Bayesian Semiparametric Modeling of Realized Covariance Matrices

Xin Jin and John M Maheu

Shanghai University of Finance and Economics, McMaster University

November 2014

Bayesian Semiparametric Modeling of Realized Covariance Matrices*

Xin Jin† John M. Maheu‡

November 2014

Abstract

This paper introduces several new Bayesian nonparametric models suitable for capturing the unknown conditional distribution of realized covariance (RCOV) matrices. Existing dynamic Wishart models are extended to countably infinite mixture models of Wishart and inverse-Wishart distributions. In addition to mixture models with constant weights we propose models with time-varying weights to capture time dependence in the unknown distribution. Each of our models can be combined with returns to provide a coherent joint model of returns and RCOV. The extensive forecast results show the new models provide very significant improvements in density forecasts for RCOV and returns and competitive point forecasts of RCOV.

Key words: multi-period density forecasts, inverse-Wishart distribution, beam sampling, hierarchical Dirichlet process, infinite hidden Markov model
JEL: C58, C32, C11, C14

*We have benefited from helpful comments from Angelo Melino, Tom McCurdy, Mark Steel, Luc Bauwins, Silvia Frühwirth-Schnatter, Arnaud Dufays and Jim Griffin as well as participants of the European Seminar on Bayesian Econometrics 2014, and seminar participants at University of Toronto. Maheu is grateful to the SSHRC for financial support.

†Shanghai University of Finance and Economics, jin.xin@mail.shufe.edu.cn

‡DeGroote School of Business, McMaster University, 1280 Main Street West, Hamilton, ON, Canada, L8S4M4 and RCEA, Italy, maheujm@mcmaster.ca

1 Introduction

This paper introduces several new Bayesian nonparametric models suitable for capturing the unknown conditional distribution of realized covariance (RCOV) matrices. The nonparametric models extend existing dynamic Wishart specifications to countably infinite mixture models of Wishart and inverse-Wishart distributions. Mixture models with constant weights as well as models with time-varying weights are introduced.

Beginning with Andersen & Bollerslev (1998) there has been a great deal of interest in estimating and modeling daily ex post measures of volatility. Recent work has focused on realized measures of multivariate covariances estimated from high frequency intraday data. The theoretical foundation for RCOV as an estimate of the quadratic variation for semimartingale processes is set forth in Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004) while the latter also establishes the asymptotic theory for the estimator. Since then, focus has shifted to improving the estimator in the presence of market microstructure dynamics (Barndorff-Nielsen et al. 2011, Hautsch et al. 2012, Corsi et al. 2013) often through a kernel based estimator. This work provides accurate estimates of ex post covariation for asset returns. A developing area of research is how to econometrically model realized covariances.

This paper contributes to the literature on time-series modeling of RCOV. There are several existing approaches. One is to use some form of a decomposition to the matrix of realized variances and covariances and then use standard time-series models for the transformed data (Bauer & Vorkink 2011, Chiriac & Voev 2011). Another strand of the literature directly models RCOV using dynamic models. Examples of this approach include the multivariate high-frequency-based volatility (HEAVY) model of Noureldin et al. (2012) which exploits ex post volatility measures in a GARCH-like setting. Extensions to this approach are Sheppard & Xu (2014) and Janusa et al. (2014) and a closely related approach is Hansen et al. (2013).

Another approach has developed around time-varying Wishart distributions. Gourieroux et al. (2009), Bonato et al. (2008), Golosnoy et al. (2012), Jin & Maheu (2013) and Bauwens et al. (2014) develop dynamic Wishart and noncentral-Wishart models for RCOV. Building on Uhlig (1997), Windle & Carvalho (2014) provide a tractable state space model that can be used to model realized covariance matrices. Wishart distributions have also been popular in traditional multivariate stochastic volatility models that only use returns (Philipov & Glickman 2006, Asai & McAleer 2009, Fox & West 2011, Lin et al. 2012, Triantafyllopoulos 2012, Asai & So 2013).

Although some of the aforementioned papers use flexible distributions for RCOV distributions they are all essentially parametric. The purpose of this paper is to provide nonparametric models for capturing the unknown conditional distribution of RCOV. It is important to allow the unknown conditional distribution to change over time as a function of observables as well as latent variables. This allows for general forms of time dependence.

By far the most popular approach to Bayesian nonparametrics uses the Dirichlet process (DP) prior. The Dirichlet process mixture (DPM) model was popularized by Escobar & West (1995) and is useful for modeling a fixed unknown continuous distribution. This model is often embedded in a richer time-series model. Some examples in finance include Delatola & Griffin (2013), Jensen & Maheu (2013*a*), Virbickaite et al. (2013) and Kalli & Griffin

(2014).[1]

Therefore, a natural starting point for our work is a DPM model that mixes over distributions defined on positive definite matrices. We extend the dynamic Wishart model of Jin & Maheu (2013) to countably infinite mixtures based on Wishart and inverse-Wishart kernels. These models give significant gains in forecast precision compared to parametric benchmarks. However, it has been recognized that the DPM model may not be adequate for financial data which display complicated dependencies. Extensions to the DPM model, featuring nonparametric dependence in time or through a covariate are pursued in Griffin & Steel (2006), Griffin & Steel (2011), Rodriguez & Dunson (2011) and Jensen & Maheu (2013*b*). Dependent Dirichlet processes are formalized by MacEachern (2000).

Our first extension to incorporate time dependence in the mixture model is based on the infinite hidden Markov model (iHMM). The iHMM generalizes the popular finite state Markov switching model of Hamilton (1989) to an infinite number of states and is well-suited for capturing changes in the unknown conditional distribution of RCOV. The iHMM is introduced in Teh et al. (2006) and uses a hierarchical Dirichlet process (HDP). The HDP serves as a prior to link the rows of the infinite-dimension transition matrix of the hidden Markov chain. Extensions of the iHMM include the sticky model of Fox et al. (2011) which allows estimation and control of state persistence and the hierarchical prior governing the data density parameters by Song (2014). Other successful applications of the iHMM to economic and financial data are Dufays (2012), Jochmann (2014) and Shi & Song (2014).

The iHMM is a nonparametric model that allows the unknown distribution to flexibly change over time. The applications to date assume once a state is entered, the observation is governed by a parametric distribution. This may be appropriate for some forms of data (macroeconomic and low frequency financial data) but if the data do not conform to this assumption the model will require rapid switching among states to approximate the true density. This will expand the size of the latent state space and increase model complexity and computation time. To avoid this we introduce a new specification in which the state dependent data density is modelled nonparametrically. In this model each state of the Markov chain follows its own DPM model. Rather than having a potentially infinite number of DPM models to keep track we use a second HDP to reuse the existing atoms of support while allowing weights in each DPM to differ. This results in a very flexible approach that combines the benefits of the iHMM and the DPM model. That is, the model combines the benefits of Markov switching with i.i.d. switching and is self-organizing in the sense that the size of each mixture is endogenously determined given a finite dataset. Of course each of the simpler specifications are also nested in this new model. We show how the beam sampler of Van Gael et al. (2008) can be extended to estimate this model efficiently. The empirical results show the new model provides very significant gains in density forecasts for RCOV and returns and competitive point forecasts of RCOV. The parametric models fail to account for extreme observations in diagonal and off-diagonal elements of realized covariance matrices while the nonparametric models do significantly better.

Each of the models introduced in this paper can be combined with returns to produce a coherent joint model of returns and realized covariances. When combined with returns, and

[1]There are numerous applications in economics, e.g. Hirano (2002), Conley et al. (2008), Burda et al. (2008), Chib & Greenberg (2010) and Bassetti et al. (2014).

conditional on appropriate quantities, all of the mixture models based on inverse-Wishart kernels imply infinite mixtures of Student-t distributions with constant or time-varying weights.

Besides introducing several new nonparametric models of RCOV matrices this paper makes several additional contributions to the literature. We empirically estimate and compare multi-period density forecasts of RCOV matrices from parametric and nonparametric models. We find that mixtures of inverse-Wishart distributions are very promising for modeling realized covariance matrices and perform better than models based on the Wishart distribution.[2] Mixture models with constant weights and time-varying weights provide such large improvements in the fit that they render the parametric specifications (based in Wishart and inverse-Wishart) uncompetitive.

This paper is organized as follows. The data and estimation of realized covariances are discussed next. Following this parametric benchmark models for RCOV based on dynamic Wishart and inverse-Wishart distributions are reviewed. Section 4 introduces the Dirichlet process mixture models based on Wishart and inverse-Wishart dynamic models. Extensions to this model are discussed in Section 5, which include an infinite hidden Markov model, a sticky variant and a model that mixes iHMM and DPM behaviour. How to extend each of the models for RCOV to joint models of returns and RCOV is discussed in Section 6. Results are in Section 7 and conclusions in Section 8. A detailed appendix of all the posterior simulation steps is found in Section 9.

2 Data

The data is the same as that used in Jin & Maheu (2013). It consists of transactions obtained from the TAQ database for Standard and Poor's Depository Receipt (SPY), General Electric Co. (GE), Citigroup Inc.(C), Alcoa Inc. (AA) and Boeing Co. (BA). The sample period is from 1998/12/04 - 2007/12/31 giving 2281 days. We follow Barndorff-Nielsen et al. (2011) and their kernel based approach to construct daily realized covariance (RCOV) matrices. We compute daily returns and RCOV matrices based on close-to-close data. For more details see Jin & Maheu (2013). Daily RCOV matrices are denoted as Σ_t and daily returns as r_t, $t = 1, \ldots, T$, and summary statistics are reported in Table 1. Time-series plots of realized volatility and associated realized correlations are displayed in Figure 1.

3 Parametric Wishart Models

In this section we review a dynamic Wishart model for RCOV and introduce a similar version based on the inverse-Wishart distribution.

[2]One reason for this is that the second moments of a Wishart variate always exist while the second moments of an inverse-Wishart variate exist only if the degree of freedom is sufficiently large. This may allow greater tail thickness from the inverse-Wishart density. We thank Silvia Frühwirth-Schnatter for bringing this to our attention.

3.1 An additive component Wishart model: W-A(M)

The additive component Wishart model of Jin & Maheu (2013) is designed to capture the strong persistence in the elements of RCOV matrices by using several components that affect the scale matrix. Each component is a rolling window average of past values of Σ_t and the window size is estimated.

Consider a time series of $k \times k$ realized covariance matrices $\Sigma_t,\ t = 1, 2, \ldots, T$ and let $\Sigma_{1:t} = \{\Sigma_1, \ldots, \Sigma_t\}$. In the W-A(M) model the conditional distribution of Σ_t is defined as

$$
\begin{aligned}
f(\Sigma_t|\Sigma_{1:t-1}, \nu, \Theta) &= \text{Wishart}_k\left(\Sigma_t|\nu, \frac{1}{\nu}V_t\right) && (1)\\
V_t &= B_0 + \sum_{j=1}^{M} B_j \odot \Gamma_{t-1,\ell_j} && (2)\\
\Gamma_{t-1,\ell} &= \frac{1}{\ell}\sum_{i=1}^{\ell}\Sigma_{t-i} && (3)\\
B_j &= b_j b_j', \quad j = 1, \ldots, M && (4)\\
1 = \ell_1 < &\ \ldots < \ \ell_M. && (5)
\end{aligned}
$$

$\text{Wishart}_k(.|\nu, \frac{1}{\nu}V_t)$ denotes the density[3] of a Wishart distribution over positive definite matrices of dimension k with $\nu > k$ degrees of freedom and scale matrix $\frac{1}{\nu}V_t$. $\odot$ denotes the element-by-element (Hadamard) product of two matrices and Θ represents all parameters concerning the dynamics of V_t and includes $B_0, b_1, \ldots, b_M, \ell_2, \ldots, \ell_M$. B_0 is a $k \times k$ symmetric positive-definite matrix, and 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 is equal to Σ_{t-1} by construction with $\ell_1 = 1$. For component $j \geq 2$, rather than preset to either weekly or monthly term, each ℓ_j is allowed to be a free parameter to be estimated. The model specification ensures that $\frac{1}{\nu}V_t$ is symmetric positive-definite. The conditional mean of Σ_t is

$$
\text{E}(\Sigma_t|\Sigma_{1:t-1}, \nu, \Theta) = V_t = B_0 + \sum_{j=1}^{M} B_j \odot \Gamma_{t-1,\ell_j}. \tag{6}
$$

Instead of estimating B_0, Jin & Maheu (2013) implement RCOV targeting by setting $B_0 = (\iota\iota' - B_1 - \cdots - B_M) \odot \overline{\Sigma}_t$, where $\overline{\Sigma}_t$ is the sample mean of Σ_t and ι is a $k \times 1$ vector of ones. This ensures that the long-run mean of Σ_t is equal to $\overline{\Sigma}_t$ and leads to improved forecasts. In estimation any posterior draws in which B_0 is not positive definite are rejected. In addition, to ensure the mean exists draws that violate $\sum_{j=1}^{M} B_j < 1$ are rejected.

For posterior simulation a Metropolis-Hastings (MH) step using a joint random walk proposal is used for $b_1, \ldots, b_M$. For each lag length, ℓ_j is sampled according to a random walk with Poisson increments that are equally likely to be positive or negative. Additional

[3]The density function of a Wishart distribution for $k \times k$ symmetric positive-definite matrix Σ with ν degrees of freedom and positive-definite scale matrix V is $\text{Wishart}_k(\Sigma|\nu, V) = \frac{|\Sigma|^{\frac{\nu-k-1}{2}}|V|^{-\frac{\nu}{2}}}{2^{\frac{\nu k}{2}}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma(\frac{\nu+1-j}{2})}e^{-\frac{1}{2}\text{tr}(V^{-1}\Sigma)}$.

details of posterior simulation are found in Jin & Maheu (2013) along with stationarity conditions.

The W-A(M) model is a very competitive specification. In Jin & Maheu (2013) the model is extensively compared to models from Gourieroux et al. (2009) and Chiriac & Voev (2011), as well as extensions of Gourieroux et al. (2009) and Philipov & Glickman (2006) to handle RCOV matrices. The W-A(3) model provides superior point forecasts of RCOV matrices and when linked with returns in a joint model gives the best density forecasts of returns at multiple horizons.

3.2 An additive component inverse-Wishart model: IW-A(M)

In an similar fashion we can replace the Wishart density of the previous model with an inverse-Wishart to obtain the following IW-A(M) specification.

$$f(\Sigma_t|\Sigma_{1:t-1},\nu,\Theta) = \text{Wishart}_k^{-1}(\Sigma_t|\nu,(\nu-k-1)V_t) \tag{7}$$

$$V_t = B_0+\sum_{j=1}^{M} B_j\odot\Gamma_{t-1,\ell_j} \tag{8}$$

$$\Gamma_{t-1,\ell_j} = \frac{1}{\ell_j}\sum_{i=1}^{\ell_j}\Sigma_{t-i}. \tag{9}$$

$\text{Wishart}_k^{-1}(.|\nu,(\nu-k-1)V_t)$ denotes the density of an inverse-Wishart distribution over $k\times k$ symmetric positive-definite matrices with $\nu > k+1$ degrees of freedom and scale matrix equal to $(\nu-k-1)V_t$.[4]

By the properties of the inverse-Wishart distribution, the conditional mean of Σ_t is

$$\text{E}(\Sigma_t|\Sigma_{1:t-1},\nu,\Theta) = V_t = B_0+\sum_{j=1}^{M} B_j\odot\Gamma_{t-1,\ell_j}. \tag{10}$$

The conditional second moments are (Press 2005)

$$\text{Cov}(\Sigma_{t,ij},\Sigma_{t,lm}|\Sigma_{1:t-1},\nu,\Theta) = \frac{2V_{t,ij}V_{t,lm}+(\nu-k-1)(V_{t,il}V_{t,jm}+V_{t,im}V_{t,jl})}{(\nu-k)(\nu-k-1)^2(\nu-k-3)}, \tag{11}$$

which exist only if $\nu > k+3$.

Similar to the Wishart case quadratic transformations of inverse-Wishart distributed matrices are themselves inverse-Wishart distributed.[5]

Property 1 *Suppose A is $l\times k$ with $l\le k$ and has full row rank. If $\Sigma\sim$ Wishart$_k^{-1}(\nu,V)$, then $A\Sigma A'\sim$ Wishart$_l^{-1}(\nu-k+l,AVA')$.*

[4]The density function of an inverse-Wishart distribution for $k\times k$ symmetric positive-definite matrix Σ with ν degrees of freedom and positive-definite scale matrix V is $\text{Wishart}_k^{-1}(\Sigma|\nu,V) = \frac{|V|^{\frac{\nu}{2}}|\Sigma|^{-\frac{\nu+k+1}{2}}}{2^{\frac{\nu k}{2}}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma(\frac{\nu+1-j}{2})}e^{-\frac{1}{2}\text{tr}(V\Sigma^{-1})}$.

[5]See Press (2005).

Let w be a $k \times 1$ portfolio allocation vector. A direct implication of Property 1 is that under IW-A(M), the conditional distribution of the realized portfolio variance $RV_{w,t} \equiv w'\Sigma_t w$ follows a univariate inverse-Wishart:

$$RV_{w,t}|\Sigma_{1:t-1}, \nu, \Theta \sim \text{Wishart}_1^{-1}(\nu - k + 1, (\nu - k - 1)w'V_t w), \tag{12}$$

its density coincides with that of an inverse-Gamma distribution

$$\text{Gamma}^{-1}\left(\frac{\nu - k + 1}{2}, \frac{(\nu - k - 1)w'V_t w}{2}\right). \tag{13}$$

From this, the diagonal elements of Σ_t, which are realized variances of individual assets, follow the inverse-Gamma distribution

$$\Sigma_{t,ii}|\Sigma_{1:t-1}, \nu, \Theta, \sim \text{Gamma}^{-1}\left(\frac{\nu - k + 1}{2}, \frac{(\nu - k - 1)V_{t,ii}}{2}\right). \tag{14}$$

The IW-A(M) and W-A(M) models are parallel to each other in the sense their conditional expectations of Σ_t share the same dynamic structure in V_t. Any difference in the two models comes from the Wishart or inverse-Wishart assumption. On the other hand, because of the one-to-one correspondence between the parameter sets of the IW-A(M) and W-A(M), posterior sampling of IW-A(M) can be carried out in the same fashion as for the W-A(M) model.

4 Semiparametric RCOV Models

This paper will focus on mixture models with an infinite number of components. Before discussing these models we give some examples of the flexibility that finite mixtures have in which the density of Σ_t has the form

$$f(\Sigma_t) \quad = \quad \sum_{j=1}^{L} \omega_j \text{Wishart}_2^{-1}(\Sigma_t|\nu_j, V_j). \tag{15}$$

Applying the results of (14) to this mixture to focus on the diagonal elements of the 2×2 Σ_t results in a mixture of corresponding inverse-Gamma distributions. Figure 2 displays the densities from a two-component mixture along with a single component model, each of which has the same mean. It is clear that the mixture provides more flexibility including bimodal behaviour.

Figure 3 is a plot of the tail of the log-density from two different mixtures along with a single-component model. Each model has the same mean and variance. However, as the plot shows one mixture has a thinner tail and the other a fatter tail than the one-component model.

Finally, Figure 4 displays simulated data for the covariance (off diagonal) element of Σ_t from an inverse-Wishart and a mixture model. Each model has an identical mean and variance. In the top panel it is clear that the mixture has fatter tails as we see more extreme realizations. In the bottom panel the mixture appears to have thinner tails than

the parametric model. In summary, finite mixture models provide a great deal of flexibility in modeling the distribution of Σ_t. They can be used to capture multimodal behaviour and various tail structures. The following sections consider the fully nonparametric model with $L \to \infty$.

4.1 Dirichlet process mixture model

This paper will focus on semiparametric models based on the Dirichlet Process mixture (DPM) and various extensions. The first specification to consider takes the following form for the unknown density $f(\Sigma_t|\cdot)$,

$$f(\Sigma_t|\Sigma_{1:t-1}, \Theta, G) = \int h(\Sigma_t|\Sigma_{1:t-1}, \Theta, \phi)G(d\phi) \quad (16)$$

$$G|G_0, \alpha \sim \mathrm{DP}(\alpha, G_0) \quad (17)$$

where $\mathrm{DP}(\alpha, G_0)$ denotes the Dirichlet process with precision parameter $\alpha > 0$ and base measure G_0. G is the unknown mixing distribution that governs ϕ and is assumed to follow a Dirichlet process. G is centered around G_0 since $E[G] = G_0$. $h(\Sigma_t|\Sigma_{1:t-1}, \Theta, \phi)$ is a kernel density defined over symmetric positive-definite matrices given $\Sigma_{1:t-1}$ and parameters Θ and ϕ. Θ collects other parameters common to each conditional density $h(\cdot|\cdot)$.

Due to the Dirichlet process prior $\mathrm{DP}(G_0, \alpha)$, the random distribution G is almost surely discrete and the model is a countably-infinite mixture:

$$f(\Sigma_t|\Sigma_{1:t-1}, \Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \omega_j h(\Sigma_t|\Sigma_{1:t-1}, \Theta, \phi_j) \quad (18)$$

$$\omega_j = v_j \prod_{l<j}(1 - v_l), \quad v_j \overset{iid}{\sim} \mathrm{Beta}(1, \alpha), \quad j = 1, 2, \ldots \quad (19)$$

$$\phi_j \overset{iid}{\sim} G_0 \quad (20)$$

where $\Omega = \{\omega_j\}_{j=1}^{\infty}$, $\Phi = \{\phi_j\}_{j=1}^{\infty}$. (19) and (20) give the stick-breaking representation (Sethuraman 1994) of $G = \sum_{j=1}^{\infty} \omega_j \delta_{\phi_j}$, where δ_{ϕ_j} is a point mass at ϕ_j, the random atoms ϕ_j are i.i.d. draws from prior distribution G_0, and the random weights ω_j are constructed using i.i.d. Beta variates v_j. In the following the stick-breaking construction of the weights are denoted as $\Omega \sim \mathbf{SBP}(\alpha)$.

To implement the model we need to select a parametric kernel density $h(.|.)$ defined over symmetric positive-definite matrices. Given the discussion on the previous models it is natural to consider the kernel as one of the models in Section 3.[6] Those specifications were designed to capture important features of the time-series properties of RCOV matrices. Mixing over them (IW-A(M) or W-A(M)) will allow for more general distributional shapes for the conditional density of Σ_t.

[6] Another choice is the noncentral Wishart density which has no closed form and can only be approximated or computed recursively, making the computation formidable. Meanwhile, the empirical results in Jin & Maheu (2013) and Chiriac & Voev (2011) show inferior results for WAR compared to other models. Thus, we exclude it from our choice of kernels.

Although several papers referenced in the introduction use Wishart distributions to model the conditional distribution of RCOV we are not aware of any papers that use the inverse-Wishart distribution. In the empirical work we found that the inverse-Wishart model dominated the Wishart counterpart and therefore we focus on it. However, all of the following models could use any kernel that is defined for positive definite matrices.

Extending the IW-A(M) to the semiparametric specification, IW-DPM, we have

$$f(\Sigma_t|\Sigma_{1:t-1},\Theta,\Omega,\Phi) = \sum_{j=1}^{\infty}\omega_j \text{Wishart}_k^{-1}(\Sigma_t|\nu_j,(\nu_j-k-1)V_t^{1/2}A_j(V_t^{1/2})') \quad (21)$$

$$V_t = B_0+\sum_{j=1}^{M}B_j\odot\Gamma_{t-1,\ell_j} \quad (22)$$

$$\Gamma_{t-1,\ell_j} = \frac{1}{\ell_j}\sum_{i=1}^{\ell_j}\Sigma_{t-i} \quad (23)$$

$$B_j = b_jb_j', \quad j=1,\ldots,M, \qquad 1=\ell_1<\ldots<\ell_M, \quad (24)$$

where the evolution of V_t is identical to the parametric model, $\Omega\sim\mathbf{SBP}(\alpha)$ and A_j are $k\times k$ symmetric positive matrices and $\phi_j\equiv(\nu_j,A_j)$. $V_t^{1/2}$ denotes the Cholesky factor of V_t. Each component of the distribution j allows for a different scale matrix, $(\nu_j-k-1)V_t^{1/2}A_j(V_t^{1/2})'$, which by construction is positive definite, and a different degree of freedom ν_j. The term $V_t^{1/2}A_j(V_t^{1/2})'$ can represent any symmetric positive definite matrix. This is a richer functional form than the parametric model. For instance, the conditional mean is a weighted average of the component means,

$$E[\Sigma_t|\Sigma_{1:t-1},\Theta,\Omega,\Phi]=\sum_{j=1}^{\infty}\omega_jV_t^{1/2}A_j(V_t^{1/2})'. \quad (25)$$

Note that in this model, the parametric version previously discussed, is nested. For instance, if $\omega_j=1$ and $A_j=I$ we have the IW-A(M) model exactly, while if only $A_j=I\ \forall j$, we have an identical conditional mean in (25) but different higher order moments.

The analogous model with a Wishart kernel replaces (21) with

$$f(\Sigma_t|\Sigma_{1:t-1},\Theta,\Omega,\Phi) = \sum_{j=1}^{\infty}\omega_j\text{Wishart}_k\left(\Sigma_t|\nu_j,\frac{1}{\nu_j}V_t^{1/2}A_j(V_t^{1/2})'\right) \quad (26)$$

The definition of V_t and other portions of the model remain the same.

To complete the DPM models, the prior distribution G_0 for the random atoms ϕ_j are defined for IW-DPM as:

$$G_0(\nu_j,A_j) \equiv \text{Exp}_{\nu>k+1}(\lambda)\times\text{Wishart}_k\left(\gamma_0,\frac{1}{\gamma_0}I\right), \quad \gamma_0\geq k \quad (27)$$

and for W-DPM as:

$$G_0(\nu_j,A_j) \equiv \text{Exp}_{\nu>k}(\lambda)\times\text{Wishart}_k^{-1}(\gamma_0,(\gamma_0-k-1)I), \quad \gamma_0\geq k+1; \quad (28)$$

Under G_0, ν_j and A_j are independently drawn from a truncated exponential distribution and a Wishart (inverse-Wishart) distribution, respectively, such that the mean of A_j satisfies $\mathrm{E}(A_j) = I$. In other words, the nonparametric model has a prior that centers the conditional mean of Σ_t to that of the parametric model.

The precision parameter α controls the distribution of the mixture weights ω_j. We include α in the posterior inference with the following prior,

$$\alpha \sim \mathrm{Gamma}(a_0, c_0). \tag{29}$$

4.2 Posterior inference

To sample from the posterior we use slice sampling techniques introduced by Walker (2007) and extended by Kalli et al. (2011) and Papaspiliopoulos (2008). This samples from the stick-breaking representation of the infinite mixture model by introducing a slice variable that randomly truncates the model to a finite mixture model. This is done in such a way that integrating out the slice variable gives the correct marginal distribution.

Recall that $\phi_j = (\nu_j, A_j)$ and in the following conditioning on $\Sigma_{1:t-1}$ is suppressed where the context is clear. The general model is

$$f(\Sigma_t|\Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \omega_j h(\Sigma_t|\Theta, \nu_j, A_j), \tag{30}$$

where $h(\Sigma_t|\Theta, \nu_j, A_j)$ corresponds to either the inverse-Wishart in (21) or Wishart kernel in (26). Introducing an auxiliary latent variable $u_t > 0$, we define the joint conditional density of Σ_t and u_t as

$$f(\Sigma_t, u_t|\Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \mathbf{1}(u_t < \omega_j) h(\Sigma_t|\Theta, \nu_j, A_j). \tag{31}$$

Note that integrating out u_t returns the original model (30). The parameter space is augmented with $u_{1:T} = \{u_1, \ldots, u_T\}$. Let $s_t = j$ assign observation Σ_t to component j with data density $h(\Sigma_t|\Theta, \nu_j, A_j)$ The target likelihood is now

$$\begin{aligned} f(\Sigma_{1:T}, u_{1:T}, s_{1:T}|\Theta, \Omega, \Phi) &= \prod_{t=1}^{T} f(\Sigma_t, u_t, s_t|\Theta, \Omega, \Phi) \\ &= \prod_{t=1}^{T} \mathbf{1}(u_t < \omega_{s_t}) h(\Sigma_t|\Theta, \nu_{s_t}, A_{s_t}), \end{aligned} \tag{32}$$

where $s_{1:T} = \{s_t\}_{t=1}^{T}$. The joint posterior is proportional to

$$p(\Theta) p(\Omega_{\overline{K}}) \left[\prod_{i=1}^{\overline{K}} p(\nu_j, A_j) \right] \prod_{t=1}^{T} \mathbf{1}(u_t < \omega_{s_t}) h(\Sigma_t|\Theta, \nu_{s_t}, A_{s_t}), \tag{33}$$

where $\Omega_{\overline{K}} = \{\omega_j\}_{j=1}^{\overline{K}}$ and $\overline{K}$ is the smallest natural number such that $\sum_{j=1}^{\overline{K}} \omega_j > 1 - \min\{u_t\}$.

The posterior sampling steps are as follows.

1. $p(\phi_j|\Sigma_{1:T}, s_{1:T}, \Theta) \propto p(\phi_j) \prod_{\{t:s_t=j\}} h(\Sigma_t|\Theta, \nu_j, A_j), j = 1, \ldots, \overline{K}$.

2. $p(v_j|s_{1:T}, \alpha) \propto \text{Beta}(v_j|a_{1,j}, a_{2,j}), j = 1, \ldots, \overline{K}$, with $a_{1,j} = 1 + \sum_{t=1}^{T} \mathbf{1}(s_t = j)$ and $a_{2,j} = \alpha + \sum_{t=1}^{T} \mathbf{1}(s_t > j)$, where Beta(.|.,.) denotes the density of a Beta distribution.

3. $p(u_t|\Omega_{\overline{K}}, s_{1:T}) \propto \mathbf{1}(0 < u_t < \omega_{s_t}), t = 1, \ldots, T$.

4. Find the smallest $\overline{K}$ such that $\sum_{j=1}^{\overline{K}} \omega_j > 1 - \min\{u_t\}$.

5. $P(s_t = j|\Sigma_{1:T}, \Phi, \Omega_{\overline{K}}, \Theta, u_{1:T}) \propto \mathbf{1}(u_t < \omega_j) h(\Sigma_t|\Theta, \nu_j, A_j)$.

6. $p(\alpha|K) \propto p(\alpha)p(K|\alpha)$, where K is the number of active clusters in $s_{1:T}$.

7. $p(\Theta|\Sigma_{1:T}, s_{1:T}, \Phi) \propto p(\Theta) \prod_{t=1}^{T} h(\Sigma_t|\Theta, \nu_{s_t}, A_{s_t})$

One sweep of the sampler delivers $\{\{(\nu_j, A_j, v_j)\}_{j=1}^{\overline{K}}, \overline{K}, u_{1:T}, s_{1:T}, \alpha, \Theta\}$. In Step 1, the conditional posterior of A_j is

$$p(A_j|\nu_j, \Sigma_{1:T}, s_{1:T}, \Theta) \propto p(A_j) \prod_{\{t:s_t=j\}} h(\Sigma_t|\Theta, \nu_j, A_j). \tag{34}$$

By conjugacy, we have for IW-DPM model

$$A_j \sim \text{Wishart}_k(\overline{\gamma}_j, \overline{Q}_j), \tag{35}$$

where $\overline{\gamma}_j = \gamma_0 + n_j\nu_j$ and $\overline{Q}_j = \left[(\nu_j - k - 1) \sum_{\{t:s_t=j\}} \left[(V_t^{1/2})\Sigma_t^{-1}((V_t^{1/2}))'\right] + \gamma_0 I\right]^{-1}$, with $n_j = \#\{t : s_t = j\}$. And for the W-DPM model

$$A_j \sim \text{Wishart}_k^{-1}(\overline{\gamma}_j, \overline{Q}_j), \tag{36}$$

where $\overline{\gamma}_j$ defined as before but $\overline{Q}_j = \nu_j \sum_{\{t:s_t=j\}} \left[(V_t^{1/2})^{-1}\Sigma_t((V_t^{1/2})^{-1})'\right] + (\gamma_0 - k - 1)I$. The conditional posterior of ν_j is

$$p(\nu_j|A_j, \Sigma_{1:T}, s_{1:T}, \Theta) \propto p(\nu_j) \prod_{\{t:s_t=j\}} h(\Sigma_t|\Theta, \nu_j, A_j). \tag{37}$$

Metropolis-Hastings (MH) steps are used to sample ν_j with Gaussian random walk proposals. In Step 4, additional ω_j and ϕ_j will need to be simulated from the prior if $\overline{K}$ is incremented. Step 6 follows Escobar & West (1995) and consists of first sampling an auxiliary variable η from Beta$(\alpha + 1, T)$, and then sampling α from a two-component mixture of Gamma distributions,

$$\alpha \sim p_\eta \text{Gamma}(a_0 + K, c_0 - \log\eta) + (1 - p_\eta)\text{Gamma}(a_0 + K - 1, c_0 - \log\eta), \tag{38}$$

where $p_\eta/(1 - p_\eta) = (a_0 + K - 1)/(T(c_0 - \log\eta))$. In Step 7, MH steps are used to sample elements of b_j's and ℓ_j. As in the benchmark models, we impose the same restriction associated with RCOV targeting in the nonparametric models. That is, we set

$B_0 = (\iota\iota' - B_1 - \cdots - B_M) \odot \overline{\Sigma}_t$ in estimation and reject any draws in which B_0 is not positive definite. This leads to significant improvements in forecasts.

After dropping a suitable number of draws as burn-in we collect the next N draws to be used for posterior inference. Each iteration of the posterior sampler delivers a draw of the unknown distribution G where

$$G^{(i)} = \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} \delta_{\phi_j^{(i)}} + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) G_0. \tag{39}$$

This can be used to form the predictive density of Σ_{T+1} which is discussed next.

4.3 Predictive density

In Bayesian nonparametrics interest focuses on the predictive density. This can be computed as follows. Given a draw $G^{(i)}$ from the posterior then

$$\begin{aligned} & p(\Sigma_{T+1}|\Sigma_{1:T}, G^{(i)}) \\ = & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \int h(\Sigma_{T+1}|\Theta^{(i)}, \phi) G_0(d\phi) \qquad (40) \\ \approx & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \frac{1}{R} \sum_{l=1}^{R} h(\Sigma_{T+1}|\Theta^{(i)}, \phi^{[l]}), \qquad (41) \end{aligned}$$

where $\phi^{[l]} \overset{iid}{\sim} G_0, l = 1, \ldots, R$. In the empirical work $R = 10$ but smaller values gave similar accuracy.[7] Finally, the predictive density with all parameter uncertainty integrated out is estimated as

$$p(\Sigma_{T+1}|\Sigma_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} p(\Sigma_{T+1}|\Sigma_{1:T}, G^{(i)}). \tag{42}$$

5 Extensions

If there are features of the unknown conditional distribution of Σ_t that change over time and cannot be captured through observables, such as V_t, then the DPM models cannot capture these. We extend the DPM specifications to have time-varying weights to allow for time variation in the conditional distribution.

5.1 Infinite hidden Markov models

The infinite hidden Markov model (iHMM) builds on a hierarchical Dirichlet process prior (HDP) of Teh et al. (2006).[8] They show that a sequence of draws from a Dirichlet process,

[7] An asymptotically equivalent, but potentially less accurate estimate in finite simulations, would be to randomly draw ϕ from each sampled $G^{(i)}$ in (39) and then average $h(\cdot|\cdot)$ over these draws.

[8] A related but different approach to allow dependence through a hierarchical structure is the nested DP of Rodriguez et al. (2008).

with a base measure that itself is a draw from a DP, can be used as a prior for the rows of the transition matrix of an infinite Markov chain. The iHMM is also reviewed in Van Gael & Ghahramani (2010). We propose the following iHMM model:

$$\begin{aligned}
\boldsymbol{\pi}_0|\alpha &\sim \mathbf{SBP}(\alpha) && (43)\\
\boldsymbol{\pi}_i|\boldsymbol{\pi}_0,\beta &\sim \mathrm{DP}(\beta,\boldsymbol{\pi}_0) && (44)\\
\phi_j &\overset{iid}{\sim} G_0, \quad j=1,2,\ldots && (45)\\
s_t|s_{t-1}=i,\Pi &\sim \boldsymbol{\pi}_i, \quad i=1,2,\ldots && (46)\\
\Sigma_t|\Sigma_{1:t-1},\Theta,\Phi,s_t &\sim \mathcal{H}_t(\Sigma_t|\phi_{s_t}) && (47)
\end{aligned}$$

The latent discrete state variable s_t follows a Markov chain on an infinite state space with doubly-infinite transition matrix $\Pi = (\boldsymbol{\pi}_1', \boldsymbol{\pi}_2', \ldots)'$, where $\boldsymbol{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \ldots)$ is the i^{th} row of Π. That is, $p(s_t = j|s_{t-1} = i) = \pi_{i,j}$, $i, j \in \{1, 2, \ldots\}$. $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, \ldots)$ denotes an infinite-dimensional vector of probability weights which are drawn from a stick-breaking process $\mathbf{SBP}(\alpha)$. Conditional on $\boldsymbol{\pi}_0$ and scalar β, $\boldsymbol{\pi}_i$ independently draws from the common Dirichlet process prior $\mathrm{DP}(\beta, \boldsymbol{\pi}_0)$ for $i = 1, 2, \ldots$. $\mathcal{H}_t(\Sigma_t|\phi_{s_t})$ would be either of the inverse-Wishart model with density $\mathrm{Wishart}_k^{-1}(\Sigma_t|\nu_{s_t}, (\nu_{s_t} - k - 1)V_t^{1/2}A_{s_t}(V_t^{1/2})')$ or the Wishart analogue. These models are labelled IW-iHMM and W-iHMM.

There is a similar stick-breaking representation of the model with weights ω_j in (21) replaced with π_{s_{t-1},s_t} as

$$\begin{aligned}
f(\Sigma_t|\Theta,\Pi,\Phi,s_{t-1}) &= \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \mathrm{Wishart}_k^{-1}(\Sigma_t|\nu_{s_t}, (\nu_{s_t} - k - 1)V_t^{1/2}A_{s_t}(V_t^{1/2})') && (48)\\
\pi_{i,j} &= \hat{\pi}_{i,j}\prod_{l=1}^{j-1}(1-\hat{\pi}_{i,l}) && (49)\\
\hat{\pi}_{i,j} &\overset{iid}{\sim} \mathrm{Beta}(\beta\pi_{0j}, \beta(1-\sum_{l=1}^{j}\pi_{0l})) && (50)
\end{aligned}$$

The definition of the weights (Van Gael & Ghahramani 2010) follows from the properties of the Dirichlet process and Dirichlet distribution and links the transition matrix Π to $\boldsymbol{\pi}_0$. From this we have $E[\pi_{i,j}] = E[\pi_{0j}] = \alpha^{j-1}/(1+\alpha)^j$. In other words, the prior centers the infinite hidden Markov model around the DPM model discussed in the last section. The parameters α and β play an important role in the distribution of the weights π_{s_{t-1},s_t}, and can be used to set various prior beliefs. We impose the following priors to learn about these parameters,

$$\alpha \sim \mathrm{Gamma}(a_1, c_1), \qquad \beta \sim \mathrm{Gamma}(a_2, c_2). \tag{51}$$

5.2 IW-sticky-iHMM

The original iHMM model does not differentiate between self-transitions and moves into different states since each $\boldsymbol{\pi}_i$ draws from the same Dirichlet prior in (44). The IW-iHMM may not capture state persistence commonly present in economic time-series data. To solve

this issue, a "sticky" version of iHMM was introduced by Fox et al. (2011) in which the prior can reinforce self-transitions. This is done by replacing (44) in the IW-iHMM model with

$$\boldsymbol{\pi}_i|\boldsymbol{\pi}_0, \beta, \kappa \sim \text{DP}\left(\beta+\kappa, \frac{\beta\boldsymbol{\pi}_0 + \kappa\delta_i}{\beta+\kappa}\right). \tag{44'}$$

The term $\beta\boldsymbol{\pi}_0 + \kappa\delta_i$ means that the amount $\kappa \geq 0$ is added to the i^{th} component of $\beta\boldsymbol{\pi}_0$. A $\kappa > 0$ increases the prior probability of self-transition and larger values impose stronger beliefs on state persistence while $\kappa = 0$ gives the benchmark iHMM specification above. The stick-breaking formulation of the weights replaces $\hat{\pi}_{i,j}$ terms in (50) with

$$\hat{\pi}_{i,j} \overset{iid}{\sim} \text{Beta}(\beta\pi_{0j} + \kappa\delta_i, \beta + \kappa - \textstyle\sum_{l=1}^{j}(\beta\pi_{0l} + \kappa\delta_i)). \tag{52}$$

Rather than setting the parameters we impose the following priors,

$$\alpha \sim \text{Gamma}(a_3, c_3), \qquad \beta+\kappa \sim \text{Gamma}(a_4, c_4), \qquad \rho = \frac{\kappa}{\beta+\kappa} \sim \text{Beta}(a_5, c_5), \tag{53}$$

which allow for learning from the data. This prior formulation is more convenient for posterior sampling. These changes give the IW-sticky-iHMM model.

5.2.1 Posterior inference

Similar to the posterior sampling methods for the DPM model of Section 4 the idea of slice sampling can be extended to the infinite hidden Markov model. Beam sampling introduced by Van Gael et al. (2008) combines slice sampling and dynamic programming. The slice sampling portion stochastically truncates the infinite dimension state space into a finite one. With a finite state space, traditional posterior sampling methods can be applied such as the forward filtering backward sampling (FFBS) of Chib (1996). This allows for the efficient sampling of the state variables as one block.

An auxiliary latent variable $u_t > 0$ is introduced such that its conditional density is

$$p(u_t|s_t, s_{t-1}, \Pi) = \frac{\mathbf{1}(u_t < \pi_{s_{t-1},s_t})}{\pi_{s_{t-1},s_t}} \tag{54}$$

and is sampled with the other model parameters. With this slice variable, Van Gael et al. (2008) show that the filtering step of the sampler becomes

$$\begin{aligned}
p(s_t|u_{1:t}, \Sigma_{1:t}) &\propto h(\Sigma_t|\phi_{s_t}) \sum_{s_{t-1}=1}^{\infty} p(u_t|s_t, s_{t-1}) p(s_t|s_{t-1}) p(s_{t-1}|\Sigma_{1:t-1}, u_{1:t-1}) && (55)\\
&\propto h(\Sigma_t|\phi_{s_t}) \sum_{s_{t-1}=1}^{\infty} \mathbf{1}(u_t < \pi_{s_{t-1},s_t}) p(s_{t-1}|u_{1:t-1}, \Sigma_{1:t-1}) && (56)\\
&\propto h(\Sigma_t|\phi_{s_t}) \sum_{s_{t-1}: u_t < \pi_{s_{t-1},s_t}} p(s_{t-1}|u_{1:t-1}, \Sigma_{1:t-1}). && (57)
\end{aligned}$$

Thus the infinite summation in this filter is reduced to a finite summation since the set $\{s_{t-1} : u_t < \pi_{s_{t-1},s_t}\}$ is finite. The backward sampling step follows

$$p(s_t|s_{t+1}, \Sigma_{1:T}, u_{1:T}) \propto p(s_t|u_{1:t}, \Sigma_{1:t}) \mathbf{1}(u_{t+1} < \pi_{s_t,s_{t+1}}). \tag{58}$$

s_T is sampled from the last step of the filter $p(s_T|u_{1:T}, \Sigma_{1:T})$ after which s_t, $t = T-1, \ldots, 1$ is sampled from (58).

It is convenient to find a finite set that includes all possible states that satisfy the condition $u_t < \pi_{s_{t-1},s_t}$. This must hold for each t and each row of the transition matrix. States that do not satisfy this condition can be ignored. We require $\overline{K}$ states to be kept track of such that the remaining states do not satisfy the condition, that is the $\overline{K}$ such that $\sum_{j=\overline{K}+1}^{\infty} \pi_{i,j} < u_t$ holds for each i and each t. This gives the following condition, $\max_{i\in\{1,\ldots,\overline{K}\}}\{1 - \sum_{j=1}^{\overline{K}} \pi_{i,j}\} < \min_{t\in\{1,\ldots,T\}}\{u_t\}$, to select $\overline{K}$.

After the states are sampled we keep track of the number of *alive* states in which at least one observation is allocated to the state. These are ordered as the first K states. Each sweep of the sampler updates the value of K.

The parameter set consists of $\{u_{1:T}, s_{1:T}, \boldsymbol{\pi}_0, \Pi, \Phi, \Theta, \alpha, \beta, \kappa\}$. In posterior sampling we keep track of $K+1$ rows for Π and $K+1$ elements of $\boldsymbol{\pi}_0$. The first K rows of Π represent the *alive* states while the $K+1$ row is the residual probability. For other parameters such as Φ we sample only the K values associated with *alive* states.

The sampling procedure sequentially simulates from the following conditional posterior densities:

1. $p(u_{1:T}|s_{1:T}, \Pi)$,
2. $p(s_{1:T}|\Pi, u_{1:T}, \Phi, \Theta, \Sigma_{1:T})$,
3. $p(\boldsymbol{\pi}_0|s_{1:T}, \alpha, \beta, \kappa)$,
4. $p(\Pi|\boldsymbol{\pi}_0, s_{1:T}, \beta, \kappa)$,
5. $p(\Phi|s_{1:T}, \Theta, \Sigma_{1:T})$,
6. $p(\alpha, \beta, \kappa|s_{1:T}, \boldsymbol{\pi}_0)$,
7. $p(\Theta|s_{1:T}, \Phi, \Sigma_{1:T})$.

The Appendix 9.1 provides full details on each of the steps. For the (non-sticky) IW-iHMM model discussed in the previous subsection the above sampling steps are used with $\kappa = 0$ and irrelevant steps are omitted.

5.2.2 Predictive density

The predictive density is computed in the following way. Given a draw from the posterior,

$$
\begin{aligned}
& p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, \Phi^{(i)}, s_{1:T}^{(i)}, \Theta^{(i)}) \\
= & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s_T^{(i)},j} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s_T^{(i)},j}\right) \int h(\Sigma_{T+1}|\Theta^{(i)}, \phi) G_0(d\phi) \qquad (59) \\
\approx & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s_T^{(i)},j} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s_T^{(i)},j}\right) \frac{1}{R} \sum_{l=1}^{R} h(\Sigma_{T+1}|\Theta^{(i)}, \phi^{[l]}), \qquad (60)
\end{aligned}
$$

where $\phi^{[l]} \overset{iid}{\sim} G_0, l = 1, \ldots, R$. Finally, the predictive density is estimated as

$$p(\Sigma_{T+1}|\Sigma_{1:T}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, \Phi^{(i)}, s_{1:T}^{(i)}, \Theta^{(i)}), \tag{61}$$

which integrates out all uncertainty.

5.3 IW-sticky-iHMM-HDP

A potential drawback of the IW-sticky-iHMM model is that it allows for persistence in states that have a fixed parametric density. However, if the conditional density of Σ_t given s_t is not close to the inverse-Wishart then rapid mixing among other states may be necessary to approximate it. This loses the interpretation of state persistence. In this section we propose a new model that allows each conditional density of Σ_t given s_t to be nonparametrically modelled as a DPM model. A related model is discussed in Fox et al. (2011). Our version employs a hierarchical DP prior that links each of the DPM models in each state, in addition to a separate HDP that governs the rows of the transition matrix as before. This second HDP improves posterior sampling efficiency and exploits parameters already in use by pooling.

This model is the following

$$\begin{aligned}
\boldsymbol{\pi}_0|\alpha &\sim \mathbf{SBP}(\alpha), && (62)\\
\boldsymbol{\pi}_i|\boldsymbol{\pi}_0, \beta, \kappa &\sim \mathrm{DP}\left(\beta+\kappa, \frac{\beta\boldsymbol{\pi}_0 + \kappa\delta_i}{\beta+\kappa}\right), && (63)\\
s_t|s_{t-1}=i, \Pi &\sim \boldsymbol{\pi}_i, \quad i = 1, 2, \ldots, && (64)\\
\phi_j &\overset{iid}{\sim} G_0, \quad j = 1, 2, \ldots, && (65)\\
\boldsymbol{\psi}_0|\alpha_\psi &\sim \mathbf{SBP}(\alpha_\psi), && (66)\\
\boldsymbol{\psi}_i|\boldsymbol{\psi}_0, \beta_\psi &\sim \mathrm{DP}(\beta_\psi, \boldsymbol{\psi}_0), && (67)\\
z_t|s_t=i, \Psi &\sim \boldsymbol{\psi}_i, \quad i = 1, 2, \ldots, && (68)\\
\Sigma_t|\Sigma_{1:t-1}, \Theta, \Phi, z_t &\sim \mathcal{H}_t(\Sigma_t|\phi_{z_t}). && (69)
\end{aligned}$$

z_t is a discrete variable taking on natural numbers indexing the component (parameter) assigned to observation t. $\boldsymbol{\psi}_i = (\psi_{i,1}, \psi_{i,2}, \ldots)$ is the state-specific discrete probability measure for state i, $i = 1, 2, \ldots$, and $\Psi = \{\boldsymbol{\psi}_i\}_{i=1}^{\infty}$. $\boldsymbol{\psi}_0 = (\psi_{01}, \psi_{02}, \ldots)$ draws from a stick-breaking process $\mathbf{SBP}(\alpha_\psi)$. Conditional on $\boldsymbol{\psi}_0$ and scalar β_ψ, all $\boldsymbol{\psi}_i$s are independently drawn from the common Dirichlet process prior $\mathrm{DP}(\beta_\psi, \boldsymbol{\psi}_0)$. Note that each of the DPM models, indexed by s_t, shares the same points of support Φ but has different weights. The weights have a common DP prior.

Using the inverse-Wishart distribution for $\mathcal{H}_t(\Sigma_t|\phi_{z_t})$ we have the IW-sticky-iHMM-HDP model. This model consists of two hierarchical Dirichlet processes. The first one includes (62)-(63), which defines the prior for the transition probabilities of the infinite hidden Markov chain. The second HDP corresponds to (65)-(67), which defines the prior for the set of state-specific countably-infinite mixture distributions $G_i = \sum_{j=1}^{\infty} \psi_{i,j}\delta_{\phi_j}$, $i = 1, 2, \ldots$.

There is a stick-breaking representation for this model. To distinguish between the two hierarchical Dirichlet processes used, conditional on s_t, we have

$$f(\Sigma_t|\Theta,\Phi,\Psi,s_t)=\sum_{z_t=1}^{\infty}\psi_{s_t,z_t}\text{Wishart}_k^{-1}(\Sigma_t|\nu_{z_t},(\nu_{z_t}-k-1)V_t^{1/2}A_{z_t}(V_t^{1/2})'). \tag{70}$$

This is a standard DPM model as discussed in Section 4. Changing s_t only changes the weights ψ_{s_t,z_t} while the mixture has the same points of support Φ. Now, mixing over states as well, conditional on s_{t-1}, gives

$$f(\Sigma_t|\Theta,\Pi,\Phi,\Psi,s_{t-1})=\sum_{s_t=1}^{\infty}\pi_{s_{t-1},s_t}\sum_{z_t=1}^{\infty}\psi_{s_t,z_t}\text{Wishart}_k^{-1}(\Sigma_t|\nu_{z_t},(\nu_{z_t}-k-1)V_t^{1/2}A_{z_t}(V_t^{1/2})'). \tag{71}$$

An important feature of this model is that it is possible to persist in a state s_t for many periods but have the parameters affecting the data density, ν_{z_t} and A_{z_t}, change. This is due to each state having a DPM model with different weights mixing over the common set of parameters Φ.

This model nests both the DPM model and the iHMM model. What this means in practice is that we can separate out the states that mix in an i.i.d fashion from the states that mix with persistence (Markov chain).

The IW-sticky-iHMM-HDP model is completed with the following priors $\alpha\sim\text{Gamma}(a_6,c_6)$, $\beta+\kappa\sim\text{Gamma}(a_7,c_7)$, $\rho\sim\text{Beta}(a_8,c_8)$, $\alpha_\psi\sim\text{Gamma}(a_9,c_9)$ and $\beta_\psi\sim\text{Gamma}(a_{10},c_{10})$.

5.3.1 Posterior inference

From the stick-breaking representation of the model in (71) we can modify the previous beam sampler to provide a valid sampler that stochastically truncates the state space to a finite one in which the FFBS can be applied. The main difference is that now the probability weights are $\pi_{s_{t-1},s_t}\psi_{s_t,z_t}$ and to truncate the state space we need to consider two dimensions, that of s_t and z_t. The auxiliary latent variable $u_t>0$ is introduced such that its conditional density is

$$p(u_t|s_t,s_{t-1},z_t,\Pi,\Psi)=\frac{\mathbf{1}(u_t<\pi_{s_{t-1},s_t}\psi_{s_t,z_t})}{\pi_{s_{t-1},s_t}\psi_{s_t,z_t}}. \tag{72}$$

We seek to find truncation variables $\overline{K}$ and $\overline{K}_Z$ such that the set $\{(s_t,z_t)|s_t\leq\overline{K},\,z_t\leq\overline{K}_Z\}$ contains all instances of $u_t<\pi_{s_{t-1},s_t}\psi_{s_t,z_t}$ for each t. First, find $\overline{K}$ such that $\max_{j\in\{1,\ldots,\overline{K}\}}\{1-\sum_{l=1}^{\overline{K}}\pi_{j,l}\}<\min_t\{u_t\}$. Given $\overline{K}$, we have $u_t>\pi_{s_{t-1},s_t}\psi_{s_t,z_t}$ for all $s_t>\overline{K}$ for any value of z_t. This holds since $\psi_{s_t,z_t}\leq 1$. Then, find $\overline{K}_Z$ such that $\max_{j\in\{1,\ldots,\overline{K}\}}\{1-\sum_{l=1}^{\overline{K}_Z}\psi_{j,l}\}<\min_t\{u_t\}$. Note that $u_t>\pi_{s_{t-1},s_t}\psi_{s_t,z_t}$ for any $z_t>\overline{K}_Z$ and $s_t\leq\overline{K}$ since $\pi_{s_{t-1},s_t}\leq 1$. Therefore, any pair (s_t,z_t) that satisfies $u_t<\pi_{s_{t-1},s_t}\psi_{s_t,z_t}$ will also satisfy $s_t\leq\overline{K},z_t\leq\overline{K}_Z$. With this, the double summation in (71) is truncated at $\overline{K}$ and $\overline{K}_Z$ and the state variables s_t and z_t can be sampled jointly with the FFBS as in the previous iHMM models. The effective dimension of the state space in the FFBS step is $\overline{K}\times\overline{K}_Z$.

As before, after the state variables are sampled we keep track of only the states (s_t,z_t) in which at least one observation is assigned. These *alive* states are ordered from 1 to K for

s_t and 1 to K_Z for z_t, respectively. Where appropriate we keep track of the first $K+1$ or K_Z+1 values of parameter vectors.

The full parameter set consists of $\{u_{1:T}, s_{1:T}, z_{1:T}, \boldsymbol{\pi}_0, \Pi, \boldsymbol{\psi}_0, \Psi, \Phi, \Theta, \alpha, \beta, \kappa, \alpha_\psi, \beta_\psi\}$, where $z_{1:T} = \{z_t\}_{t=1}^T$. The sampling procedure sequentially simulates from the following conditional posterior distributions:

1. $p(u_{1:T}|s_{1:T}, z_{1:T}, \Pi, \Psi)$,
2. $p(s_{1:T}, z_{1:T}|\Pi, \Psi, u_{1:T}, \Phi, \Theta, \Sigma_{1:T})$,
3. $p(\boldsymbol{\pi}_0|s_{1:T}, \alpha, \beta, \kappa)$,
4. $p(\Pi|\boldsymbol{\pi}_0, s_{1:T}, \beta, \kappa)$,
5. $p(\boldsymbol{\psi}_0|z_{1:T}, \alpha_\psi, \beta_\psi)$,
6. $p(\Psi|\boldsymbol{\psi}_0, z_{1:T}, \beta_\psi)$,
7. $p(\Phi|z_{1:T}, \Theta, \Sigma_{1:T})$,
8. $p(\alpha, \beta, \kappa|s_{1:T}, \boldsymbol{\pi}_0)$,
9. $p(\alpha_\psi, \beta_\psi|z_{1:T}, \boldsymbol{\psi}_0)$,
10. $p(\Theta|z_{1:T}, \Phi, \Sigma_{1:T})$.

See the Appendix for details of each of the sampling steps. Note that compared to the iHMM models, the conditional posteriors of Φ and Θ depend on $z_{1:T}$, rather than $s_{1:T}$.

5.3.2 Predictive density

The predictive density can be computed in a similar way as before. Given a draw from the posterior,

$$
\begin{aligned}
& p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, s_{1:T}^{(i)}, \Psi^{(i)}, z_{1:T}^{(i)}, \Phi^{(i)}, \Theta^{(i)}) \\
= & \sum_{j=1}^{K^{(i)}} \left(\pi^{(i)}_{s_T^{(i)},j} \sum_{q=1}^{K_Z^{(i)}} \psi_{j,q}^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_q^{(i)}) \right) \\
& + \left(1 - \sum_{j=1}^{K^{(i)}} \left(\pi^{(i)}_{s_T^{(i)},j} \sum_{q=1}^{K_Z^{(i)}} \psi_{j,q}^{(i)} \right) \right) \int h(\Sigma_{T+1}|\Theta^{(i)}, \phi) G_0(d\phi) \qquad (73) \\
\approx & \sum_{j=1}^{K^{(i)}} \left(\pi^{(i)}_{s_T^{(i)},j} \sum_{q=1}^{K_Z^{(i)}} \psi_{j,q}^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_q^{(i)}) \right) \\
& + \left(1 - \sum_{j=1}^{K^{(i)}} \left(\pi^{(i)}_{s_T^{(i)},j} \sum_{q=1}^{K_Z^{(i)}} \psi_{j,q}^{(i)} \right) \right) \frac{1}{R} \sum_{l=1}^{R} h(\Sigma_{T+1}|\Theta^{(i)}, \phi^{[l]}), \qquad (74)
\end{aligned}
$$

where $\phi^{[l]} \overset{iid}{\sim} G_0, l = 1, \ldots, R$. Thus, the predictive density with all parameter uncertainty integrated out is obtained as

$$p(\Sigma_{T+1}|\Sigma_{1:T}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, s_{1:T}^{(i)}, \Psi^{(i)}, z_{1:T}^{(i)}, \Phi^{(i)}, \Theta^{(i)}). \tag{75}$$

6 Joint Modeling of Return and RCOV

Although our focus is on nonparametric modeling of RCOV we also evaluate the performance of our models through density forecasts of returns. Better models of RCOV should translate into better density forecasts of returns. We link any of the previous specifications of Σ_t with the following model to determine the distribution of returns given Σ_t,[9]

$$p(r_t|\Sigma_t) \quad = \quad \mathrm{N}(r_t|\mu, \Lambda^{1/2}\Sigma_t(\Lambda^{1/2})'). \tag{76}$$

As in Jin & Maheu (2013) Λ is a symmetric positive-definite matrix that can scale up or down Σ_t. This is a slightly different specification than Jin & Maheu (2013) which leads to better empirical performance. If $\Lambda = I$ then Σ_t is synonymous with the variance of returns, but we do not assume this is true in our analysis and place a prior on $\Lambda^{1/2}$ to estimate it. Besides allowing for additional flexibility an advantage of this approach is that (76) can be estimated independently once and used for any of our models of Σ_t. Each element of $\Lambda^{1/2}$ is assigned an independent normal prior with diagonal elements restricted to be positive for identification purpose. Estimation is conducted with a random walk proposal in an MH sampler.

If Σ_t is modelled parametrically or nonparametrically using the inverse-Wishart kernel instead of the Wishart kernel, it can be shown that linking r_t and Σ_t with (76) (instead of the other alternative in Jin & Maheu (2013)) renders a special implication for the conditional distribution of r_t. For example, if we assume (76) and Σ_t follows the IW-A(M) model, then after integrating out Σ_t we have

$$f(r_t|\Sigma_{1:t-1}, \nu, \Theta) = \mathrm{St}_k\left(r_t\middle|\mu, \frac{\nu-k-1}{\nu-k+1}\Lambda^{1/2}V_t(\Lambda^{1/2})', \nu-k+1\right). \tag{77}$$

$\mathrm{St}_k(.|\mu, \frac{\nu-k-1}{\nu-k+1}\Lambda^{1/2}V_t(\Lambda^{1/2})', \nu-k+1)$ denotes the density of a multivariate Student-t distribution with mean μ, $\nu-k+1$ degrees of freedom and the scale matrix equal to $\frac{\nu-k-1}{\nu-k+1}\Lambda^{1/2}V_t(\Lambda^{1/2})'$. Similarly, if Σ_t obeys the IW-DPM model, we will have

$$f(r_t|\Sigma_{1:t-1}, \Theta, \Omega, \Phi) = \sum_{j=1}^{\infty}\omega_j \mathrm{St}_k\left(r_t\middle|\mu, \frac{\nu_j-k-1}{\nu_j-k+1}\Lambda^{1/2}V_t^{1/2}A_j(V_t^{1/2})'(\Lambda^{1/2})', \nu_j-k+1\right), \tag{78}$$

which means that conditionally r_t follows an infinite mixture of multivariate Student-t distributions, with each component distribution having a different scale matrix $\frac{\nu_j-k-1}{\nu_j-k+1}\Lambda^{1/2}V_t^{1/2}A_j(V_t^{1/2})'(\Lambda^{1/2})'$ and a different degree of freedom $\nu_j - k + 1$.

[9] In estimation we fix $\mu = 0$ but additional dynamics such as an autoregressive process could be used.

These results also apply for the infinite hidden Markov models with an inverse-Wishart kernel and conditioning on the last state s_{t-1}. That is, combining returns with any of the inverse-Wishart based mixture models (with constant or time-varying weights) results in a countably infinite mixture of Student-t distributions for returns.

7 Estimation Results

The full sample estimates are reported in Tables 2, 3 and 4. In the latter table we confine our attention to iHMM models with the inverse-Wishart kernel as these performed better in forecasting than the Wishart alternatives. All models have an identical specification for V_t with 3 components which have the lag length estimated. Other specifications for V_t are possible, such as adding in an asymmetric effect from lagged returns[10] but this provided little gains. Therefore, we focus on the best model in Jin & Maheu (2013).

The following prior parameters are used. For IW-DPM and W-DPM, $\alpha \sim \text{Gamma}(2, 8)$. For IW-iHMM, $\alpha \sim \text{Gamma}(2, 8)$, $\beta \sim \text{Gamma}(2, 8)$; for IW-sticky-iHMM, $\alpha \sim \text{Gamma}(2, 8)$, $\beta + \kappa \sim \text{Gamma}(2, 8)$, $\rho \sim \text{Beta}(30, 0.1)$; for IW-sticky-iHMM-HDP, $\alpha \sim \text{Gamma}(2, 20)$, $\beta + \kappa \sim \text{Gamma}(2, 20)$, $\rho \sim \text{Beta}(30, 0.1)$, $\alpha_\psi \sim \text{Gamma}(2, 8)$, $\beta_\psi \sim \text{Gamma}(2, 8)$. For all the nonparametric models, $\gamma_0 = 10$ and $\lambda = 10$. For IW-A(3) and W-A(3), $\nu \sim \text{Exp}_{\nu>k+1}(10)$ and $\nu \sim \text{Exp}_{\nu>k}(10)$, respectively. The priors for Θ and $\Lambda^{1/2}$ are the same for all models. In particular, the priors for the elements of b_j's are all $\text{N}(0, 100)$, except that the first elements of b_j are truncated to be positive for identification purposes. The priors for ℓ_2 and ℓ_3 are uniform discrete with support $\{2, 3, \ldots, 200\}$, with the restriction $\ell_2 < \ell_3$. Each of the elements of $\Lambda^{1/2}$ are assumed to have a $\text{N}(0, 100)$ prior with diagonal elements restricted to be positive. In posterior simulation the first 10000 draws are discarded and the next 10000 for used for inference.

Table 2 reports estimates for the parametric models. The estimates of the component impacts, b_{ij}, indicate significant persistence. The lag lengths ℓ_2 and ℓ_3 are consistent with 2 weeks and just under 3 months for the IW-A(3) model.

Table 3 contains estimates of the DPM models IW-DPM and W-DPM. These models assume the conditional density of RCOV is unknown and approximate it with an i.i.d. mixture. All time variation in this specification comes through V_t as in the parametric models. Estimates of b_{ij} and ℓ_2 and ℓ_3 are similar to those in Table 2 and very precisely estimated. The parameter K is the number of alive clusters used. On average, the IW-DPM uses about 38 components in the mixture, much less than the W-DPM model which uses about 56. α, the precision parameter is also larger in the Wishart DPM model. According to these estimates the IW-DPM is using approximately 646 parameters[11], on average, to capture the unknown distribution.

Estimates of b_{ij} and the lag lengths of components are broadly similar in the iHMM models reported in Table 4. The IW-iHMM model and the IW-sticky-iHMM use an HDP to model the transition matrix of the iHMM. The *sticky* version introduces a prior to estimate the importance of state persistence. The final model, IW-sticky-iHMM-HDP employs two HDPs to model the infinite transition matrix of the iHMM and to model the linked DPM

[10] See Jin & Maheu (2013) for an example.

[11] $38 \times (k+1)k/2 + 38 + 38$ is the number of parameters in all A_j plus all v_j plus weights.

models in each state. The first two models in this table are similar and their hyperparameters estimates are close. For instance, the number of active states (K) is about 28, while α is just over 2.0. The first two iHMM specifications are using approximately 28 components to capture the time-varying structure in the conditional density of Σ_t.

The hierarchical structure in the IW-sticky-iHMM-HDP is quite different from the other two infinite hidden Markov models. The number of active states in the Markov chain drops to 10. On the other hand, the number of active states in the DPM portion of the model is around 26. If some of the structure in the conditional density of RCOV is constant over time the IW-sticky-iHMM-HDP model estimates this more parsimoniously than the iHMM specification. The precision parameters for the top level of the two hierarchical Dirichlet processes are very different with $\alpha = 0.4806$ for the Markov chain and $\alpha_\psi = 2.8799$ for the Dirichlet process mixtures. The estimates point to a clear distinction in the mixture model dynamics needed to capture the conditional density of Σ_t.

Figure 5 displays the estimate of states changes, $P(s_t \neq s_{t-1}|\Sigma_{1:T})$ for the IW-sticky-iHMM-HDP model. There are regular state changes and the model identifies these clearly. Adding the HDP structure to the model also increases persistence of states. The persistence parameter for states goes from 0 in the IW-iHMM to 0.57 ($\kappa = (\beta + \kappa) * \rho$ Table 4) for the IW-sticky-iHMM-HDP. Average state durations go from 1.095 (IW-DPM), 1.434 (IW-sticky-iHMM) to 3.456 (IW-sticky-iHMM-HDP). Finally, Figure 6 shows the impact of time variation in the IW-sticky-iHMM-HDP model versus the parametric version. Both panels indicate significant moves in the degree of freedom parameter and the log-determinant of A_j over time. The parametric model sets these to constants across time.

7.1 Forecasts

To compare the models we focus on out-of-sample density forecasts and also evaluate point forecasts. The out-of-sample period extends from $T_0 = 2006/03/31$ to $2007/12/31$, for a total of 441 observations. Each of the models is recursively estimated at each t in the out-of-sample period and forecasts are computed. To reduce some of the computational burden the first 3000 iterations of the posterior sampler are discard as burn-in and the next 5000 are used for inference.

As in Jin & Maheu (2013) a term structure of predictive likelihoods is computed for each model. This evaluates out-of-sample density forecasts of Σ_{t+h} for $h = 1, 5, 10, 20, 60$ given time t information. The cumulative log-predictive likelihood for model $\mathcal{A}$ and forecast horizon h is defined as

$$\sum_{t=T_0-h}^{T-h} \log p(\Sigma_{t+h}|\Sigma_{1:t}, \mathcal{A}). \tag{79}$$

This allows for comparison of the quality of density forecasts from each model from one day out to about 3 months out.[12] The log-predictive likelihoods for the full out-of-sample period are found in Table 5. Larger values indicate better models and log-predictive Bayes factors for the comparison of two models can be formed by subtracting the entries in the table for a

[12]For $h > 1$, the predictive density is computed in a similar way as with $h = 1$ in Sections 4 and 5, but requires simulating out the latent variables.

fixed h. A positive log-predictive Bayes factor favours the first model and a value in excess of 5 is considered strong.

The first point to note from Table 5 is that the inverse-Wishart specification is uniformly better than its Wishart counterpart. Yet, most of literature has used Wishart type models (Gourieroux et al. 2009, Golosnoy et al. 2012, Jin & Maheu 2013). The improvements are not minor. For instance, the log-Bayes factor for the IW-A(3) versus the W-A(3) is 1029.98 ($h = 1$), 813.97 ($h = 5$) and 754.69 ($h = 60$).

Moving from the IW-A(3) model to the IW-DPM model gives a huge improvement in the accuracy of density forecasts. Log-predictive Bayes factors are in excess of 1000. The difference in the individual log-predictive likelihoods at each time t is displayed in Figure 7. Except for a few periods the IW-DPM model is better and often significantly better with improvements in excess of 5 common. What we can conclude from this is that the parametric models are simply not competitive in terms of density forecasts. Mixture models offer large improvements and suggest important deviations from parametric distributional assumptions.

Table 5 also shows that when serving as a kernel in the DPM model, the inverse-Wishart distribution again provides significantly better results than the Wishart distribution (IW-DPM versus W-DPM) in all cases except for the longest forecast horizons $h = 20, 60$.

The ranking of the different nonparametric models is much closer, but successively more sophisticated models yield substantial improvements. For instance, the IW-sticky-iHMM-HDP against the IW-DPM has a log-Bayes factor of 84.48 ($h = 1$) while against the IW-sticky-iHMM is still significantly better with a log-Bayes factor of 40.2. With the exception of the forecast horizon of $h = 60$ the IW-sticky-iHMM-HDP significantly dominates every other model in the table.

A display of various cumulative log-Bayes factors is shown in Figure 8. Also shown is the log-generalized variance $\log|\Sigma_t|$, which is a single measure of variability. It is clear that no individual time period drives the results and that the final values reported on Table 5 are the result of regular ongoing gains.

Table 6 provides further details on model differences. Although there are general gains in moving to the nonparametric models this table identifies the differences in log-predictive likelihoods for outliers of Σ_t that are beyond 3 and 5 standard deviations from their sample mean. The outliers span the whole out-of-sample period and are not confined to diagonal elements only. It is clear that the the nonparametric model always does better for these outliers compared to the IW-A(3) model. In August 2007 the gains are particularly large. The last column compares the differences between two nonparametric models. The improvements from the IW-sticky-iHMM-HDP model are mixed but again this model does well in August 2007. In summary, the parametric models fail to account for extreme observations in diagonal and off-diagonal elements of realized covariance matrices while the nonparametric models do significantly better.

Differences in the densities can be seen from Figure 9 which displays (in-sample) density estimates of an equally weighted portfolio from selected models for three consecutive days in the out-of-sample period. The vertical line is the realized variance of the portfolio. Just before this period the realized variance takes a dramatic drop from about 10 to 1. The density of the IW-iHMM is quite different than the other models and is able to adapt quickly to this change in volatility.

Point forecasts in the form of predictive means are computed and their root mean squared

errors are found in Table 7. Results are consistent with the density forecasts. That is, the IW-sticky-iHMM-HDP performs the best at most forecast horizons. Compared to the IW-A(3) model our preferred nonparametric model achieves reductions of 7% ($h = 1$), 4.7% ($h = 5$), 5.4% ($h = 10$), 5.7% ($h = 20$) and 7.4% ($h = 60$) in root mean squared error.

Results on density forecasts for returns are found in Table 8. Each of the models for Σ_t is linked to returns through (76) with Λ estimated in general. This table reports multi-period log-predictive likelihood values for returns from each of the joint return-RCOV models. Also included is an asymmetric VD-GARCH-t model which uses only daily returns

$$r_t|r_{1:t-1} \sim \mathrm{St}_k(0, H_t, \zeta) \tag{80}$$

$$H_t = CC' + aa' \odot r_{t-1}r'_{t-1} + bb' \odot H_{t-1} + ee' \odot \eta_{t-1}\eta'_{t-1} \tag{81}$$

where $\eta_t = \max[\mathbf{0}, -r_t]$, and a, b, e are all $k \times 1$ vectors.

For the improvements in modeling RCOV to translate into improvements in return density forecasts it is important to estimate Λ. Generally, the ranking of the models is similar to our previous discussion, however, the gains are smaller but still significant. For instance, the log-Bayes factor for IW-sticky-iHMM-HDP versus IW-A(3) is 27.33 ($h = 1$), 28.39 ($h = 5$) and 37.24 ($h = 60$). The gains from modeling RCOV nonparametrically improve return density forecasts for each forecast horizon as compared to the GARCH model.

In summary, the time-varying mixture models presented in this paper provide large gains in density forecasts of RCOV and smaller, but still significant gains, for density forecasts for returns. The improvements that the nonparametric models provide in fitting the data are so substantial as to essentially make the parametric models we consider of little value.

7.2 Robustness

In this section we report on the sensitivity of the predictive likelihood results to various prior configurations. We focus on the IW-DPM and IW-sticky-iHMM-HDP specifications. The results are found in Tables 9 – 14. Tables 9 and 12 display the different prior assumptions for the two models. To match the previous results we estimate and compute forecast quantities for each t in the out-of-sample period for the the models with each new prior specification.

The log-predictive likelihoods of Σ_{t+h} for both models are fairly robust. There is some variation with the different priors, often leading to improved performance, but overall the gains discussed in the previous section are found here as well.

In both Tables 10 and 13 the final columns report the posterior mean of the number of alive clusters (K for IW-DPM and K and K_Z for IW-sticky-iHMM-HDP) from full sample estimation. The first value is from the benchmark prior and discussed above. The number of active clusters in the IW-DPM model is very robust to changes in the prior. The IW-sticky-iHMM-HDP model shows more changes but it is always the case that K_Z is about twice or more the size of K. In other words, the Markov switching component of the model has a much smaller dimension once each state s_t is modelled as a nonparametric DPM model.

The other Tables 11 and 14 report the cumulative log-predictive likelihoods for returns. In contrast to predictive likelihoods for Σ_{t+h}, different prior assumptions results in very little changes.

In summary, the dominance of the new nonparametric models is robust to different prior assumptions and the importance of combining Markov-switching behaviour with a DPM model is preserved in the IW-sticky-iHMM-HDP model.

8 Conclusion

This paper introduces several new Bayesian nonparametric models suitable for capturing the unknown conditional distribution of realized covariance (RCOV) matrices. Existing dynamic Wishart models are extended to countably infinite mixture models. We consider mixture models with constant weights as well as time-varying weights to capture time dependence in the unknown distribution. Each of our models can be combined with returns to provide a coherent joint model of returns and RCOV. The extensive forecast results show the new models provide very significant improvements in density forecasts for RCOV and returns and competitive point forecasts of RCOV. The parametric models fail to account for extreme observations in diagonal and off-diagonal elements of realized covariance matrices while the nonparametric models do significantly better.

The best performing model combines mixture dynamics from an infinite hidden Markov model and a Dirichlet process mixture. Our conclusion is that dynamic mixtures of inverse-Wishart distributions are a very promising area of research for modeling the conditional density of realized covariance matrices.

9 Appendix

In this section we provide details for the beam samplers for IW-sticky-iHMM and IW-sticky-iHMM-HDP.

9.1 Sampling details for IW-sticky-iHMM

Let K denote the number of active states in the state sequence $s_{1:T}$. Let n_{jl} denote the number of transitions from state j to state l in $s_{1:T}$, that is, $n_{jl} = \#\{t : s_{t-1} = j, s_t = l\}$. Also let $n_{j\cdot} = \sum_l n_{jl}$, $n_{\cdot l} = \sum_j n_{jl}$. A set of auxiliary variables, $\boldsymbol{m} = \{m_{jl}\}$, $\widetilde{\boldsymbol{m}} = \{\widetilde{m}_j\}$, $\overline{\boldsymbol{m}} = \{\overline{m}_{jl}\}$, are introduced to facilitate the sampling. We use the notation $m_{j\cdot} = \sum_l m_{jl}$, $m_{\cdot l} = \sum_j m_{jl}$, $m_{\cdot\cdot} = \sum_j \sum_l m_{jl}$. Similar notations are used for $\widetilde{\boldsymbol{m}}$ and $\overline{\boldsymbol{m}}$.

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. Initialize $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_j$ for $j = 1, \ldots, K$, all of which have $K + 1$ elements; Initialize ϕ_j for $j = 1, \ldots, K$; Initialize $\alpha, \beta, \kappa, \Theta$.

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

3. Sampling $s_{1:T}$:

(a) Set the initial value of $\overline{K}$ equal to K and if $\max\{\pi_{j,\overline{K}+1}\}_{j=1}^{\overline{K}} > \min\{u_t\}_{t=1}^{T}$, repeat the following steps:

i. Draw $\boldsymbol{\pi}_{\overline{K}+1} \sim \text{Dirichlet}(\beta\boldsymbol{\pi}_0)$.

ii. Break the last probability weight of $\boldsymbol{\pi}_0$, $\pi_{0\overline{K}+1}$:

A. Draw $\zeta \sim \text{Beta}(1, \alpha)$.

B. Add new probability weight $\pi_{0\overline{K}+2} = (1-\zeta)\pi_{0\overline{K}+1}$.

C. Update $\pi_{0\overline{K}+1} = \zeta\pi_{0\overline{K}+1}$.

iii. Break the last probability weight of $\boldsymbol{\pi}_j$ for $j = 1, \ldots, \overline{K}+1$:

A. Draw $\zeta_j \sim \text{Beta}(\beta\pi_{0\overline{K}+1}, \beta\pi_{0\overline{K}+2})$.

B. Add new probability weight $\pi_{j,\overline{K}+2} = (1-\zeta_j)\pi_{j,\overline{K}+1}$.

C. Update $\pi_{j,\overline{K}+1} = \zeta_j\pi_{j,\overline{K}+1}$.

iv. Draw $A_{\overline{K}+1} \sim \text{Wishart}_k(\gamma_0, \frac{1}{\gamma_0}I)$, $\nu_{\overline{K}+1} \sim \text{Exp}_{\nu>k+1}(\lambda)$.

v. Increment $\overline{K}$.

(b) Sample $s_{1:T}$ from $p(s_{1:T}|\Pi, u_{1:T}, \Phi, \Theta, \Sigma_{1:T})$ using the forward filtering and backward smoothing method based on Chib (1996):

i. Working sequentially forwards in time for $t = 1, \ldots, T$, repeat the following steps:

Prediction step: for $j = 1, \ldots, \overline{K}$, calculate

$$p(s_t = j|u_{1:T}, \Pi, \Phi, \Theta, \Sigma_{1:t-1}) \propto \sum_{i=1}^{\overline{K}} \mathbf{1}(u_t < \pi_{i,j}) p(s_{t-1} = i|u_{1:T}, \Pi, \Phi, \Theta, \Sigma_{1:t-1}). \quad (82)$$

Update step: for $j = 1, \ldots, \overline{K}$, calculate

$$p(s_t = j|u_{1:T}, \Pi, \Phi, \Theta, \Sigma_{1:t}) \propto p(s_t = j|u_{1:T}, \Pi, \Phi, \Theta, \Sigma_{1:t-1}) h(\Sigma_t|\Sigma_{1:t-1}, \Theta, \phi_j). \quad (83)$$

ii. Working sequentially backwards in time, sample $s_{1:T}$:

A. Sample s_T from $p(s_T|u_{1:T}, \Pi, \Phi, \Theta, \Sigma_{1:T})$.

B. Sample s_t from $p(s_t|u_{1:T}, \Pi, \Phi, \Theta, \Sigma_{1:t})\mathbf{1}(u_{t+1} < \pi_{s_t,s_{t+1}})$ for $t = T-1, \ldots, 1$.

(c) Cleaning up: Update K given $s_{1:T}$, re-label all the active states in $s_{1:T}$ in the order of $1, \ldots, K$ and remove the inactive states; Adapt $\boldsymbol{\pi}_0$, Π, A, ν according to the new labelling; Collapse π_{0K+1} and $\pi_{j,K+1}$ for $j = 1, \ldots, K$.

4. Sampling auxiliary variables $\boldsymbol{m}$, $\widetilde{\boldsymbol{m}}$, $\overline{\boldsymbol{m}}$:

(a) Sample $\boldsymbol{m}$: For $j = 1, \ldots, K$ and $l = 1, \ldots, K$, sample m_{jl} as follows: Set $m_{jl} = 0$. For $i = 1, \ldots, n_{jl}$, draw $x_i \sim \text{Bernoulli}(\frac{\beta\pi_{0l}+\kappa\delta(j,l)}{i-1+\beta\pi_{0l}+\kappa\delta(j,l)})$, where $\delta(.,.)$ denotes the discrete Kronecker delta. If $x_i = 1$, increment m_{jl}.

(b) Sampling $\widetilde{\boldsymbol{m}}$: For $j = 1, \ldots, K$, sample $\widetilde{m}_j \sim \text{Binomial}(m_{jj}, \frac{\rho}{\rho+\pi_{0j}(1-\rho)})$, where $\rho = \frac{\kappa}{\beta+\kappa}$.

(c) Update $\overline{\boldsymbol{m}}$: For $j = 1, \ldots, K$ and $l = 1, \ldots, K$, set $\overline{m}_{jl} = m_{jl}$ if $j \neq l$; set $\overline{m}_{jj} = m_{jj} - \widetilde{m}_j$.

5. Sampling $\boldsymbol{\pi}_0$: Draw

$$\boldsymbol{\pi}_0 \sim \text{Dirichlet}(\overline{m}_{\cdot 1}, \ldots, \overline{m}_{\cdot K}, \alpha). \tag{84}$$

6. Sampling Π: For $j = 1, \ldots, K$, sample

$$\boldsymbol{\pi}_j \sim \text{Dirichlet}(\beta\pi_{01} + n_{j1}, \ldots, \beta\pi_{0j} + \kappa + n_{jj}, \ldots, \beta\pi_{0K} + n_{jK}, \beta\pi_{0K+1}). \tag{85}$$

7. Sampling Φ: for $j = 1, \ldots, K$,

(a) draw

$$A_j \sim \text{Wishart}_k(\overline{\gamma}_j, \overline{Q}_j), \tag{86}$$

where $\overline{\gamma}_j = \gamma_0 + n_{\cdot j}\nu_j$, and $\overline{Q}_j = \left[(\nu_j - k - 1)\sum_{\{t:s_t=j\}}\left[(V_t^{1/2})\Sigma_t^{-1}((V_t^{1/2}))'\right] + \gamma_0 I\right]^{-1}$;

(b) sample

$$\begin{aligned} \nu_j &\sim p(\nu_j|\Sigma_{1:T}, s_{1:T}, A_j, \Theta) \\ &\propto p(\nu_j)\prod_{\{t:s_t=j\}} h(\Sigma_t|\Theta, \nu_j, A_j). \end{aligned} \tag{87}$$

An MH step with Gaussian random walk proposal is used.

8. Sampling hyperparameters α, β and κ:

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

i. For $j = 1, \ldots, K$, draw $\overline{\eta}_j \sim \text{Bernoulli}(\frac{n_{j\cdot}}{n_{j\cdot}+\beta+\kappa})$.

ii. For $j = 1, \ldots, K$, draw $\widetilde{\eta}_j \sim \text{Beta}(\beta + \kappa + 1, n_{j\cdot})$.

iii. Sample $\beta + \kappa \sim \text{Gamma}(a_4 + m_{\cdot\cdot} - \sum_{j=1}^K \overline{\eta}_j, c_4 - \sum_{l=1}^K \log\widetilde{\eta}_l)$.

(b) Sample ρ: Sample $\rho \sim \text{Beta}(a_5 + \widetilde{m}_{\cdot}, c_5 + m_{\cdot\cdot} - \widetilde{m}_{\cdot})$.

(c) Sample α:

i. Draw $\widetilde{\omega} \sim \text{Bernoulli}(\frac{\overline{m}_{\cdot\cdot}}{\overline{m}_{\cdot\cdot}+\alpha})$.

ii. Draw $\overline{\omega} \sim \text{Beta}(\alpha + 1, \overline{m}_{\cdot\cdot})$.

iii. Sample $\alpha \sim \text{Gamma}(a_3 + \widetilde{K} - \widetilde{\omega}, c_3 - \log(\overline{\omega}))$, where $\widetilde{K} = \sum_{l=1}^K \mathbf{1}(\overline{m}_{\cdot l} > 0)$.

9. Sample Θ: Note $p(\Theta|s_{1:T}, \Phi, \Sigma_{1:T}) \propto \prod_{t=1}^T h(\Sigma_t|\Theta, \phi_{s_t})p(\Theta)$. MH steps are used to sample elements of b_j's and ℓ_j as discussed in the benchmark models.

10. Repeat 2-9.

For IW-iHMM, fix $\kappa = 0$ and omit 4b, 8b while replace $\overline{\boldsymbol{m}}$ with $\boldsymbol{m}$.

9.2 Sampling details for IW-sticky-iHMM-HDP

Let K denote the number of active states in the state sequence $s_{1:T}$, and n_{jl} denote the number of transitions from state j to state l in $s_{1:T}$. Let K_Z denote the number of active clusters in sequence $z_{1:T}$, and n_{Zjl} denote the number of times cluster l is visited in state j, that is, $n_{Zjl} = \#\{t : s_t = j, z_t = l\}$. An extra auxiliary variable $\boldsymbol{m}_Z = \{m_{Zjl}\}$ is introduced.

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 value for K_Z and a starting $z_{1:T}$ sequence consisting of K_Z clusters which are labelled $1, \ldots, K_Z$; The infinite many unvisited components are merged into one component. Initialize $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_j$ for $j = 1, \ldots, K$, all of which have $K + 1$ elements; Initialize $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_j$ for $j = 1, \ldots, K$, all of which have $K_Z + 1$ elements; Initialize ϕ_j for $j = 1, \ldots, K_Z$; Initialize $\alpha, \beta, \kappa, \alpha_\psi, \beta_\psi, \Theta$.

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

3. Sampling $s_{1:T}, z_{1:T}$:

   (a) Set the initial value of $\overline{K}$ equal to K and if $\max\{\pi_{j,\overline{K}+1}\}_{j=1}^{\overline{K}} > \min\{u_t\}_{t=1}^{T}$, repeat the following steps:

      i. Draw $\boldsymbol{\pi}_{\overline{K}+1} \sim \mathrm{Dirichlet}(\beta\boldsymbol{\pi}_0)$.

      ii. Break the last probability weight of $\boldsymbol{\pi}_0$, $\pi_{0\overline{K}+1}$:

         A. Draw $\zeta \sim \mathrm{Beta}(1, \alpha)$.

         B. Add new probability weight $\pi_{0\overline{K}+2} = (1 - \zeta)\pi_{0\overline{K}+1}$.

         C. Update $\pi_{0\overline{K}+1} = \zeta\pi_{0\overline{K}+1}$.

      iii. Break the last probability weight of $\boldsymbol{\pi}_j$ for $j = 1, \ldots, \overline{K} + 1$:

         A. Draw $\zeta_j \sim \mathrm{Beta}(\beta\pi_{0\overline{K}+1}, \beta\pi_{0\overline{K}+2})$.

         B. Add new probability weight $\pi_{j,\overline{K}+2} = (1 - \zeta_j)\pi_{j,\overline{K}+1}$.

         C. Update $\pi_{j,\overline{K}+1} = \zeta_j\pi_{j,\overline{K}+1}$.

      iv. Draw $\boldsymbol{\psi}_{\overline{K}+1} \sim \mathrm{Dirichlet}(\beta_\psi\boldsymbol{\psi}_0)$.

      v. Increment $\overline{K}$.

   (b) Set the initial value of $\overline{K}_Z$ equal to K_Z and if $\max\{\psi_{j,\overline{K}_Z+1}\}_{j=1}^{\overline{K}} > \min\{u_t\}_{t=1}^{T}$, repeat the following steps:

      i. Break the last probability weight of $\boldsymbol{\psi}_0$, $\psi_{0\overline{K}_Z+1}$:

         A. Draw $\zeta \sim \mathrm{Beta}(1, \alpha_\psi)$.

         B. Add new probability weight $\psi_{0\overline{K}_Z+2} = (1 - \zeta)\psi_{0\overline{K}_Z+1}$.

         C. Update $\psi_{0\overline{K}_Z+1} = \zeta\psi_{0\overline{K}_Z+1}$.

      ii. Break the last probability weight of $\boldsymbol{\psi}_j$ for $j = 1, \ldots, \overline{K}$:

         A. Draw $\zeta_j \sim \mathrm{Beta}(\beta_\psi\psi_{0\overline{K}_Z+1}, \beta_\psi\psi_{0\overline{K}_Z+2})$.

         B. Add new probability weight $\psi_{j,\overline{K}_Z+2} = (1 - \zeta_j)\psi_{j,\overline{K}_Z+1}$.

         C. Update $\psi_{j,\overline{K}_Z+1} = \zeta_j\psi_{j,\overline{K}_Z+1}$.

iii. Draw $A_{\overline{K}_Z+1} \sim \text{Wishart}_k(\gamma_0, \frac{1}{\gamma_0} I)$, $\nu_{\overline{K}_Z+1} \sim \text{Exp}_{\nu>k+1}(\lambda)$.

iv. Increment $\overline{K}_Z$.

(c) Sample $s_{1:T}, z_{1:T}$ from $p(s_{1:T}, z_{1:T}|\Pi, \Psi, u_{1:T}, \Phi, \Theta, \Sigma_{1:T})$ using the forward filtering and backward smoothing method:

i. Working sequentially forwards in time for $t = 1, \ldots, T$, repeat the following steps:

Prediction step: for $j = 1, \ldots, \overline{K}$, $l = 1, \ldots, \overline{K}_Z$ calculate

$$\begin{aligned} &p(s_t = j, z_t = l|u_{1:T}, \Pi, \Psi, \Phi, \Theta, \Sigma_{1:t-1}) \\ &\propto \sum_{i=1}^{\overline{K}} \sum_{q=1}^{\overline{K}_Z} \mathbf{1}(u_t < \pi_{i,j}\psi_{j,l}) p(s_{t-1} = i, z_{t-1} = q|u_{1:T}, \Pi, \Psi, \Phi, \Theta, \Sigma_{1:t-1}) \end{aligned} \quad (88)$$

Update step: for $j = 1, \ldots, \overline{K}$, $l = 1, \ldots, \overline{K}_Z$ calculate

$$\begin{aligned} &p(s_t = j, z_t = l|u_{1:T}, \Pi, \Psi, \Phi, \Theta, \Sigma_{1:t}) \\ &\propto p(s_t = j, z_t = l|u_{1:T}, \Pi, \Psi, \Phi, \Theta, \Sigma_{1:t-1}) h(\Sigma_t|\Theta, \phi_l, \Sigma_{1:t-1}). \end{aligned} \quad (89)$$

ii. Working sequentially backwards in time for $t = 1, \ldots, T$, sample $s_{1:T}, z_{1:T}$:

A. Sample (s_T, z_T) from $p(s_T, z_T|u_{1:T}, \Pi, \Psi, \Phi, \Theta, \Sigma_{1:T})$.

B. Sample (s_t, z_t) from $p(s_t, z_t|u_{1:T}, \Pi, \Psi, \Phi, \Theta, \Sigma_{1:t}) \mathbf{1}(u_{t+1} < \pi_{s_t,s_{t+1}}\psi_{s_{t+1},z_{t+1}})$ for $t = T-1, \ldots, 1$.

(d) Cleaning up: Update K given $s_{1:T}$, re-label all the active states in $s_{1:T}$ in the order of $1, \ldots, K$ and remove the inactive states; Update K_Z given $z_{1:T}$, re-label all the alive clusters in $z_{1:T}$ in the order of $1, \ldots, K_Z$ and remove the unvisited components; Adapt $\boldsymbol{\pi}_0$, Π, $\boldsymbol{\psi}_0$,Ψ, A, ν according to the new labelling; Collapse π_{0K+1} and $\pi_{j,K+1}$ for $j = 1, \ldots, K$, and collapse ψ_{0K_Z+1} and ψ_{l,K_Z+1} for $l = 1, \ldots, K$.

4. Sampling auxiliary variables $\boldsymbol{m}$, $\widetilde{\boldsymbol{m}}$, $\overline{\boldsymbol{m}}$:

(a) Sample $\boldsymbol{m}$: For $j = 1, \ldots, K$ and $l = 1, \ldots, K$, sample m_{jl} as follows: Set $m_{jl} = 0$. For $i = 1, \ldots, n_{jl}$, draw $x_i \sim \text{Bernoulli}(\frac{\beta\pi_{0l} + \kappa\delta(j,l)}{i-1+\beta\pi_{0l}+\kappa\delta(j,l)})$. If $x_i = 1$, increment m_{jl}.

(b) Sampling $\widetilde{\boldsymbol{m}}$: For $j = 1, \ldots, K$, sample $\widetilde{m}_j \sim \text{Binomial}(m_{jj}, \frac{\rho}{\rho+\pi_{0j}(1-\rho)})$.

(c) Update $\overline{\boldsymbol{m}}$: For $j = 1, \ldots, K$ and $l = 1, \ldots, K$, set $\overline{m}_{jl} = m_{jl}$ if $j \neq l$; set $\overline{m}_{jj} = m_{jj} - \widetilde{m}_j$.

5. Sampling $\boldsymbol{\pi}_0$: Draw

$$\boldsymbol{\pi}_0 \sim \text{Dirichlet}(\overline{m}_{\cdot 1}, \ldots, \overline{m}_{\cdot K}, \alpha). \quad (90)$$

6. Sampling Π: For $j = 1, \ldots, K$, sample

$$\boldsymbol{\pi}_j \sim \text{Dirichlet}(\beta\pi_{01} + n_{j1}, \ldots, \beta\pi_{0j} + \kappa + n_{jj}, \ldots, \beta\pi_{0K} + n_{jK}, \beta\pi_{0K+1}). \quad (91)$$

7. Sampling auxiliary variables $\boldsymbol{m}_Z$: For $j = 1, \ldots, K$ and $l = 1, \ldots, K_Z$, sample m_{Zjl} as follows: Set $m_{Zjl} = 0$. For $i = 1, \ldots, n_{Zjl}$, draw $x_i \sim \text{Bernoulli}(\frac{\beta_\psi \psi_{0l}}{i-1+\beta_\psi \psi_{0l}})$. If $x_i = 1$, increment m_{Zjl}.

8. Sampling $\boldsymbol{\psi}_0$: Draw

$$\boldsymbol{\psi}_0 \sim \text{Dirichlet}(m_{Z.1}, \ldots, m_{Z.K_Z}, \alpha_\psi). \tag{92}$$

9. Sampling Ψ: For $j = 1, \ldots, K$, sample

$$\boldsymbol{\psi}_j \sim \text{Dirichlet}(\beta_\psi \psi_{01} + n_{Zj1}, \ldots, \beta_\psi \psi_{0K_Z} + n_{ZjK_Z}, \beta_\psi \psi_{0K_Z+1}). \tag{93}$$

10. Sampling Φ: for $j = 1, \ldots, K_Z$,

    (a) draw

$$A_j \sim \text{Wishart}_k(\overline{\gamma}_j, \overline{Q}_j), \tag{94}$$

    where $\overline{\gamma}_j = \gamma_0 + n_{Z.j}\nu_j$ and $\overline{Q}_j = \left[(\nu_j - k - 1)\sum_{\{t:z_t=j\}}\left[(V_t^{1/2})\Sigma_t^{-1}((V_t^{1/2}))'\right] + \gamma_0 I\right]^{-1}$;

    (b) sample

$$\begin{aligned} \nu_j &\sim p(\nu_j | \Sigma_{1:T}, z_{1:T}, A_j, \Theta) \\ &\propto p(\nu_j) \prod_{\{t:z_t=j\}} h(\Sigma_t | \Theta, \nu_j, A_j). \end{aligned} \tag{95}$$

    An MH step with Gaussian random walk proposal is used.

11. Sampling hyperparameters α, β and κ:

    (a) Sample $\beta + \kappa$:
        i. For $j = 1, \ldots, K$, draw $\overline{\eta}_j \sim \text{Bernoulli}(\frac{n_{j.}}{n_{j.}+\beta+\kappa})$.
        ii. For $j = 1, \ldots, K$, draw $\widetilde{\eta}_j \sim \text{Beta}(\beta + \kappa + 1, n_{j.})$.
        iii. Sample $\beta + \kappa \sim \text{Gamma}(a_7 + m_{..} - \sum_{j=1}^K \overline{\eta}_j, c_7 - \sum_{l=1}^K \log\widetilde{\eta}_l)$.

    (b) Sample ρ: Sample $\rho \sim \text{Beta}(a_8 + \widetilde{m}_{.}, c_8 + m_{..} - \widetilde{m}_{.})$.

    (c) Sample α:
        i. Draw $\widetilde{\omega} \sim \text{Bernoulli}(\frac{\overline{m}_{..}}{\overline{m}_{..}+\alpha})$.
        ii. Draw $\overline{\omega} \sim \text{Beta}(\alpha + 1, \overline{m}_{..})$.
        iii. Sample $\alpha \sim \text{Gamma}(a_6 + \widetilde{K} - \widetilde{\omega}, c_6 - \log(\overline{\omega}))$, where $\widetilde{K} = \sum_{l=1}^K \mathbf{1}(\overline{m}_{.l} > 0)$.

12. Sampling hyperparameters α_ψ, β_ψ:

    (a) Sample β_ψ:
        i. For $j = 1, \ldots, K$, draw $\overline{\eta}_j \sim \text{Bernoulli}(\frac{n_{Zj.}}{n_{Zj.}+\beta_\psi})$.
        ii. For $j = 1, \ldots, K$, draw $\widetilde{\eta}_j \sim \text{Beta}(\beta_\psi + 1, n_{Zj.})$.

iii. Sample $\beta_\psi \sim \text{Gamma}(a_{10} + m_{Z_{\cdot\cdot}} - \sum_{j=1}^{K} \overline{\eta}_j, c_{10} - \sum_{l=1}^{K} \log\widetilde{\eta}_l)$

(b) Sample α_ψ:

i. Draw $\widetilde{\omega} \sim \text{Bernoulli}(\frac{m_{Z_{\cdot\cdot}}}{m_{Z_{\cdot\cdot}} + \alpha_\psi})$.

ii. Draw $\overline{\omega} \sim \text{Beta}(\alpha_\psi + 1, m_{Z_{\cdot\cdot}})$.

iii. Sample $\alpha_\psi \sim \text{Gamma}(a_9 + K_Z - \widetilde{\omega}, c_9 - \log(\overline{\omega}))$.

13. Sample Θ: Note $p(\Theta|\Sigma_{1:T}, z_{1:T}, \Phi) \propto \prod_{t=1}^{T} h(\Sigma_t|\Theta, \phi_{z_t}) p(\Theta)$. MH steps are used to sample elements of b_j's and ℓ_j as discussed in the benchmark models.

14. Repeat 2-13.

To initialize parameters in this model we start with a fixed truncation version of the model and iterate on this for several hundred draws. After this we switch to the beam sampling approach discussed above.

References

Andersen, T. G. & Bollerslev, T. (1998), 'Answering the skeptics: Yes, standard volatility models do provide accurate forecasts', *International Economic Review* **39**(4), 885–905.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**(2), 579–625.

Asai, M. & McAleer, M. (2009), 'The structure of dynamic correlations in multivariate stochastic volatility models', *Journal of Econometrics* **150**(2), 182 – 192.

Asai, M. & So, M. K. P. (2013), 'Stochastic covariance models', *Journal of the Japan Statistical Society* **43**(2), 127–162.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2011), 'Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading', *Journal of Econometrics* **162**(2), 149 – 169.

Barndorff-Nielsen, O. E. & Shephard, N. (2004), 'Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics', *Econometrica* **72**(3), 885–925.

Bassetti, F., Casarin, R. & Leisen, F. (2014), 'Beta-product dependent Pitman-Yor processes for Bayesian inference', *Journal of Econometrics* **180**(1), 49 – 72.

Bauer, G. H. & Vorkink, K. (2011), 'Forecasting multivariate realized stock market volatility', *Journal of Econometrics* **160**(1), 93 – 101.

Bauwens, L., Braione, M. & Storti, G. (2014), Forecasting comparison of long term component dynamic models for realized covariance matrices. CORE Discussion Paper 2014/53.

Bonato, M., Caporin, M. & Ranaldo, A. (2008), Forecasting realized (co)variances with a block structure Wishart autoregressive model. Available at SSRN: http://ssrn.com/abstract=1282254.

Burda, M., Harding, M. & Hausman, J. (2008), 'A Bayesian mixed logit–probit model for multinomial choice', *Journal of Econometrics* **147**(2), 232 – 246.

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

Chib, S. & Greenberg, E. (2010), 'Additive cubic spline regression with Dirichlet process mixture errors', *Journal of Econometrics* **156**(2), 322 – 336.

Chiriac, R. & Voev, V. (2011), 'Modelling and forecasting multivariate realized volatility', *Journal of Applied Econometrics* **26**(6), 922–947.

Conley, T. G., Hansen, C. B., McCulloch, R. E. & Rossi, P. E. (2008), 'A semi-parametric Bayesian approach to the instrumental variable problem', *Journal of Econometrics* **144**(1), 276 – 305.

Corsi, F., Peluso, S. & Audrino, F. (2013), Missing in asynchronicity: A Kalman–EM approach for multivariate realized covariance estimation. Available at SSRN: http://ssrn.com/abstract=2000996.

Delatola, E.-I. & Griffin, J. (2013), 'A Bayesian semiparametric model for volatility with a leverage effect', *Computational Statistics & Data Analysis* **60**(0), 97 – 110.

Dufays, A. (2012), 'Infinite state Markov switching for dynamic volatility and correlation models', *CORE discussion paper 2012/43* .

Escobar, M. & West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association* **90**, 577–588.

Fox, E. B. & West, M. (2011), Autoregressive models for variance matrices: Stationary inverse Wishart processes. ArXiv e-prints, http://arxiv.org/abs/1107.5239.

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.

Golosnoy, V., Gribisch, B. & Liesenfeld, R. (2012), 'The conditional autoregressive Wishart model for multivariate stock market volatility', *Journal of Econometrics* **167**(1), 211–223.

Gourieroux, C., Jasiak, J. & Sufana, R. (2009), 'The Wishart autoregressive process of multivariate stochastic volatility', *Journal of Econometrics* **150**, 167–181.

Griffin, J. E. & Steel, M. F. J. (2006), 'Order-based dependent Dirichlet processes', *Journal of the American Statistical Association* **101**(473), 179–194.

Griffin, J. & Steel, M. (2011), 'Stick-breaking autoregressive processes', *Journal of Econometrics* **162**(2), 383 – 396.

Hamilton, J. D. (1989), 'A new approach to the economic analysis of non-stationary time series and the business cycle', *Econometrica* **57**, 357–384.

Hansen, P. R., Lunde, A. & Voev, V. (2013), 'Realized beta GARCH: A multivariate GARCH model with realized measures of volatility', *forthcoming Journal of Econometrics* .

Hautsch, N., Kyj, L. M. & Oomen, R. C. A. (2012), 'A blocking and regularization approach to high-dimensional realized covariance estimation', *Journal of Applied Econometrics* **27**(4), 625–645.

Hirano, K. (2002), 'Semiparametric Bayesian inference in autoregressive panel data models', *Econometrica* **70**, 781–799.

Janusa, P., Lucas, A. & Opschoor, A. (2014), New HEAVY models for fat-tailed returns and realized covariance kernels. Tinbergen Institute Discussion Paper 2014-073.

Jensen, M. J. & Maheu, J. M. (2013*a*), 'Bayesian semiparametric multivariate GARCH modeling', *Journal of Econometrics* **176**(1), 3 – 17.

Jensen, M. J. & Maheu, J. M. (2013*b*), Risk, return and volatility feedback: A Bayesian nonparametric analysis. MPRA Working Paper No. 52132.

Jin, X. & Maheu, J. M. (2013), 'Modeling realized covariances and returns', *Journal of Financial Econometrics* **11**(2), 335–369.

Jochmann, M. (2014), 'Modeling U.S. inflation dynamics: a Bayesian nonparametric approach', *forthcoming Econometric Reviews* .

Kalli, M. & Griffin, J. E. (2014), 'Flexible modelling of dependence in volatility processes', *forthcoming Journal of Business & Economic Statistics* .

Kalli, M., Griffin, J. & Walker, S. (2011), 'Slice sampling mixture models', *Statistics and Computing* **21**, 93–105.

Lin, M., Liu, C. & Niu, L. (2012), Bayesian estimation of Wishart autoregressive stochastic volatility model. manuscript, WISE, Xiamen University.

MacEachern, S. (2000), Dependent Dirichlet processes. Technical Report, Ohio State University.

Noureldin, D., Shephard, N. & Sheppard, K. (2012), 'Multivariate high-frequency-based volatility (HEAVY) models', *Journal of Applied Econometrics* **27**(6), 907–933.

Papaspiliopoulos, O. (2008), A note on posterior sampling from dirichlet mixture models. manuscript, Department of Economics, Universitat Pompeu Fabra.

Philipov, A. & Glickman, M. E. (2006), 'Multivariate stochastic volatility via Wishart processes', *Journal of Business & Economic Statistics* **24**(3), 313–328.

Press, S. J. (2005), *Applied multivariate analysis: using Bayesian and frequentist methods of inference*, Dover Publications.

Rodriguez, A. & Dunson, D. B. (2011), 'Nonparametric Bayesian models through probit stick-breaking processes', *Bayesian Analysis* **6**(1), 145–177.

Rodriguez, A., Dunson, D. B. & Gelfand, A. E. (2008), 'The nested Dirichlet process', *Journal of the American Statistical Association* **103**(483), 1131–1154.

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

Sheppard, K. & Xu, W. (2014), Factor high-frequency based volatility (HEAVY) models. Available at SSRN: http://ssrn.com/abstract=2442230.

Shi, S. & Song, Y. (2014), 'Identifying speculative bubbles using an infinite hidden Markov model', *forthcoming Journal of Financial Econometrics* .

Song, Y. (2014), 'Modelling regime switching and structural breaks with an infinite hidden Markov model', *Journal of Applied Econometrics* **29**(5), 825–842.

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

Triantafyllopoulos, K. (2012), 'Multivariate stochastic volatility modelling using Wishart autoregressive processes', *Journal of Time Series Analysis* **33**(1), 48–60.

Uhlig, H. (1997), 'Bayesian vector autoregressions with stochastic volatility', *Econometrica* **65**(1), 59–73.

Van Gael, J. & Ghahramani, Z. (2010), Nonparametric hidden markov models, *in* 'Inference and Estimation in Probabilistic Time-Series Models', Cambridge University Press, Cambridge.

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.

Virbickaite, A., Ausn, M. C. & Galeano, P. (2013), 'Bayesian inference methods for univariate and multivariate GARCH models: A survey', *forthcoming Journal of Economic Surveys* .

Walker, S. G. (2007), 'Sampling the Dirichlet mixture model with slices', *Communications in Statistics – Simulation and Computation* **36**, 45–54.

Windle, J. & Carvalho, C. M. (2014), A tractable state-space model for symmetric positive-definite matrices. forthcoming *Bayesian Analysis*.

Table 1: Summary Statistics: Daily Returns and RCOV

	Sample covariance from daily returns					Average of realized covariances				
	SPY	GE	C	AA	BA	SPY	GE	C	AA	BA
SPY	1.30	1.52	1.70	1.29	1.07	1.27	1.39	1.56	1.11	0.96
GE		3.18	2.19	1.66	1.32		3.18	1.83	1.29	1.12
C			4.35	1.75	1.38			4.62	1.41	1.21
AA				5.13	1.46				5.08	1.01
BA					4.14					3.84

This table reports the sample covariance from daily returns and the sample average of the realized covariances. The data are Standard and Poor's Depository Receipt (SPY), General Electric Co. (GE), Citigroup Inc.(C), Alcoa Inc. (AA) and Boeing Co. (BA). Total observations is 2281.

Table 2: Full Sample Estimates for IW-A(3) and W-A(3)

	IW-A(3)			W-A(3)		
	Mean	Stdev	0.95DI	Mean	Stdev	0.95DI
b_{11}	0.3580	0.0128	$(0.3330, 0.3870)$	0.4226	0.0163	$(0.3906, 0.4522)$
b_{12}	0.4048	0.0127	$(0.3796, 0.4286)$	0.4601	0.0161	$(0.4284, 0.4929)$
b_{13}	0.3987	0.0122	$(0.3732, 0.4192)$	0.4410	0.0201	$(0.4021, 0.4753)$
b_{14}	0.3265	0.0179	$(0.2895, 0.3597)$	0.3824	0.0249	$(0.3324, 0.4293)$
b_{15}	0.5232	0.0118	$(0.4993, 0.5448)$	0.5587	0.0173	$(0.5250, 0.5926)$
b_{21}	0.7154	0.0090	$(0.6973, 0.7312)$	0.6651	0.0130	$(0.6387, 0.6893)$
b_{22}	0.6896	0.0109	$(0.6685, 0.7134)$	0.6045	0.0154	$(0.5782, 0.6383)$
b_{23}	0.7252	0.0093	$(0.7055, 0.7415)$	0.6547	0.0170	$(0.6150, 0.6881)$
b_{24}	0.5723	0.0169	$(0.5378, 0.6056)$	0.4825	0.0241	$(0.4368, 0.5291)$
b_{25}	0.6009	0.0180	$(0.5624, 0.6350)$	0.5291	0.0225	$(0.4840, 0.5719)$
b_{31}	0.5005	0.0122	$(0.4771, 0.5239)$	0.5696	0.0111	$(0.5474, 0.5938)$
b_{32}	0.5235	0.0148	$(0.4941, 0.5547)$	0.6187	0.0130	$(0.5907, 0.6433)$
b_{33}	0.4970	0.0131	$(0.4716, 0.5209)$	0.5882	0.0138	$(0.5611, 0.6157)$
b_{34}	0.6863	0.0158	$(0.6541, 0.7176)$	0.6575	0.0161	$(0.6233, 0.6858)$
b_{35}	0.4906	0.0208	$(0.4494, 0.5297)$	0.5612	0.0161	$(0.5271, 0.5888)$
ν	12.5621	0.0494	$(12.4619, 12.6562)$	10.6395	0.0639	$(10.5127, 10.7640)$
ℓ_2	13.0000	0.0000	$(13.0000, 13.0000)$	5.0000	0.0000	$(5.0000, 5.0000)$
ℓ_3	62.3684	0.8608	$(61.0000, 63.0000)$	65.9914	0.0923	$(66.0000, 66.0000)$

This table reports the posterior mean, standard deviation and 0.95 probability density intervals for model parameters.

Table 3: Full Sample Estimates for IW-DPM and W-DPM

	IW-DPM			W-DPM		
	Mean	Stdev	0.95DI	Mean	Stdev	0.95DI
b_{11}	0.2419	0.0164	(0.2091, 0.2720)	0.2817	0.0168	(0.2484, 0.3143)
b_{12}	0.2802	0.0227	(0.2342, 0.3274)	0.2918	0.0194	(0.2513, 0.3267)
b_{13}	0.2751	0.0210	(0.2290, 0.3174)	0.3026	0.0233	(0.2553, 0.3462)
b_{14}	0.2478	0.0365	(0.1766, 0.3162)	0.2349	0.0335	(0.1671, 0.2997)
b_{15}	0.4209	0.0207	(0.3795, 0.4608)	0.4430	0.0214	(0.4004, 0.4830)
b_{21}	0.6532	0.0112	(0.6318, 0.6751)	0.6264	0.0129	(0.6014, 0.6519)
b_{22}	0.6274	0.0140	(0.6016, 0.6551)	0.6020	0.0170	(0.5740, 0.6366)
b_{23}	0.6742	0.0127	(0.6515, 0.7003)	0.6679	0.0162	(0.6341, 0.6953)
b_{24}	0.4888	0.0251	(0.4394, 0.5368)	0.4994	0.0286	(0.4391, 0.5534)
b_{25}	0.4577	0.0274	(0.4003, 0.5099)	0.4600	0.0276	(0.4021, 0.5167)
b_{31}	0.6302	0.0133	(0.6024, 0.6556)	0.6386	0.0137	(0.6114, 0.6648)
b_{32}	0.6473	0.0149	(0.6177, 0.6749)	0.6647	0.0157	(0.6297, 0.6914)
b_{33}	0.6283	0.0146	(0.5997, 0.6583)	0.6191	0.0165	(0.5866, 0.6547)
b_{34}	0.7376	0.0191	(0.6993, 0.7729)	0.7150	0.0224	(0.6687, 0.7550)
b_{35}	0.7035	0.0228	(0.6580, 0.7430)	0.6802	0.0210	(0.6357, 0.7199)
ℓ_2	13.0000	0.0000	(13.0000, 13.0000)	13.2716	0.5052	(13.0000, 15.0000)
ℓ_3	64.6217	0.9445	(64.0000, 67.0000)	64.2398	0.6671	(63.0000, 66.0000)
α	2.7012	0.4635	(1.8680, 3.6653)	4.0587	0.5904	(2.9989, 5.2988)
K	38.1623	1.8080	(35.0000, 42.0000)	56.5360	2.9715	(51.0000, 63.0000)

This table reports the posterior mean, standard deviation and 0.95 probability density intervals for model parameters. K is the number of alive clusters.

Table 4: Full Sample Estimates

	IW-iHMM		IW-sticky-iHMM		IW-sticky-iHMM-HDP	
	Mean	0.95DI	Mean	0.95DI	Mean	0.95DI
b_{11}	0.1871	$(0.1509, 0.2245)$	0.1386	$(0.0960, 0.1837)$	0.2430	$(0.2078, 0.2773)$
b_{12}	0.1984	$(0.1406, 0.2469)$	0.1728	$(0.1062, 0.2300)$	0.2903	$(0.2507, 0.3298)$
b_{13}	0.2438	$(0.1891, 0.2960)$	0.2013	$(0.1462, 0.2481)$	0.2890	$(0.2500, 0.3303)$
b_{14}	0.2609	$(0.1533, 0.3225)$	0.1317	$(0.0180, 0.2260)$	0.2021	$(0.1444, 0.2584)$
b_{15}	0.2707	$(0.1850, 0.3346)$	0.2920	$(0.2384, 0.3369)$	0.3616	$(0.3155, 0.4031)$
b_{21}	0.6212	$(0.5941, 0.6449)$	0.6096	$(0.5819, 0.6354)$	0.6028	$(0.5753, 0.6303)$
b_{22}	0.5922	$(0.5621, 0.6183)$	0.5759	$(0.5451, 0.6126)$	0.5777	$(0.5437, 0.6116)$
b_{23}	0.6502	$(0.6163, 0.6862)$	0.6407	$(0.6131, 0.6659)$	0.6083	$(0.5728, 0.6400)$
b_{24}	0.4572	$(0.4078, 0.5095)$	0.4738	$(0.4259, 0.5199)$	0.4829	$(0.4230, 0.5456)$
b_{25}	0.4311	$(0.3665, 0.4992)$	0.4281	$(0.3728, 0.4838)$	0.4076	$(0.3332, 0.4805)$
b_{31}	0.6698	$(0.6444, 0.6938)$	0.6877	$(0.6636, 0.7142)$	0.6653	$(0.6371, 0.6890)$
b_{32}	0.6937	$(0.6660, 0.7192)$	0.7135	$(0.6817, 0.7406)$	0.6719	$(0.6417, 0.7009)$
b_{33}	0.6577	$(0.6225, 0.6850)$	0.6786	$(0.6533, 0.7048)$	0.6707	$(0.6447, 0.6991)$
b_{34}	0.7434	$(0.7051, 0.7821)$	0.7569	$(0.7163, 0.7957)$	0.7360	$(0.6904, 0.7780)$
b_{35}	0.7760	$(0.7341, 0.8138)$	0.7611	$(0.7248, 0.7992)$	0.7550	$(0.7094, 0.7958)$
ℓ_2	14.2747	$(13.0000, 16.0000)$	13.4932	$(13.0000, 14.0000)$	15.9964	$(16.0000, 16.0000)$
ℓ_3	64.2628	$(63.0000, 68.0000)$	64.0265	$(64.0000, 64.0000)$	67.0946	$(64.0000, 68.0000)$
α	2.0792	$(1.3270, 2.9635)$	2.0203	$(1.2841, 2.9260)$	0.4806	$(0.2390, 0.8078)$
$\beta + \kappa$	1.9951	$(1.6541, 2.3818)$	2.1287	$(1.7650, 2.5303)$	0.6802	$(0.4273, 0.9891)$
ρ	0.0000	$(0.0000, 0.0000)$	0.2682	$(0.2106, 0.3288)$	0.6814	$(0.5636, 0.7911)$
K	28.5290	$(28.0000, 30.0000)$	27.2671	$(27.0000, 28.0000)$	10.4747	$(10.0000, 11.0000)$
α_ψ	$--$	$--$	$--$	$--$	2.8799	$(1.9606, 3.9650)$
β_ψ	$--$	$--$	$--$	$--$	1.8194	$(1.3499, 2.3314)$
K_Z	$--$	$--$	$--$	$--$	26.4110	$(26.0000, 28.0000)$

This table reports the posterior mean and 0.95 probability density intervals for model parameters. K is the number of alive clusters in the infinite hidden Markov model and K_Z is the number of alive clusters in the state dependent DPM mixture in the IW-sticky-iHMM-HDP specification.

Table 5: Cumulative Log-predictive Likelihoods for RCOV

Model	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW-A(3)	-968.51	-1344.51	-1471.53	-1708.97	-1959.79
W-A(3)	-1998.49	-2158.48	-2315.78	-2558.65	-2714.48
IW-DPM	110.41	-82.15	-169.58	-335.31	-523.33
W-DPM	-6.30	-161.38	-204.20	-333.36	$\mathbf{-383.15}$
IW-iHMM	137.10	-70.32	-157.66	-313.68	-492.52
IW-sticky-iHMM	154.69	-75.01	-161.56	-296.96	-467.45
IW-sticky-iHMM-HDP	**194.89**	**−39.08**	**−107.01**	**−260.69**	-397.63

The table reports the cumulative log-predictive likelihoods for RCOV at different forecast horizon h. The first two models are parametric while the remainder are nonparametric.

Table 6: Difference in Log-predictive Likelihoods for Outliers

	Outlier		Difference in log-predictive likelihoods	
date	3 stdev	5 stdev	IW-DPM/IW-A(3)	IW-sticky-iHMM-HDP/IW-DPM
060411		$\Sigma_{4,4}$	5.65	0.04
060515	$\Sigma_{4,4}$		4.09	−0.24
061011		$\Sigma_{4,4}$	3.39	−1.49
070213		$\Sigma_{4,4}$	3.59	0.40
070425	$\Sigma_{4,4}$		4.80	−0.67
070507		$\Sigma_{4,4}$	3.50	−2.31
070712		$\Sigma_{4,4}$	5.91	0.82
070719	$\Sigma_{4,4}$		6.56	0.29
070809	$\Sigma_{4,1}, \Sigma_{4,3}$		12.67	2.48
070810	$\Sigma_{4,1}$		14.24	2.11
070816	$\Sigma_{4,1}$	$\Sigma_{4,4}$	17.99	3.77
070817	$\Sigma_{3,1}, \Sigma_{4,1}$		6.93	1.67
071101		$\Sigma_{3,3}$	5.89	−1.48
071108	$\Sigma_{4,4}$		11.20	−1.46
average of above			7.6	0.28
average of whole out-of-sample			2.44	0.19

The table reports the differences in log-predictive likelihoods between IW-DPM and IW-A(3) models; and IW-sticky-iHMM-HDP and IW-DPM models for outliers in the out-of-sample period. The RCOV element Σ_{ij} identifies the observations that are 3 and 5 standard deviations away from their sample means. The second last row gives the average for the outliers. The last row gives the average among the whole out-of-sample period.

Table 7: Root Mean Squared Error $RMSE_h$ for the Predictive Mean of RCOV

Model	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW-A(3)	5.3140	5.7362	5.9561	6.2834	6.4737
W-A(3)	5.2909	5.7633	6.0153	6.2777	6.4130
IW-DPM	5.2409	5.5306	5.7038	5.9637	6.0080
W-DPM	5.4044	5.7769	6.0192	6.4580	7.2087
IW-iHMM	4.9687	5.5081	5.6800	5.9403	**5.9734**
IW-sticky-iHMM	4.9592	5.5089	5.6950	5.9417	5.9970
IW-sticky-iHMM-HDP	**4.9395**	**5.4623**	**5.6318**	**5.9246**	5.9883

The table reports the root mean squared error for predictive mean of RCOV at different forecast horizon h. $RMSE_h = \frac{1}{T-T_0+1}\sum_{t=T_0-h}^{T-h}\|\Sigma_{t+h} - E[\Sigma_{t+h}|\Sigma_{1:t}]\|$, where $\|A\| = \sqrt{\sum_i \sum_j |a_{ij}|^2}$, and $E[\Sigma_{t+h}|\Sigma_{1:t}]$ denotes a model's predictive mean. The first two models are parametric while the remainder are nonparametric.

Table 8: Cumulative Log-predictive Likelihoods for Returns

Model	$h=1$	$h=5$	$h=10$	$h=20$	$h=60$
asymmetric VDGARCH-t	−2989.07	−3008.07	−3013.54	−3038.39	−3083.56
IW-A(3) $\Lambda = I$	−2989.19	−3025.39	−3037.46	−3069.36	−3095.36
IW-A(3)	−2965.69	−3002.47	−3014.11	−3048.35	−3077.35
IW-DPM	−2945.71	**−2973.57**	−2985.46	−3015.57	−3042.97
IW-iHMM	−2944.43	−2974.58	−2986.59	−3016.89	−3044.78
IW-sticky-iHMM	−2946.01	−2973.73	−2985.07	−3015.11	−3040.68
IW-sticky-iHMM-HDP	**−2938.36**	−2974.08	**−2984.69**	**−3011.38**	**−3040.11**

The table reports the cumulative log-predictive likelihoods for return data at different forecast horizon h. The first three models are parametric while the remainder are nonparametric.

Table 9: Prior Specifications for IW-DPM

prior 0	$\gamma_0 = 10, \lambda = 10, \alpha \sim \text{Gamma}(2, 8)$
prior 1	$\gamma_0 = 15, \lambda = 15, \alpha \sim \text{Gamma}(2, 8)$
prior 2	$\gamma_0 = 10, \lambda = 10, \alpha \sim \text{Gamma}(1, 12)$
prior 3	$\gamma_0 = 15, \lambda = 15, \alpha \sim \text{Gamma}(1, 12)$

Prior 0 is the benchmark prior used in the paper.

Table 10: Cumulative Log-predictive Likelihoods for RCOV, IW-DPM

IW-DPM	$h=1$	$h=5$	$h=10$	$h=20$	$h=60$	K
prior 0	110.41	−82.15	−169.58	−335.31	−523.33	38
prior 1	124.96	−77.10	−164.10	−336.93	−511.05	43
prior 2	113.88	−77.05	−164.22	−335.76	−513.05	40
prior 3	120.32	−82.69	−172.32	−337.81	−519.10	43

The table reports the cumulative log-predictive likelihoods for RCOV at different forecast horizon h for the IW-DPM model. Prior 0 is the benchmark prior used in the paper. The last column, K, records the posterior mean of the number of alive clusters based on a full sample estimation.

Table 11: Cumulative Log-predictive Likelihoods for Returns, IW-DPM

IW-DPM	$h=1$	$h=5$	$h=10$	$h=20$	$h=60$
prior 0	−2945.71	−2973.57	−2985.46	−3015.57	−3042.97
prior 1	−2945.27	−2974.18	−2985.39	−3015.56	−3042.42
prior 2	−2945.65	−2974.12	−2985.02	−3015.22	−3041.67
prior 3	−2945.60	−2974.67	−2985.44	−3015.37	−3042.56

The table reports the cumulative log-predictive likelihoods for return data at different forecast horizon h for the IW-DPM model. Prior 0 is the benchmark prior used in the paper.

Table 12: Prior Specifications for IW-sticky-iHMM-HDP

prior 0	$\gamma_0 = 10, \lambda = 10, \alpha \sim \text{Gamma}(2, 20), \beta + \kappa \sim \text{Gamma}(2, 20), \rho \sim \text{Beta}(30, 0.1),$ $\alpha_\psi \sim \text{Gamma}(2, 8), \beta_\psi \sim \text{Gamma}(2, 8)$
prior 1	$\gamma_0 = 15, \lambda = 15, \alpha \sim \text{Gamma}(2, 20), \beta + \kappa \sim \text{Gamma}(2, 20), \rho \sim \text{Beta}(30, 0.1),$ $\alpha_\psi \sim \text{Gamma}(2, 8), \beta_\psi \sim \text{Gamma}(2, 8)$
prior 2	$\gamma_0 = 10, \lambda = 10, \alpha \sim \text{Gamma}(2, 8), \beta + \kappa \sim \text{Gamma}(2, 8), \rho \sim \text{Beta}(30, 0.1),$ $\alpha_\psi \sim \text{Gamma}(2, 8), \beta_\psi \sim \text{Gamma}(2, 8)$
prior 3	$\gamma_0 = 10, \lambda = 10, \alpha \sim \text{Gamma}(2, 20), \beta + \kappa \sim \text{Gamma}(2, 20), \rho \sim \text{Beta}(10, 0.1),$ $\alpha_\psi \sim \text{Gamma}(2, 8), \beta_\psi \sim \text{Gamma}(2, 8)$
prior 4	$\gamma_0 = 10, \lambda = 10, \alpha \sim \text{Gamma}(2, 20), \beta + \kappa \sim \text{Gamma}(2, 20), \rho \sim \text{Beta}(30, 0.1),$ $\alpha_\psi \sim \text{Gamma}(1, 12), \beta_\psi \sim \text{Gamma}(1, 12)$
prior 5	$\gamma_0 = 15, \lambda = 15, \alpha \sim \text{Gamma}(2, 8), \beta + \kappa \sim \text{Gamma}(2, 8), \rho \sim \text{Beta}(10, 0.1),$ $\alpha_\psi \sim \text{Gamma}(1, 12), \beta_\psi \sim \text{Gamma}(1, 12)$

Prior 0 is the benchmark prior used in the paper.

Table 13: Cumulative Log-predictive Likelihoods for RCOV, IW-sticky-iHMM-HDP

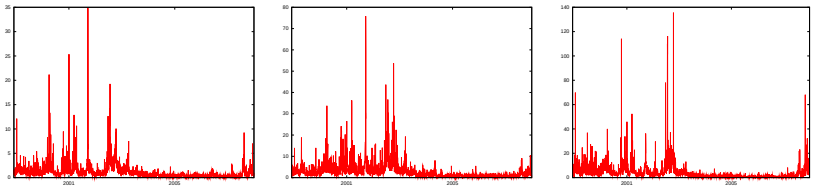
IW-sticky-iHMM-HDP	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$	K	K_Z
prior 0	194.89	−39.08	−107.01	−260.69	−397.63	10	26
prior 1	203.69	−49.04	−122.58	−279.99	−392.22	10	30
prior 2	197.99	−41.48	−137.86	−250.54	−376.65	14	28
prior 3	180.12	−54.10	−130.80	−249.36	−404.28	13	29
prior 4	200.62	−40.67	−119.68	−256.28	−407.29	10	27
prior 5	190.29	−60.65	−135.12	−270.67	−397.97	14	28

The table reports the cumulative log-predictive likelihoods for RCOV at different forecast horizon h for the IW-sticky-iHMM-HDP model. Prior 0 is the benchmark prior used in the paper. The last two columns, K and K_Z, report the posterior mean of the number of alive clusters for s_t and z_t, respectively, based on a full sample estimation.

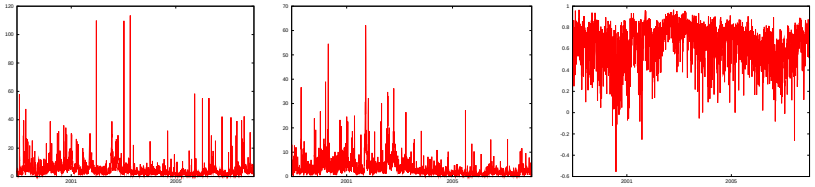
Table 14: Cumulative Log-predictive Likelihoods for Returns, IW-sticky-iHMM-HDP

IW-sticky-iHMM-HDP	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
prior 0	−2938.36	−2974.08	−2984.69	−3011.38	−3040.11
prior 1	−2941.34	−2975.59	−2986.56	−3018.58	−3037.28
prior 2	−2940.27	−2976.00	−2984.90	−3014.64	−3037.54
prior 3	−2943.41	−2975.64	−2983.53	−3013.30	−3039.17
prior 4	−2941.97	−2975.68	−2982.07	−3012.40	−3042.08
prior 5	−2943.29	−2979.02	−2986.43	−3016.49	−3035.63

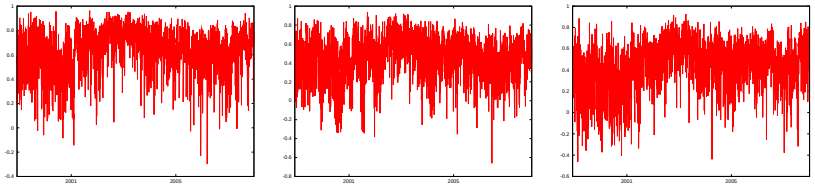
The table reports the cumulative log-predictive likelihoods for return data at different forecast horizon h for the IW-sticky-iHMM-HDP model. Prior 0 is the benchmark prior used in the paper.



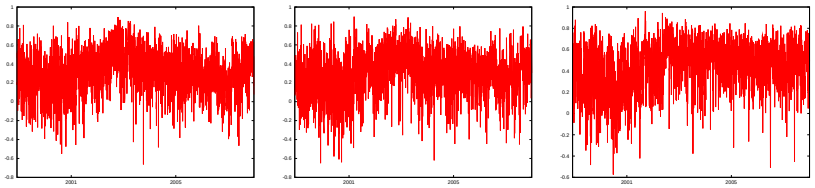
(a) RV:SPY (b) RV:GE (c) RV:C



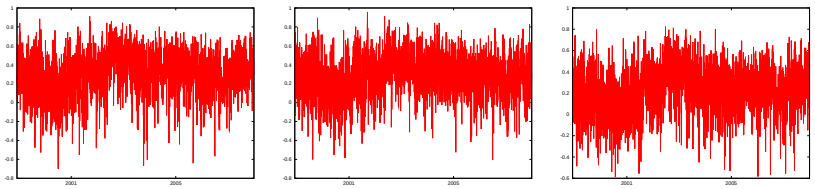
(d) RV:AA (e) RV:BA (f) RCORR:SPY-GE



(g) RCORR:SPY-C (h) RCORR:GE-C (i) RCORR:SPY-AA

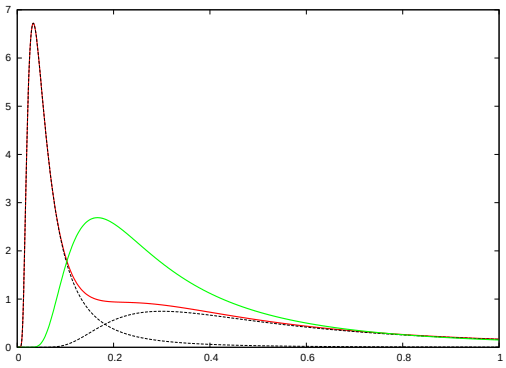


(j) RCORR:GE-AA (k) RCORR:C-AA (l) RCORR:SPY-BA

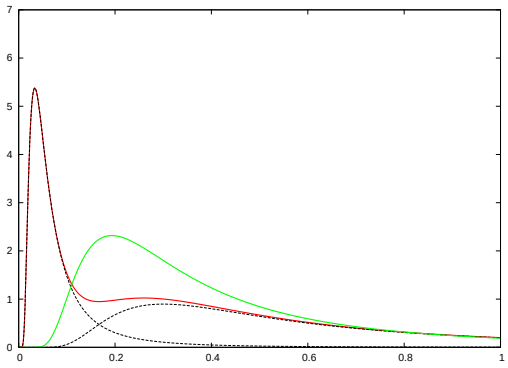


(m) RCORR:GE-BA (n) RCORR:C-BA (o) RCORR: AA-BA

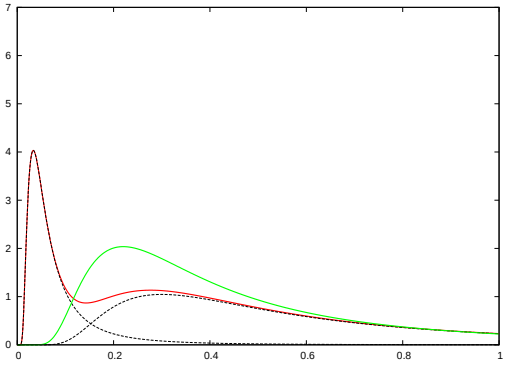
Figure 1: Realized Variances (RV) and Realized Correlations (RCORR) 1998-2007



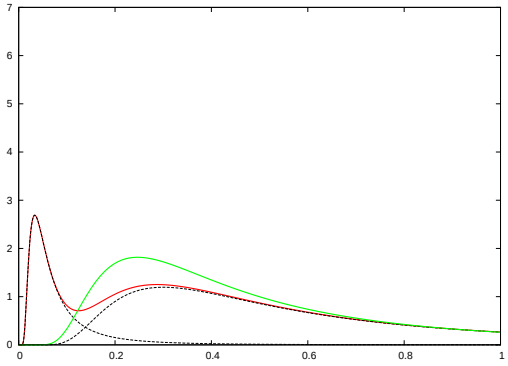
(a) $\omega_1 = 0.5$



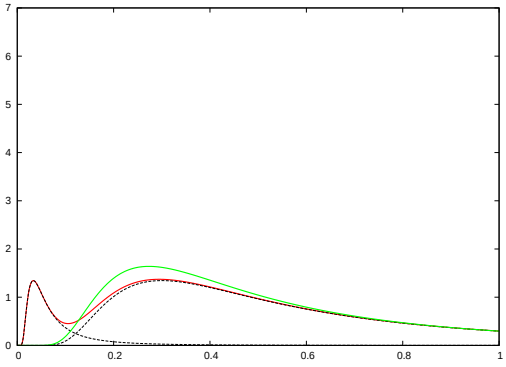
(b) $\omega_1 = 0.6$



(c) $\omega_1 = 0.7$



(d) $\omega_1 = 0.8$



(e) $\omega_1 = 0.9$

Figure 2: Densities of Two-component Mixture

The red solid line is the density of the two-component mixture of inverse-Gamma $\omega_1 \text{IG}(2, 0.9) + (1 - \omega_1)\text{IG}(2, 0.1)$. The black dotted line is the component density of the mixture scaled by weight (ω_1 or $1 - \omega_1$). The green solid line is the density of the inverse-Gamma $\text{IG}(2, 0.9\omega_1 + 0.1(1 - \omega_1))$, which has the same mean as the mixture.

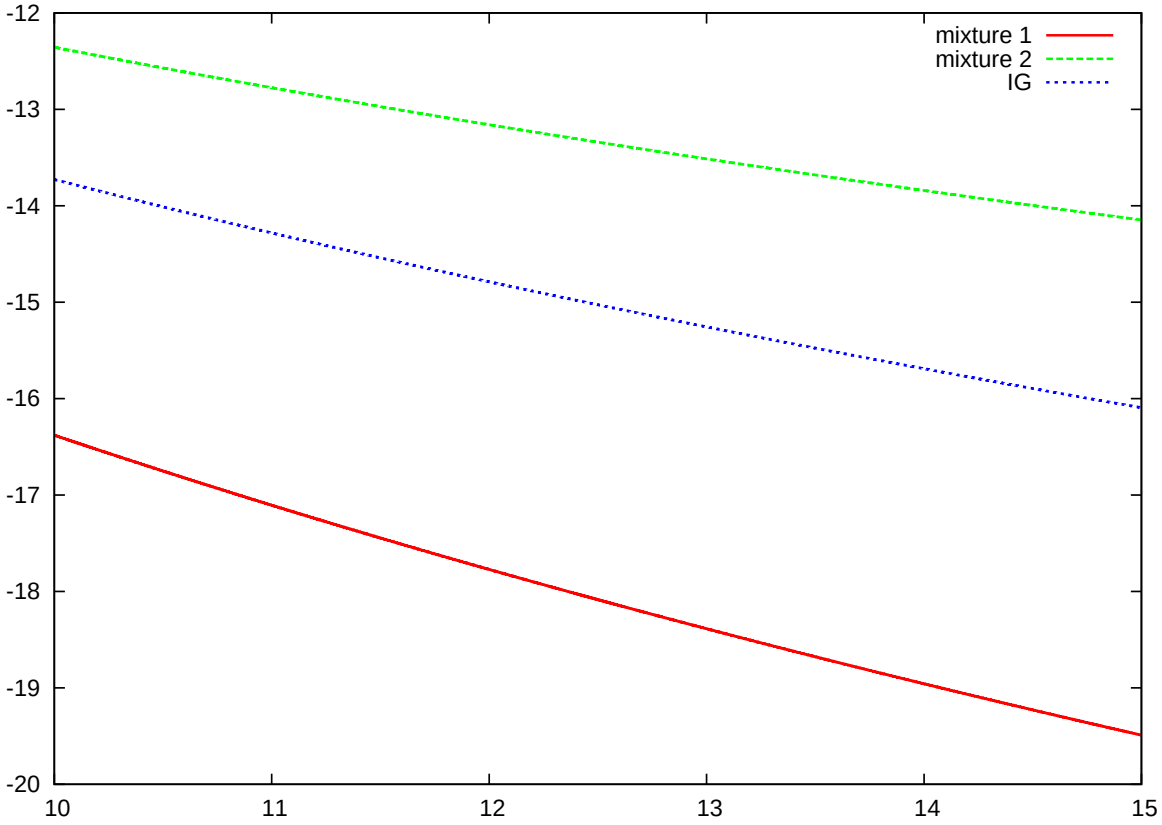


Figure 3: Log-density for the Tail

This figure displays the log-density of: mixture 1, $0.5\text{IG}(7,2) + 0.5\text{IG}(7,4)$; mixture 2, $0.5\text{IG}(3.5,1) + 0.5\text{IG}(11,6)$; and $\text{IG}(5,2)$.

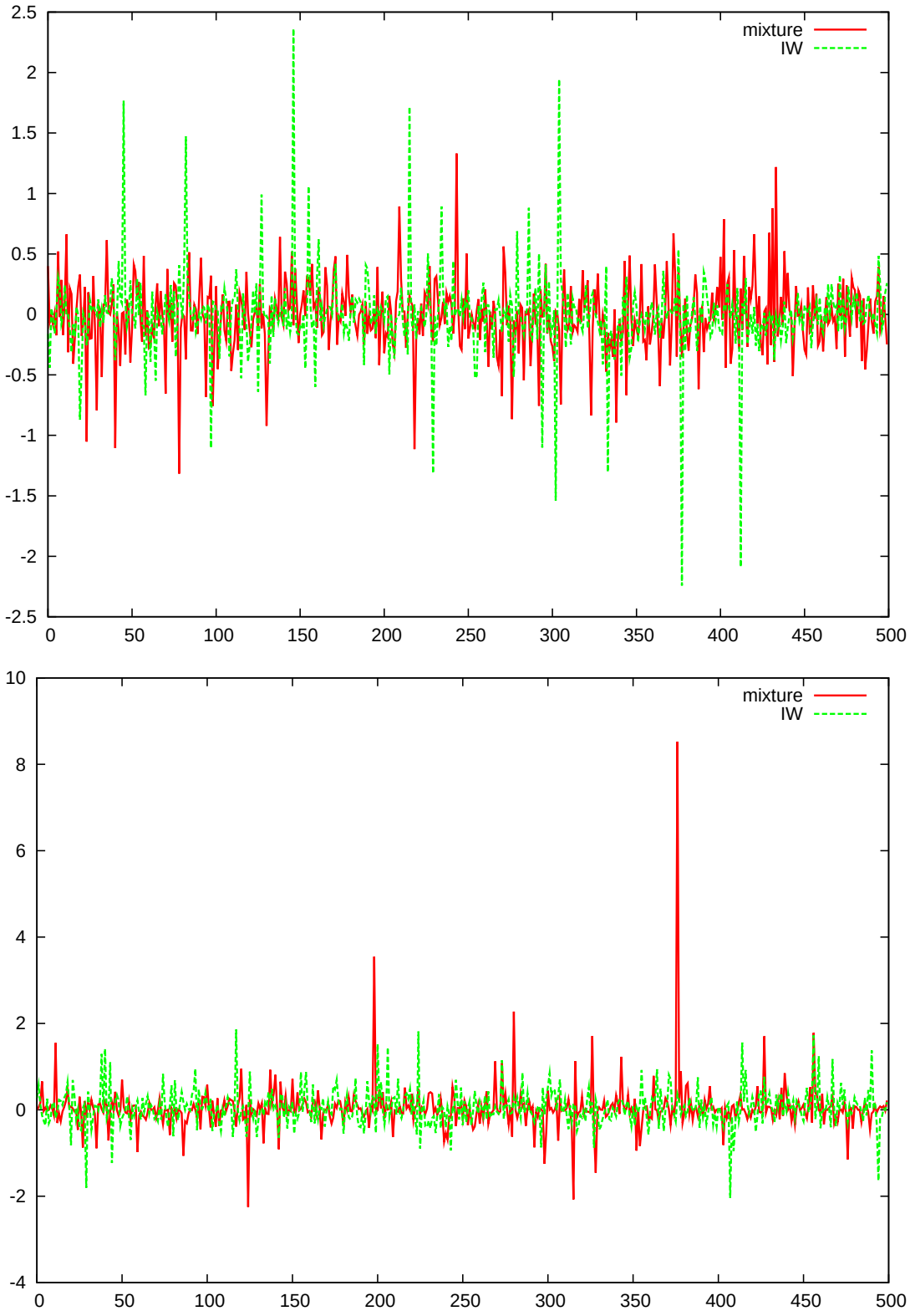


Figure 4: Simulated Data on Covariances

The top panel displays the simulated values from the off-diagonal of $\Sigma_t \sim 0.5\text{IW}_2(15, 10I) + 0.5\text{IW}_2(15, \sqrt{212}I)$ and $\Sigma_t \sim \text{IW}_2(7, 2I)$. The bottom panel displays the simulated values from the off-diagonal of $\Sigma_t \sim 0.5\text{IW}_2(6, I) + 0.5\text{IW}_2(8, \sqrt{24}I)$ and $\Sigma_t \sim \text{IW}_2(10, 7I)$. In each panel the covariance from each model has the same mean and variance.

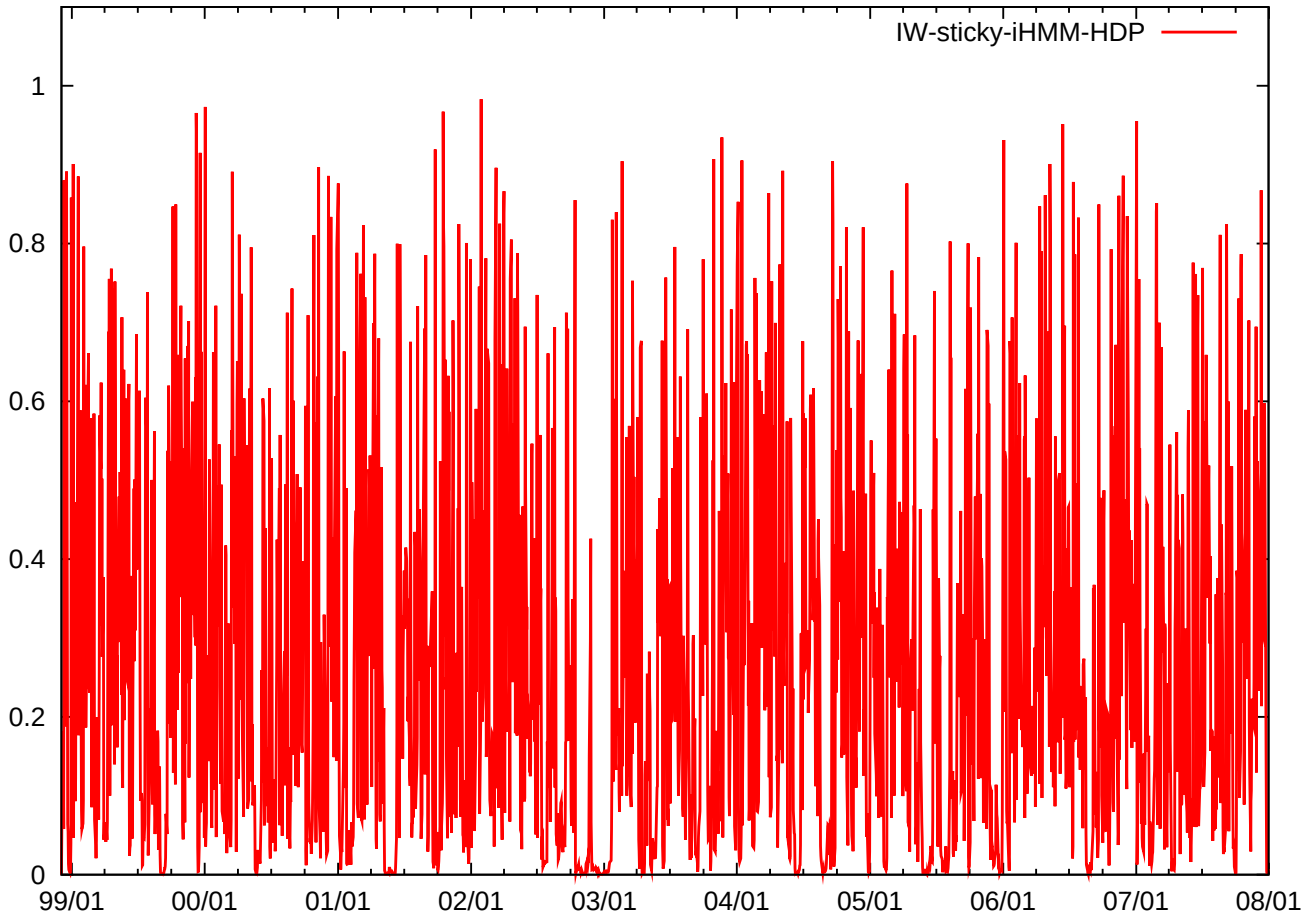
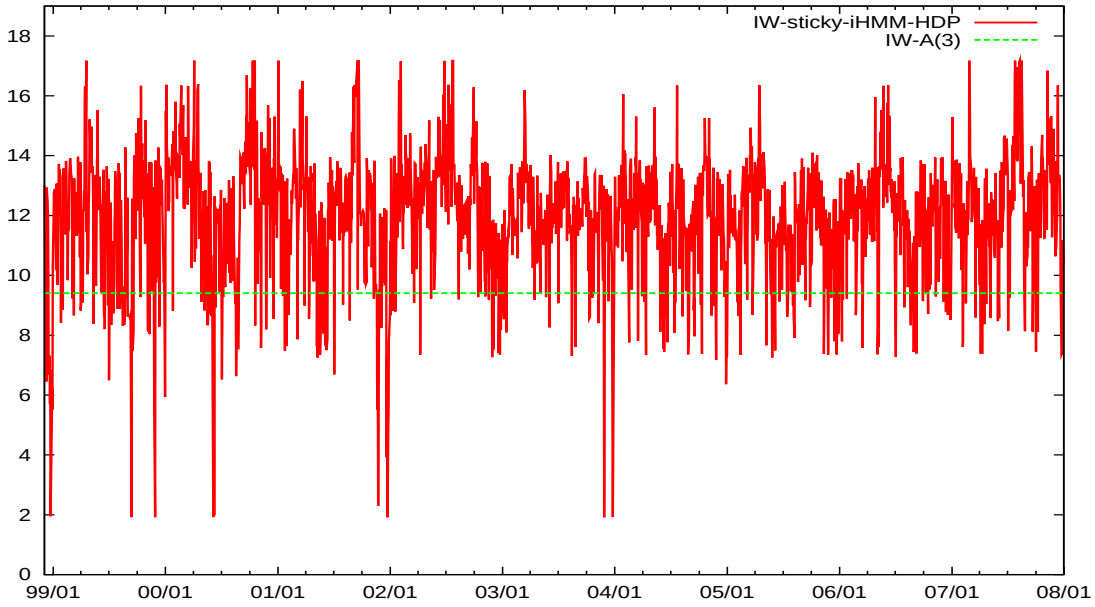
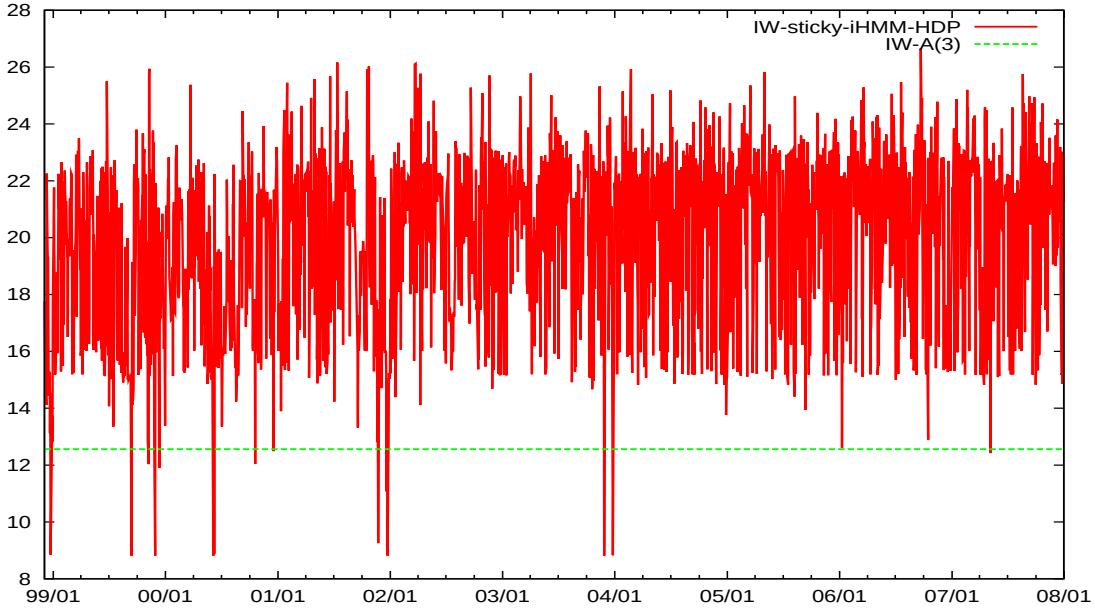


Figure 5: $P(s_t \neq s_{t-1}|\Sigma_{1:T})$

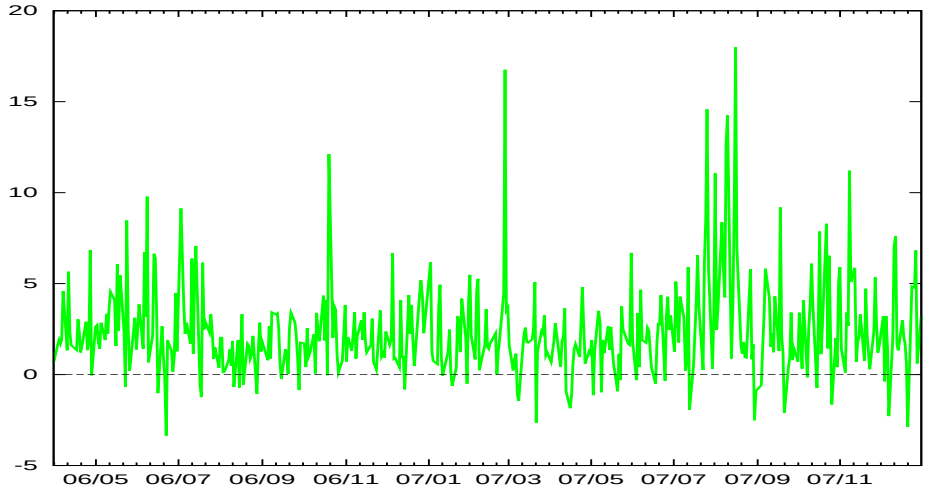


(a) $E[\log(|(\nu_{z_t} - k - 1)A_{z_t}|)|\Sigma_{1:T}]$

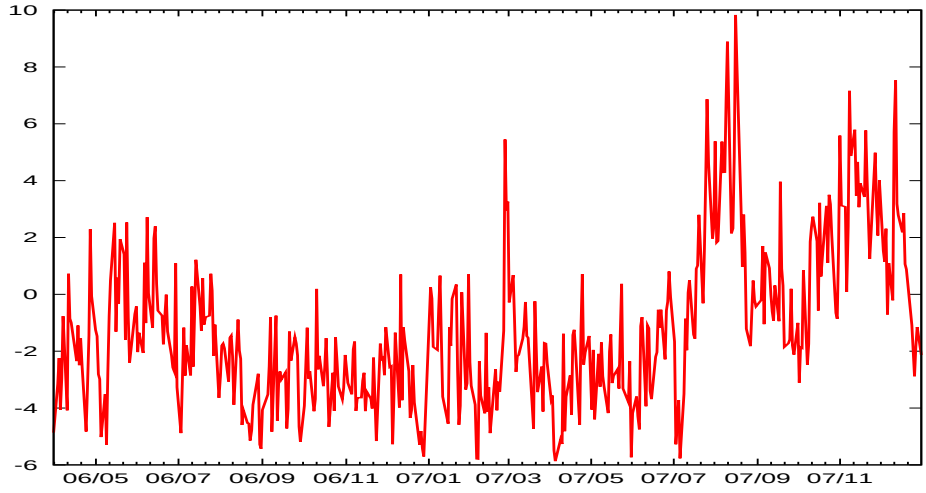


(b) $E[v_{z_t}|\Sigma_{1:T}]$

Figure 6: Comparison of IW-sticky-iHMM-HDP with IW-A(3)



(a) Difference in Log-predictive Likelihoods: IW-DPM - IW-A(3)



(b) Log determinants of RCOV

Figure 7: Difference in Log-predictive Likelihoods for RCOV

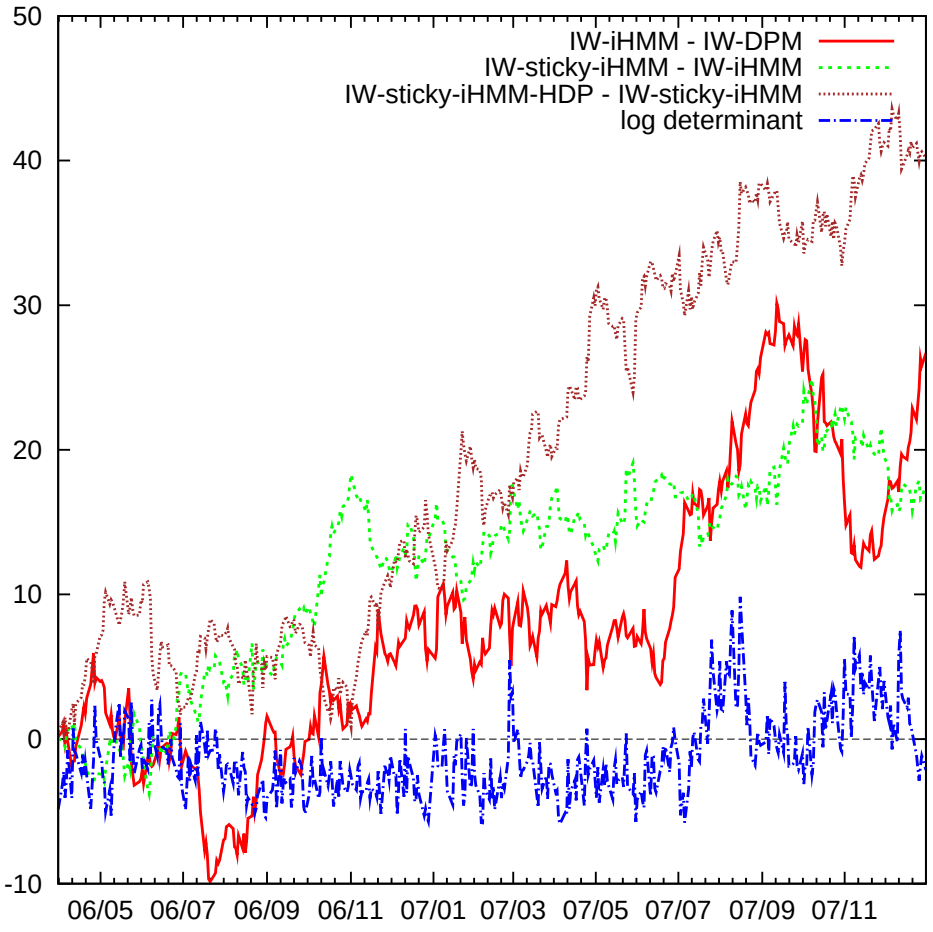
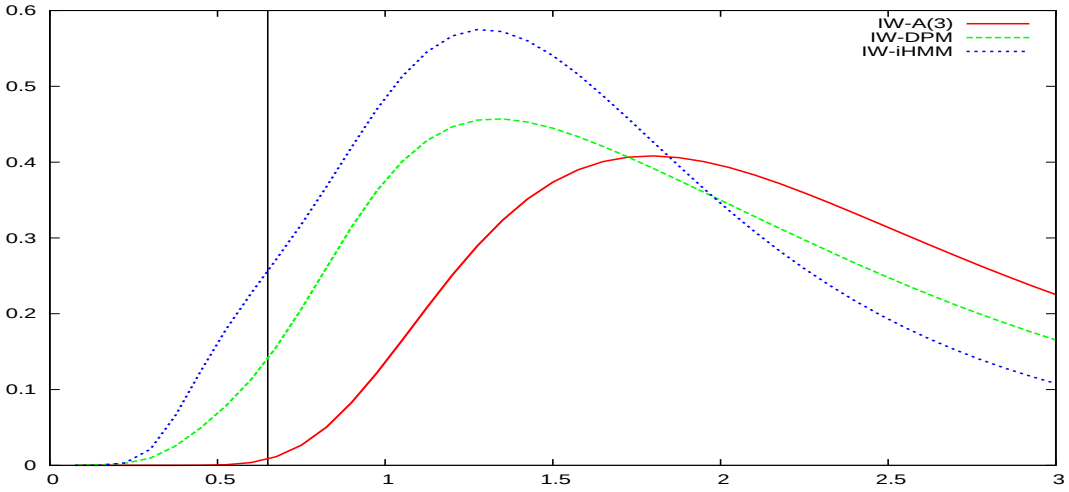
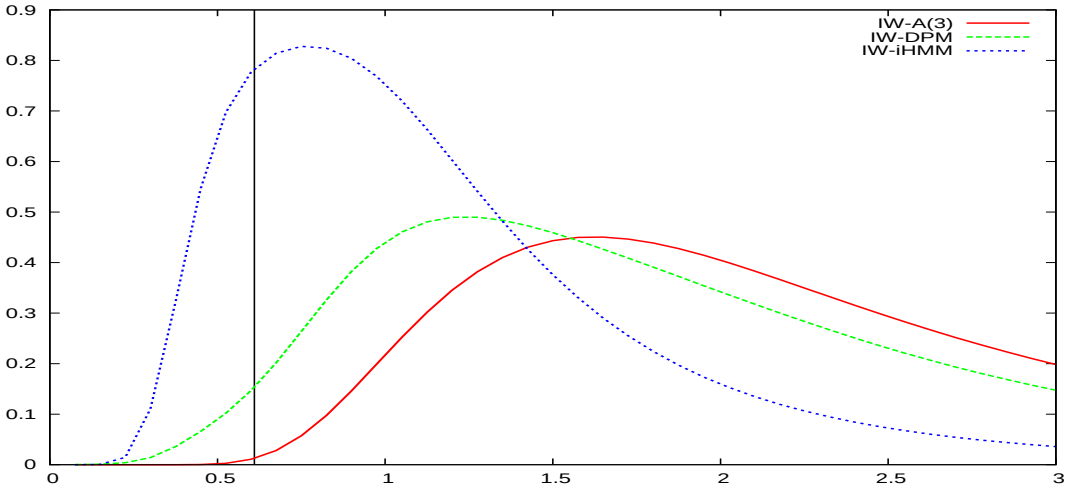


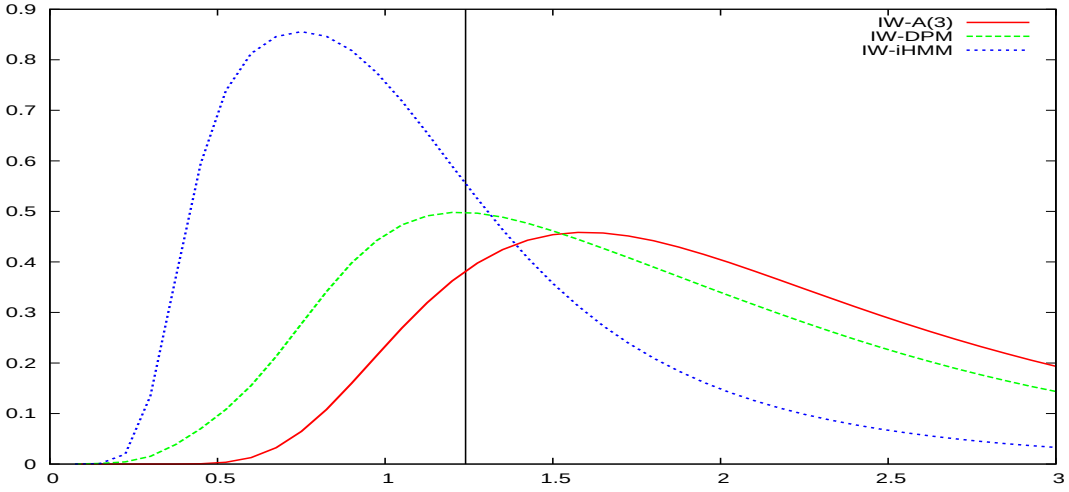
Figure 8: Cumulative Difference in Log-predictive Likelihoods for RCOV



(a) Density plots: $t = 2007/08/24$



(b) Density plots: $t = 2007/08/27$



(c) Density plots: $t = 2007/08/28$

Figure 9: Densities of Realized Variance from an Equally-weighted Portfolio