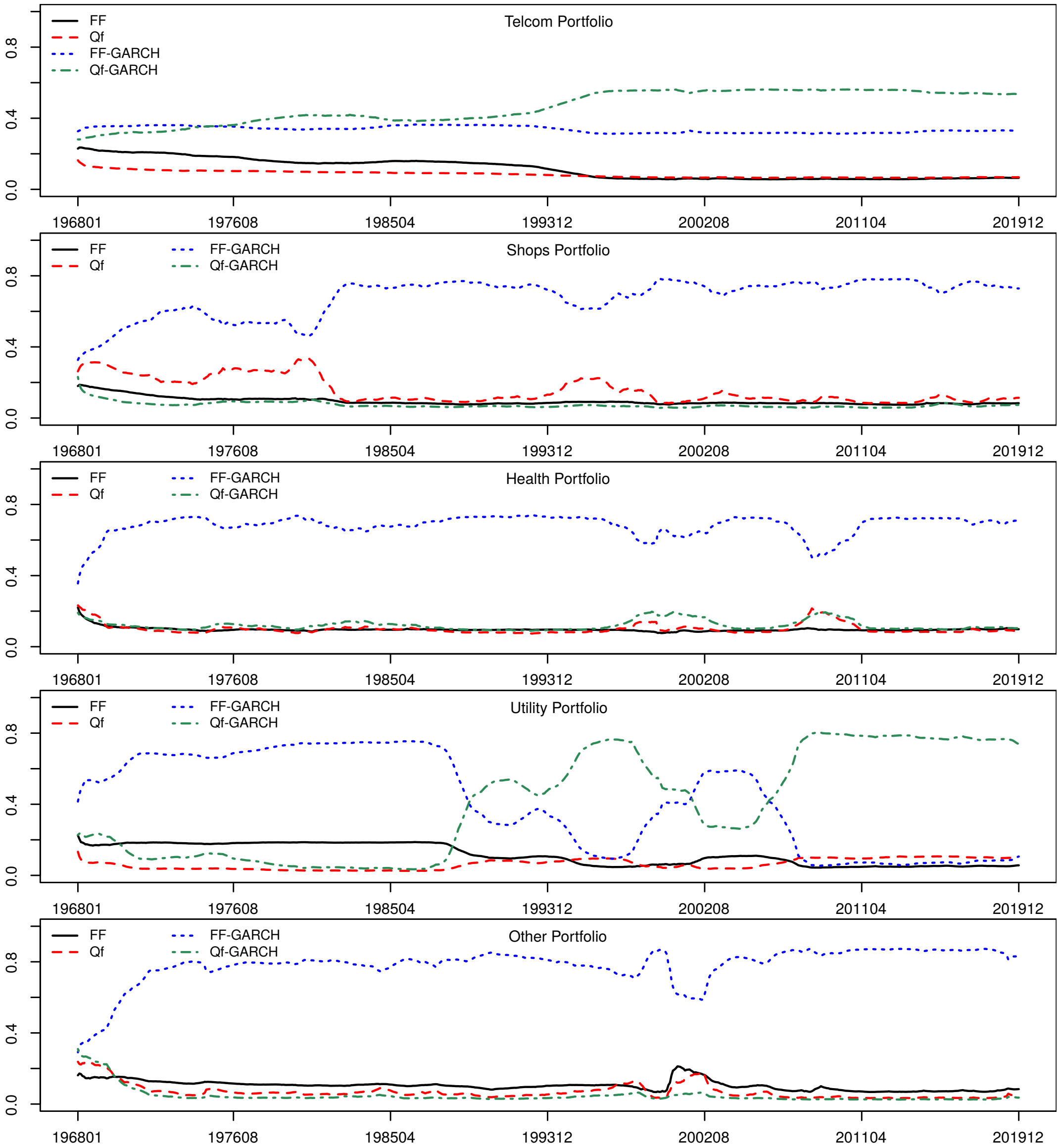
This table shows the posterior average of $\omega_{1:T}$ from pooling the five RCOV models. Each color represents the associated weight given to the indicated model.

Figure 7: Posterior Average of Weights for IMP on 10 Portfolios



This figure displays the posterior average weights from the IMP specification for 5 of the 10 portfolios. Combined individual models are: Fama and French (FF) 5-factor model, Q-factor (Qf) model and GARCH extensions (FF-GARCH, Qf-GARCH).

Figure 8: Posterior Average Weights for IMP on 10 Portfolios



This figure displays the posterior average weights from the IMP specification for 5 of the 10 portfolios. Combined individual models are: Fama and French (FF) 5-factor model, Q-factor (Qf) model and GARCH extensions (FF-GARCH, Qf-GARCH).

8 Appendix

In this section we provide details for the beam sampler.

Let K denote the number of active states in the state sequence $s_{1:T}$, and n^s_{ji} denote the number of transitions from state j to state i from $s_{1:T}$. Let n^z_{jl} denote the number of observation is assigned to model l through state j according to $s_{1:T}$ and $z_{1:T}$ that is, $n^z_{jl} = \#\{t : s_t = j, z_t = l\}$.

1. Initializing: Choose a starting value for K and a starting state sequence $s_{1:T}$ consisting of K active states which are labelled $1, \ldots, K$; The infinite many inactive states are merged into one state. Choose a starting $z_{1:T}$ sequence consisting of L models which are labelled $1, \ldots, L$; Initialize Γ and Π_k for $k = 1, \ldots, K$, all of which have $K + 1$ elements; Initialize ω_k for $k = 1, \ldots, K$; Initialize θ_l for $l = 1, \ldots, L$; Initialize $\eta, \alpha, \kappa, \alpha_\omega$.

2. Sampling $u_{1:T}$: For $t = 1, \ldots, T$, sample u_t from $\mathrm{U}(0, \pi_{s_{t-1},s_t}\omega_{s_t,z_t})$.

3. Sampling Π, expanding K: If $\max\{\pi_{k,K+1}\}_{k=1}^K > \min\{u_t\}_{t=1}^T$, repeat the following steps:

   (a) Draw $\Pi_{K+1} \sim \mathrm{Dirichlet}(\alpha\Gamma)$.

   (b) Break the last probability weight of Γ, $\Gamma_{\overline{K}+1}$:

      i. Draw $\zeta \sim \mathrm{Beta}(1, \eta)$.

      ii. Add new probability weight $\Gamma_{K+2} = (1 - \zeta)\Gamma_{K+1}$.

      iii. Update $\Gamma_{K+1} = \zeta\Gamma_{K+1}$.

   (c) Break the last probability weight of Π_k for $k = 1, \ldots, K + 1$:

      i. Draw $\zeta_k \sim \mathrm{Beta}(\alpha\Gamma_{K+1}, \alpha\Gamma_{K+2})$.

      ii. Add new probability weight $\pi_{k,K+2} = (1 - \zeta_k)\pi_{k,K+1}$.

      iii. Update $\pi_{k,K+1} = \zeta_k\pi_{k,K+1}$.

   (d) Draw $\omega_{K+1} \sim \mathrm{Dir}_L(\frac{\alpha_\omega}{L})$.

   (e) Increment K.

4. Sample $s_{1:T}, z_{1:T}$ from $p(s_{1:T}, z_{1:T}|\Pi, \omega, u_{1:T}, y_{1:T}, I_{1:T})$ using the forward filtering and backward smoothing method:

   (a) Working sequentially forwards in time for $t = 1, \ldots, T$, repeat the following steps:

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

$$\begin{aligned} &p(s_t = k, z_t = q|u_{1:T}, \Pi, \omega, y_{1:t-1}) \\ &\propto \sum_{j=1}^{K}\sum_{l=1}^{L} \mathbf{1}(u_t < \pi_{j,k}\omega_{k,q}) p(s_{t-1} = j, z_{t-1} = l|u_{1:T}, \Pi, \omega, y_{1:t-1}). \end{aligned} \tag{31}$$

Update step: for $k = 1, \dots, K$, $l = 1, \dots, L$ calculate

$$\begin{aligned} & p(s_t = k, z_t = q | u_{1:T}, \Pi, \omega, y_{1:t}) \\ & \propto p(s_t = k, z_t = q | u_{1:T}, \Pi, \omega, y_{1:t-1}) f(y_t | y_{1:t-1}, M_q). \end{aligned} \tag{32}$$

(b) Working sequentially backwards in time for $t = 1, \dots, T$, sample $s_{1:T}, z_{1:T}$:

i. Sample (s_T, z_T) from $p(s_T, z_T | u_{1:T}, \Pi, \omega, y_{1:T})$.

ii. 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 $t = T - 1, \dots, 1$.

5. Cleaning up: Update K given $s_{1:T}$, re-label all the active states in $s_{1:T}$ in the order of $1, \dots, K$ and remove the inactive states where none of the combination is assigned; Adapt Γ, Π, ω according to the new labelling; Collapse Γ_{K+1} and $\pi_{k,K+1}$ for $k = 1, \dots, K$.

6. Sampling auxiliary variables o, $\hat{o}$, $\overline{o}$: such that $o = \{o_{ji}\}$, $\hat{o} = \{\hat{o}_j\}$, $\overline{o} = \{\overline{o}_{ji}\}$

   (a) Sample o: For $j = 1, \dots, K$ and $i = 1, \dots, K$, sample o_{ji} as follows: Set $o_{ji} = 0$. For $k = 1, \dots, n^s_{ji}$, draw $x_k \sim$ Bernoulli$(\frac{\alpha\Gamma_l + \kappa\delta(j,i)}{k-1+\alpha\Gamma_i+\kappa\delta(j,i)})$. If $x_k = 1$, increment o_{ji}.

   (b) Sampling $\hat{o}$: For $j = 1, \dots, K$, sample $\hat{o}_j \sim$ Binomial$(o_{jj}, \frac{\rho}{\rho+\Gamma_j(1-\rho)})$.

   (c) Update $\overline{o}$: For $j = 1, \dots, K$ and $i = 1, \dots, K$, set $\overline{o}_{ji} = o_{ji}$ if $j \neq i$; set $\overline{o}_{jj} = o_{jj} - \hat{o}_j$.

7. Sampling Γ: let $\overline{o}_{.i} = \sum_j \overline{o}_{ji}$ for $i = 1, \dots, K$

$$\Gamma \sim \text{Dirichlet}(\overline{o}_{.1}, \dots, \overline{o}_{.K}, \eta). \tag{33}$$

8. Sampling Π: For $k = 1, \dots, K$, sample

$$\Pi_k \sim \text{Dirichlet}(\alpha\Gamma_1 + n^s_{k1}, \dots, \alpha\Gamma_k + \kappa + n^s_{kk}, \dots, \alpha\Gamma_K + n^s_{kK}, \alpha\Gamma_{K+1}). \tag{34}$$

9. For a given $s_{1:T}$ and $z_{1:T}$ and exploiting conjugacy $\omega_{1:K}$ is sampled as

$$\omega_k \sim Dir\left(n^z_{k1} + \frac{\alpha_\omega}{L}, \dots, n^z_{kL} + \frac{\alpha_\omega}{L}\right),$$

   where n^z_{kl} and n^z_{kq} respectively denote the number of observations assigned to model l and q by state k according to $s_{1:T}$ and $z_{1:T}$. That is, $n^z_{kl} = \#\{t : s_t = k, z_t = l\}$.

10. Sampling hyperparameters η, α and κ: let $n_{k.} = \sum_i n^z_{ki}$, $o_{..} = \sum_j \sum_i o_{ji}$, $\hat{o}_{.} = \sum_j \hat{o}_j$, $\overline{o}_{..} = \sum_j \sum_i \overline{o}_{ji}$,

    (a) Sample $\alpha + \kappa$:

    i. For $k = 1, \dots, K$, draw $\overline{\xi}_k \sim$ Bernoulli$(\frac{n_{k.}}{n_{k.}+\alpha+\kappa})$.

    ii. For $k = 1, \dots, K$, draw $\widetilde{\xi}_k \sim$ Beta$(\alpha + \kappa + 1, n_{k.})$.

    iii. Sample $\alpha + \kappa \sim$ Gamma$(c_2 + o_{..} - \sum_{k=1}^K \overline{\xi}_k, c_3 - \sum_{j=1}^K \log\widetilde{\xi}_j)$.

(b) Sample ρ: Sample $\rho \sim \text{Beta}(c_4 + \hat{o}_{.}, c_5 + o_{..} - \hat{o}_{.})$.

(c) Sample η:

i. Draw $\widetilde{\tau} \sim \text{Bernoulli}(\frac{\overline{o}_{..}}{\overline{o}_{..}+\eta})$.

ii. Draw $\overline{\tau} \sim \text{Beta}(\eta + 1, \overline{o}_{..})$.

iii. Sample $\eta \sim \text{Gamma}(c_0 + \overline{K} - \widetilde{\tau}, c_1 - \log(\overline{\tau}))$, where $\overline{K} = \sum_{i=1}^{K} \mathbf{1}(\overline{o}_{.i} > 0)$.

11. Sampling α_ω:let $\alpha_\omega \sim Gamma(c_6, c_7)$, sample α_ω from the following density using a random walk Metropolis-Hasting,

$$p(\alpha_\omega|\omega, c_6, c_7) \propto \prod_{j=1}^{K} \Big(\frac{\Gamma(\alpha_\omega)}{\Gamma(\frac{\alpha_\omega}{L})^L} \omega_{j1}^{\frac{\alpha_\omega}{L}-1} \ldots \omega_{jL}^{\frac{\alpha_\omega}{L}-1} \Big) p(\alpha_\omega|c_6, c_7)$$

12. Repeat 2-11.