Bayesian Parametric and Semiparametric Factor Models for Large Realized Covariance Matrices

Jin, Xin and Maheu, John M and Yang, Qiao

Shanghai University of Finance and Economics, McMaster University, ShanghaiTech University

12 October 2017

Bayesian Parametric and Semiparametric Factor Models for Large Realized Covariance Matrices*

Xin Jin† John M. Maheu‡ Qiao Yang§

September 2017

Abstract

This paper introduces a new factor structure suitable for modeling large realized covariance matrices with full likelihood based estimation. Parametric and nonparametric versions are introduced. Due to the computational advantages of our approach we can model the factor nonparametrically as a Dirichlet process mixture or as an infinite hidden Markov mixture which leads to an infinite mixture of inverse-Wishart distributions. Applications to 10 assets and 60 assets show the models perform well. By exploiting parallel computing the models can be estimated in a matter of a few minutes.

JEL Classification: G17, C11, C14, C32, C58

key words: infinite hidden Markov model, Dirichlet process mixture, inverse-Wishart, predictive density, high-frequency data

*We are grateful for helpful comments from participants at the CFIRM conference Western University and the RCEA Bayesian Econometric Workshop University of Melbourne. Maheu thanks SSHRC for financial support.

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

‡corresponding author: DeGroote School of Business, McMaster University, 1280 Main Street West, Hamilton, ON, Canada, L8S4M4, maheujm@mcmaster.ca

§School of Entrepreneurship and Management, ShanghaiTech University, yangqiao@shanghaitech.edu.cn

1 Introduction

Modeling realized covariance (RCOV) matrices constructed from high-frequency data offers considerable improvements over conventional multivariate GARCH and stochastic volatility models.[1] Besides providing an accurate measure of ex post covariation that is observable, time-series methods can be applied directly to RCOV data to capture their conditional distribution. However, RCOV matricies are positive definite and present unique challenges to time-series modeling. This paper introduces a new factor structure that can be used in parametric (inverse-) Wishart models as well as infinite mixtures models for RCOV matrices.

The initial literature on modeling RCOV focused on capturing the time-series structure through different parametric distributions such as the Wishart, non-central Wishart and inverse-Wishart distributions (Gourieroux et al. 2009, Golosnoy et al. 2012, Asai & So 2013, Jin & Maheu 2013, Yu et al. 2017). Decompositions of the RCOV matrix so that standard time-series methods can be applied are pursued in Bauer & Vorkink (2011), Chiriac & Voev (2011) and Cech & Barunik (2017). Another branch of the literature links RCOV to multivariate GARCH models in Noureldin et al. (2012) and Hansen et al. (2014). The strong persistence patterns in RCOV matrix elements are recognized in Bauwens et al. (2016, 2017) while the importance of fatter tails is shown in Jin & Maheu (2016) and Opschoor et al. (2017).

Applications of factors methods which should be natural for the large dimensions involved are complicated by the positive definite matrix restriction. The approach by Tao et al. (2011) and extensions in Shen et al. (2015) and Asai & McAleer (2015) decompose the RCOV matrix in a similar fashion to Engle et al. (1990). Asai & McAleer (2015) model the decomposed factor in a number of ways including time-series models with long-memory, asymmetric effects and as a conditional autoregressive Wishart model. Shen et al. (2015) focus on a diagonal model of the latter Wishart specification. Sheppard & Xu (2014) propose a GARCH type factor model that incorporates RCOV information.

Our approach differs in several respects. First, we work with a factor structure inside an dynamic inverse-Wishart model and extend it to infinite mixture models. As such, the predictive distributions of both RCOV and returns are fully specified given parameter values. This leads us to move beyond model assessment that focuses on point forecasts (predictive mean) and to comparisons that evaluate the relative accuracy of the whole distribution

[1] RCOV models are easier to estimate than stochastic volatility specifications since *volatility* is observable. For instance, econometric forecasting gains are demonstrated in Golosnoy et al. (2012), Asai & McAleer (2015) and Jin & Maheu (2013, 2016) while improvements in portfolio choice are found in Fleming et al. (2003), Jin & Maheu (2013) and Callot et al. (2017).

through density forecasts of RCOV. Due to the analytic convenience of our model this extends to a known predictive density for returns as well and allows for model assessment based on returns.

The nonparametric approach of Jin & Maheu (2016) is based on time-varying mixtures of inverse-Wishart distributions. This likelihood approach is very flexible and the empirical applications show large improvements in forecast precision of daily RCOV matrices and daily returns for five assets. Due to the multivariate nature of the data, parametric distributions are unlikely to provide a good fit for RCOV data. Mixture models offer a tractable approach to leverage our knowledge from parametric approaches to span the complex unknown distributions of RCOV matrices. Jin & Maheu (2016) was the first paper to introduce mixture modeling to RCOV data. Although feasible for small dimensions this approach is not immediately applicable to larger systems.

The purpose of this paper is to extend the nonparametric methods of Jin & Maheu (2016) to a factor setting capable of modeling larger RCOV matrices. We begin by proposing a factor structure for an inverse-Wishart distribution which we extend to a mixture setting. To this end we design a Dirichlet process mixture (DPM) model and an infinite hidden Markov model (IHMM) that operate on a smaller factor dimension than the data dimension. Both of these approaches are based on countably infinite mixtures. The former has fixed weights while the latter has time-varying weights in the mixture. There are several computational benefits to this approach. First, computation of the data density is significantly reduced using the factor structure. Second, mixture models from a Bayesian posterior sampling perspective can easily take advantage of parallel computing. Conditional on the state variable that assigns observations to a component in the discrete mixture, sampling parameters of each component can be done independently. Finally, the factor approach could be applied to the other inverse-Wishart and Wishart based models in the literature.

Using inverse-Wishart or Wishart distributions as building blocks in a mixture is convenient. These distributions are closed under linear transformation. As a result, predictive inference is independent of asset order in the RCOV matrix. That is, we obtain the same predictive distribution subject to a permutation matrix for different asset orderings in RCOV matrices. This applies to predictive distributions of RCOV and returns. Moreover, assuming a multivariate normal distribution for returns given RCOV results in a marginal distribution of returns that is a mixture of Student-t distributions.

The trade-off of using a factor structure against more highly parameterized models is measured first in a 10 asset application. Generally, the full DPM and IHMM versions of the

model perform best but the nonparametric factor models are not far behind. Moving to a larger 60 asset application the full DPM and IHMM specification are not feasible.

The IHMM factor model is the dominant specification when we consider density forecasts of RCOV matrices and return vectors, point forecasts of RCOV and the global minimum variance portfolio selection for 60 assets. A 5 to 10 factor dimension results in large improvements in forecast accuracy compared to a number of benchmarks. By keeping the factor dimension small we can exploit the benefits of infinite mixture models for modeling the conditional distribution of RCOV matrices and maintain reasonable computational times. For instance, all of the models have computing time less that 13 minutes for one full sample estimation using conventional desktop Intel Xeon hardware. The data processed are just over 4 million individual observations. The number of active clusters in the mixture is around 15 for most factor models. Thus, a time-varying mixture model with 15 components is sufficient to provide large gains in forecast precision.

This paper is organized as follows. The next section reviews the models introduced in Jin & Maheu (2016). Parametric factor models are discussed in Section 3 followed by their nonparametric extensions in Section 4. Benchmark models used for comparison are briefly reviewed in Section 5. Application to 10 asset RCOV data and 60 asset RCOV data are in Section 6 followed by the conclusion. An appendix collects additional posterior simulation details for estimation.

2 Review of Inverse Wishart models for RCOV

This section briefly reviews the parametric and semiparametric models for RCOV from Jin & Maheu (2016). In the following section we extend this to the factor setting.

Let $\Sigma_t,\ t = 1, 2, \ldots, T$ denote a time-series of $k \times k$ realized covariance matrices and define $\Sigma_{1:t} = \{\Sigma_1, \ldots, \Sigma_t\}$. The additive component model based on the inverse-Wishart

(IW) distribution defines the conditional distribution of Σ_t as:[2]

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

$\text{Wishart}_k^{-1}(.|\nu,(\nu-k-1)V_t)$ 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 $(\nu-k-1)V_t$.[3] The operator $\odot$ 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} component defined as the average of past Σ_t over ℓ_j observations and captures persistence in V_t. The first component is equal to Σ_{t-1} while for $j\geq 2$, each ℓ_j is a free parameter to be estimated. In this paper we restrict attention to three components, $M=3$. The conditional mean of Σ_t is

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

It is straightforward to define an analogous model replacing the inverse-Wishart with a Wishart distribution. Since Jin & Maheu (2016) find the inverse-Wishart versions empirically superior we focus on building models with that density. Nevertheless, all the forgoing analysis could be done with a Wishart distribution instead.

We implement RCOV targeting by setting $B_0=(\iota\iota'-B_1-\cdots-B_M)\odot\overline{\Sigma}$, where $\overline{\Sigma}$ is the sample mean of Σ_t and ι is a k vector of ones. This ensures that the long-run mean of Σ_t is equal to $\overline{\Sigma}$ and leads to improved forecasts. In estimation any posterior draw in which B_0 is not positive definite is rejected. In addition, to ensure the mean exists, draws in which

[2]In Jin & Maheu (2016) this model is labelled as IW-A(M). In the following we drop A(M) since it is common to all specifications.

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

any element of $\sum_{j=1}^{M} B_j$ is not less than 1 in modulus are rejected.

The likelihood function for the IW model is

$$\begin{aligned} f(\Sigma_{1:T}|\nu,\Theta) &= \prod_{t=1}^{T} f(\Sigma_t|\Sigma_{1:t-1},\nu,\Theta), \\ &= \frac{(\nu-k-1)^{\frac{Tk\nu}{2}}\prod_{t=1}^{T}|V_t|^{\frac{\nu}{2}}\prod_{t=1}^{T}|\Sigma_t|^{-\frac{\nu+k+1}{2}}}{2^{\frac{T\nu k}{2}}\Gamma_k(\frac{\nu}{2})^T}\exp\left(-\frac{1}{2}(\nu-k-1)tr\left(\sum_{t=1}^{T}V_t\Sigma_t^{-1}\right)\right). \end{aligned} \tag{7}$$

Posterior simulation is conducted with a Metropolis-Hastings (MH) step using a joint random walk proposal for $b_1,\ldots,b_M$ and ν. 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 details of posterior simulation and model details are found in Jin & Maheu (2013, 2016).

The second class of model we will extend to a factor structure is based on a Dirichlet process (DPM) mixture. The generic model for the unknown conditional density of Σ_t takes the following form:

$$f(\Sigma_t|\Sigma_{1:t-1},\Theta,G) = \int h(\Sigma_t|\Sigma_{1:t-1},\Theta,\phi)G(d\phi), \tag{8}$$

$$G|G_0,\alpha \sim \mathrm{DP}(\alpha,G_0), \tag{9}$$

where $\mathrm{DP}(\alpha,G_0)$ denotes the Dirichlet process (DP) with $\alpha>0$ the precision parameter and base distribution 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 $h(\cdot|\cdot)$.

From the constructive definition of the DP (Sethuraman 1994) the model is a countably-infinite mixture defined as:

$$f(\Sigma_t|\Sigma_{1:t-1},\Theta,\Omega,\Phi) = \sum_{j=1}^{\infty}\omega_j h(\Sigma_t|\Sigma_{1:t-1},\Theta,\phi_j), \tag{10}$$

$$\omega_j = v_j\prod_{l<j}(1-v_l), \quad v_j \overset{iid}{\sim} \mathrm{Beta}(1,\alpha), \quad j=1,2,\ldots, \tag{11}$$

$$\phi_j \overset{iid}{\sim} G_0, \quad j=1,2,\ldots, \tag{12}$$

where $\Omega=\{\omega_j\}_{j=1}^{\infty}$, $\Phi=\{\phi_j\}_{j=1}^{\infty}$. $G=\sum_{j=1}^{\infty}\omega_j\delta_{\phi_j}$, where δ_{ϕ_j} is a point mass at ϕ_j. The random atoms ϕ_j are i.i.d. draws from the base distribution G_0, and the random weights ω_j

are constructed using i.i.d. beta variates v_j. In the following we abuse notation somewhat and let $\mathbf{SBP}(\alpha)$ denote the stick-breaking construction of the weights as well as a distribution with support on the natural numbers, $\Omega \sim \mathbf{SBP}(\alpha)$.

It is natural to replace the kernel $h(\Sigma_t|\Sigma_{1:t-1}, \Theta, \phi_j)$ with that of the IW model discussed above, resulting in the IW-DPM model

$$f(\Sigma_t|\Sigma_{1:t-1}, \Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \omega_j \text{Wishart}_k^{-1}(\Sigma_t|\nu_j, (\nu_j - k - 1)V_t^{1/2} A_j (V_t^{1/2})'), \quad (13)$$

$$V_t = B_0 + \sum_{j=1}^{M} B_j \odot \Gamma_{t-1,\ell_j}. \quad (14)$$

$\Omega \sim \mathbf{SBP}(\alpha)$ and A_j's are $k \times k$ symmetric positive-definite matrices and $\phi_j \equiv (\nu_j, A_j)$. $V_t^{1/2}$ denotes the Cholesky factor of V_t and the definitions of the remaining terms are identical to the IW parametric model. Each component of the mixture j has a different scale matrix, $(\nu_j - k - 1)V_t^{1/2} A_j (V_t^{1/2})'$, which is positive definite by construction, and a different degree of freedom ν_j. This is a rich functional form since any symmetric positive-definite matrix can be represented by $V_t^{1/2} A_j (V_t^{1/2})'$ by the appropriate choice of A_j. In addition, the model nests the parametric version when $\omega_j = 1$, $\omega_i = 0$, $i \neq j$ and $A_j = I$. The conditional mean takes the form,

$$E[\Sigma_t|\Sigma_{1:t-1}, \Theta, \Omega, \Phi] = \sum_{j=1}^{\infty} \omega_j V_t^{1/2} A_j (V_t^{1/2})' = V_t^{1/2} \left[\sum_{j=1}^{\infty} \omega_j A_j \right] (V_t^{1/2})'. \quad (15)$$

The second nonparametric specification we will extend to a factor structure is an infinite hidden Markov (IHMM) model. Unlike the DPM model which assumes the latent features of the unknown distribution are fixed over time the IHMM allows for change according to a first order Markov chain. The IHMM is constructed from the hierarchical Dirichlet process (HDP) prior of Teh et al. (2006). To allow for estimation of self-transitions we focus on the sticky version of the IHMM introduced by Fox et al. (2011). Jin & Maheu (2016) propose

the following generic sticky IHMM model for Σ_t:

$$\begin{aligned}
\boldsymbol{\pi}_0|\alpha &\sim \mathbf{SBP}(\alpha), && (16)\\
\boldsymbol{\pi}_i|\boldsymbol{\pi}_0,\beta,\kappa &\sim \mathrm{DP}\left(\beta+\kappa, \frac{\beta\boldsymbol{\pi}_0+\kappa\delta_i}{\beta+\kappa}\right), && (17)\\
\phi_j &\overset{iid}{\sim} G_0, \quad j=1,2,\ldots, && (18)\\
s_t|s_{t-1}=i,\Pi &\sim \boldsymbol{\pi}_i, \quad i=1,2,\ldots, && (19)\\
\Sigma_t|\Sigma_{1:t-1},\Theta,\Phi,s_t &\sim \mathcal{H}(\Sigma_t|\phi_{s_t}), && (20)
\end{aligned}$$

where 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 $\boldsymbol{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \ldots)$ and is the i^{th} row of Π. The conditional distribution of Σ_t is governed by the distribution $\mathcal{H}(\Sigma_t|\phi_{s_t})$ given s_t and ϕ_{s_t}. Each row of the transition matrix $\boldsymbol{\pi}_i$ is generated from an associated stick breaking process that is centered on $\frac{\beta\boldsymbol{\pi}_0+\kappa\delta_i}{\beta+\kappa}$. The term $\beta\boldsymbol{\pi}_0 + \kappa\delta_i$ means that the amount $\kappa \geq 0$ is added to the i^{th} component of $\beta\boldsymbol{\pi}_0$. β controls how close each row is to the base distribution $\boldsymbol{\pi}_0$ while a larger κ increases the prior probability of self-transition and a $\kappa = 0$ reverts to the benchmark non-sticky IHMM specification.

This model admits a density very much like (13) with the important difference that the fixed weights ω_j, are replaced by time-varying weights as in

$$\begin{aligned}
f(\Sigma_t|\Theta,\Pi,\Phi,s_{t-1}) &= \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} h(\Sigma_t|\Sigma_{1:t-1},\Theta,\phi_{s_t}), && (21)\\
\pi_{i,j} &= \hat{\pi}_{i,j}\prod_{l=1}^{j-1}(1-\hat{\pi}_{i,l}), && (22)\\
\hat{\pi}_{i,j} &\overset{iid}{\sim} \mathrm{Beta}(\beta\pi_{0j}+\kappa\delta_i, \beta(1-\sum_{l=1}^{j}\pi_{0l})+\kappa\mathbf{1}(j<i)), && (23)
\end{aligned}$$

where $h(\Sigma_t|\Sigma_{1:t-1},\Theta,\phi_{s_t})$ is the density associated with the distribution $\mathcal{H}(\Sigma_t|\phi_{s_t})$. As in the IW-DPM model $h(\Sigma_t|\Sigma_{1:t-1},\Theta,\phi_j) \equiv \mathrm{Wishart}_k^{-1}(\Sigma_t|\nu_j, (\nu_j-k-1)V_t^{1/2}A_j(V_t^{1/2})')$. Full details of posterior simulation for these models can be found in Jin & Maheu (2016).

Let r_t denote the $k \times 1$ daily return vector. With the assumption of[4]

$$r_t|\Sigma_t \sim N(0,\Sigma_t) \quad (24)$$

[4]In our empirical analysis we use demeaned returns but a conditional mean could be included.

this leads to mixtures of Student-t distributions for returns. For instance, in the case of the IW-IHMM model, given s_{t-1} and integrating out Σ_t gives

$$f(r_t|\Sigma_{1:t-1},\nu,\Theta,s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \mathrm{St}_k \left(r_t|0, \frac{\nu_{s_t}-k-1}{\nu_{s_t}-k+1} V_t^{1/2} A_{s_t} (V_t^{1/2})', \nu_{s_t}-k+1 \right), \quad (25)$$

where $\mathrm{St}_k(r_t|\mu,V,\nu)$ denotes a multivariate Student-t density of dimension k with mean vector μ, scale matrix V and degree of freedom ν.

Jin & Maheu (2016) document huge improvements in density forecasts and point forecasts for Σ_t and returns in moving from the parametric IW specifications to the nonparametric versions IW-DPM and IW-IHMM. Nevertheless, the models are highly parameterized, particularly the nonparametric models. For instance, the IW model contains $3k+3$ parameters with RCOV targeting which is quite manageable. However, if the nonparametric model uses an average of K components, there are a minimum of $3k+2+K(1+1+k(k+1)/2)$ parameters for the IW-DPM model and $3k+2+K(1+K+k(k+1)/2)$ parameters for the IW-IHMM model. Each of these parameters must be sampled from the posterior density. The second computational issue is the need to efficiently evaluate the likelihood function of each model. The main bottleneck is the evaluation of $\prod_{t=1}^{T} |V_t|^{\frac{\nu}{2}}$ which is generally accomplished through a matrix decomposition such as the Cholesky which is an $O(k^3)$ computational cost for each V_t. For large realized covariance matrices estimation of these models will be infeasible. In the following sections we introduce factor models that exploit the benefits of these models but minimize parameters and speed up likelihood evaluation.

3 Parametric Factor Models of RCOV

An important property of the family of (inverse-) Wishart distributions is that they are closed under linear transformations. That is, linear transformations of (inverse-) Wishart distributed matrices are themselves (inverse-) Wishart distributed.[5]

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

To carry out our factor approach, instead of modeling the dynamics of the original RCOV itself, we first apply a linear transformation to Σ_t, the dynamics of which are then modeled using an inverse-Wishart distribution with a factor structure. The dynamics of the raw

[5]See Press (2012).

RCOV are readily available according to Property 1 by applying the inverse transformation, and forecasts of future Σ_t (and returns) can be obtained similarly.

Let $V = \mathrm{E}(\Sigma_t)$ denote the unconditional mean of Σ_t. Applying a spectral decomposition to V gives

$$V = WDW' = \sum_{i=1}^{k} d_i w_i w_i', \tag{26}$$

where $D = \mathrm{diag}\{d_1 \geq d_2 \geq \ldots \geq d_k > 0\}$ is a diagonal matrix with $d_1, d_2, \ldots d_k$ being the eigenvalues of V, and $W = (w_1, w_2, \ldots, w_k)$ is a $k \times k$ orthogonal matrix with the column w_i being the corresponding eigenvector of d_i and satisfying $W'W = WW' = I$.[6] Define the orthogonally transformed Σ_t denoted as Σ_t^* by[7]

$$\Sigma_t^* = W'\Sigma_t W. \tag{27}$$

The uniqueness of Σ_t^* is determined by the uniqueness of W. In particular, the order/positions of the elements of Σ_t^* are determined by the order of the column vectors $w_1, \ldots, w_k$ in W, which corresponds to the order of $d_1, \ldots, d_k$ listed in the diagonal of D. This is easy to see since the (i, j) element of $\Sigma_t^* \equiv (\sigma_{t,ij}^*)$ is $\sigma_{t,ij}^* = w_i'\Sigma_t w_j$. Note the unconditional mean of Σ_t^* is the diagonal matrix D by definition:

$$E(\Sigma_t^*) = E(W'\Sigma_t W) = W'E(\Sigma_t)W = W'VW = W'WDW'W = D. \tag{28}$$

So regardless of the order of w_i, the off-diagonal elements of Σ_t^* always have zero unconditional mean $E(\sigma_{t,ij}^*) = E(w_i'\Sigma_t w_j) = 0, i \neq j$, while the diagonal elements have d_i as their unconditional mean $E(\sigma_{t,ii}^*) = E(w_i'\Sigma_t w_i) = d_i$.

In this paper we sort d_i along the diagonal of D from top-left to bottom right(and hence w_i in W from left to right) according to the descending order.[8] Under this ordering scheme,

[6] If the eigenvalues are distinct w_i is unique up to sign. If there are repeated eigenvalues then W is not unique but this causes no issue for inference.

[7] A similar transformation is used in Noureldin et al. (2014) in the context of multivariate GARCH modeling.

[8] An alternative sorting would be according to the variance. Let g_i denote the unconditional variance of the diagonal elements of Σ_t^*,

$$g_i = \mathrm{Var}(\sigma_{t,ii}^*) = \mathrm{Var}(w_i'\Sigma_t w_i), \tag{29}$$

which is like the variance of the realized variance of a portfolio but with weight vector w_i and condition $w_i'w_i = 1$. Under this ordering scheme, the resulting diagonal elements of Σ_t^* are decreasing in the unconditional variance. Our empirical studies indicate sorting D based on d_i was preferred.

the resulting diagonal elements of Σ_t^* are decreasing in the unconditional mean, which will be convenient later when introducing the factor structure as it will operate on a block of Σ_t^* associated with the largest d_i values. In addition, our analysis is invariant to the asset order in the Σ_t matrix. If the asset order is permuted to the new RCOV matrix $\hat{\Sigma}_t = P\Sigma_t P'$, with $E[\hat{\Sigma}] = \hat{W}\hat{D}\hat{W}'$ then $\hat{D} = D$ and $\hat{W} = PW$, where P is the permutation matrix.

Our factor approach will model the dynamics of Σ_t through Σ_t^*. As in the IW model for Σ_t, an inverse-Wishart distribution is assumed for the conditional distribution of Σ_t^*, however, the conditional mean of Σ_t^* is restricted to a special form to allow for a factor structure. Partition Σ_t^* as follows

$$\Sigma_t^* = \begin{pmatrix} \Sigma_{t,11}^* & \Sigma_{t,21}^{*\prime} \\ \Sigma_{t,21}^* & \Sigma_{t,22}^* \end{pmatrix}, \tag{30}$$

where $\Sigma_{t,11}^*$ is $k_1 \times k_1$, $\Sigma_{t,22}^*$ is $k_2 \times k_2$, and k_1, k_2 satisfy $k_1 > 0, k_2 \geq 0, k_1 + k_2 = k$.

3.1 Block-Diagonal Factor Model (IW-F)

This section introduces a factor model based on the inverse-Wishart distribution. As before, the factor model applies to the Wishart distribution as well, although we will focus attention on the inverse-Wishart version. In the IW-F model the conditional distribution of Σ_t^* is specified as follows:

$$f(\Sigma_t^*|\Sigma_{1:t-1}^*, \nu, C, \Theta) = \text{Wishart}_k^{-1}(\Sigma_t^*|\nu, (\nu-k-1)V_t), \tag{31}$$

$$V_t = \begin{pmatrix} V_t^* & 0 \\ 0 & C \end{pmatrix}, \tag{32}$$

$$V_t^* = B_0 + \sum_{j=1}^{M} B_j \odot \Gamma_{t-1,\ell_j}^*, \tag{33}$$

$$\Gamma_{t-1,\ell_j}^* = \frac{1}{\ell_j} \sum_{i=1}^{\ell_j} \Sigma_{t-i,11}^*. \tag{34}$$

In this model the time-varying V_t^* operates on the lower dimension $k_1 \times k_1$ matrix with associated lower dimension parameter matrices $B_0, B_1, \ldots, B_M$, and $B_j = b_j b_j', j = 1, \ldots, M$. In general $C = \text{diag}\{c_1, \ldots, c_{k_2}\}$ is a $k_2 \times k_2$ matrix.[9]

V_t^* can be viewed as the set of dynamic factors, which contains $k_1(k_1+1)/2$ unique scalar

[9] Alternatively C could be specified as a full positive definite matrix.

elements and satisfy

$$E(\Sigma^*_{t,11}|\Sigma^*_{1:t-1}) = V^*_t. \tag{35}$$

Meanwhile, $\Sigma^*_{t,11}$ is the observed counterpart of V^*_t, and conditionally follows an inverse-Wishart with dimension $k_1 \times k_1$,

$$\Sigma^*_{t,11}|\Sigma^*_{1:t-1}, \nu, \Theta \sim \text{Wishart}^{-1}_{k_1}(\nu - k_2, (\nu - k - 1)V^*_t). \tag{36}$$

On the other hand, C contains the static scalar factors $c_1, \ldots, c_{k_2}$ along the diagonal and has zero everywhere else. As a result, the observed static factors appearing on the diagonal elements of $\Sigma^*_{t,22}$ all follow time-invariant inverse-Gamma distributions,

$$\sigma^*_{t,ii}|\Sigma^*_{1:t-1}, \nu, \Theta \;\sim\; \text{Gamma}^{-1}\left(\frac{\nu - k + 1}{2}, \frac{\nu - k - 1}{2} c_{i-k_1}\right), \quad i = k_1 + 1, \ldots, k. \tag{37}$$

In particular, they have both conditional and unconditional mean equal to the respective c_j, $E(\sigma^*_{t,ii}|\Sigma^*_{1:t-1}) = E(\sigma^*_{t,ii}) = c_{i-k_1}, \quad i = k_1 + 1, \ldots, k$.

The unconditional moment condition for Σ^*_t in (28) requires all the off-diagonal elements to have zero unconditional mean. The IW-F model allows the off-diagonal elements of $\Sigma^*_{t,11}$ to have non-zero conditional means, which depend on their own histories and hence time-varying. RCOV targeting can be implemented in model estimation to ensure the off-diagonal elements of $\Sigma^*_{t,11}$ have zero unconditional mean to satisfy (28). Meanwhile, the factor model still imposes zero conditional mean for off-diagonal blocks, $\Sigma^*_{t,21}$ and $\Sigma^*_{t,21}{}'$, and off-diagonal elements of $\Sigma^*_{t,22}$. This is a stronger restriction than (28) but the trade-off here is that we can retain the factor structure which at the same time alleviates computation burden in high dimensional cases.

With these assumptions the total number of parameters is $3k_1 + 3$ with RCOV targeting. Besides reducing the number of parameters a potentially more important aspect of this model is the reduced computational burden in the likelihood evaluation. As discussed above the inverse-Wishart density requires a Cholesky decomposition to compute the determinant of V_t and the computational complexity is $O(k^3)$ but the factor structure reduces this to a Cholesky decomposition on V^*_t which is of $O(k_1^3)$ computations. This makes a significant difference in large k applications.

Properties of the inverse-Wishart distribution imply

$$\Sigma_t|V_t,\nu,\Theta \sim \text{Wishart}_k^{-1}(\nu,(\nu-k-1)WV_tW^{'}). \tag{38}$$

W can be interpreted as factor loadings and imply

$$E[\Sigma_t|V_t,\nu,\Theta] \;=\; WV_tW^{'} = W_1V_t^*W_1^{'} + W_2CW_2^{'}, \tag{39}$$

where $W=(W_1,W_2)$, W_1 is $k\times k_1$ and W_2 is $k\times k_2$.

More insight into the factor structure can be shown by linking returns to RCOV. In the following we set the mean of returns to zero and work with demeaned returns, however, all the results carry through with more general conditional mean dynamics such as an intercept or lagged returns. Assume

$$E(r_t|\Sigma_t,\mathcal{F}_{t-1})=0, \quad \text{Var}(r_t|\Sigma_t,\mathcal{F}_{t-1})=\Sigma_t, \tag{40}$$

where $\mathcal{F}_{t-1}=\{\Sigma_{1:t-1},r_{1:t-1}\}$ is the information set up to time $t-1$. The unconditional variance of r_t is V,

$$\text{Var}(r_t)=E(r_tr_t')=E(E(r_tr_t'|\Sigma_t))=E(\Sigma_t)=V. \tag{41}$$

The $t-1$ conditional variance of r_t is

$$\begin{aligned}\text{Var}(r_t|\mathcal{F}_{t-1}) &= E(r_tr_t'|\mathcal{F}_{t-1})=E(E(r_tr_t'|\mathcal{F}_{t-1},\Sigma_t)|\mathcal{F}_{t-1})=E(\Sigma_t|\mathcal{F}_{t-1})\\ &= WV_tW'\\ &= W_1V_t^*W_1'+W_2CW_2'.\end{aligned} \tag{42}$$

This shows the time $t-1$ conditional covariance matrix of r_t is exactly determined by a set of time-varying factors V_t^* and a constant set $\{c_j\}$ through transformation. To see this more clearly, define $r_t^*\equiv W'r_t$.[10] Then $\text{Var}(r_t^*|\Sigma_t)=W'\Sigma_tW=\Sigma_t^*$, and $\text{Var}(r_t^*)=D$. And it is easy to show that the $t-1$ conditional variance of r_t^* is V_t. Further partition $r_t^*=({r_{t,1}^*}',{r_{t,2}^*}')'$, where $r_{t,1}^*$ is $k_1\times 1$ and $r_{t,2}^*$ is $k_2\times 1$. This model imposes the restrictions

$$\text{Var}(r_{1,t}^*|\mathcal{F}_{t-1})=V_t^*,\ \text{Var}(r_{2,t}^*|\mathcal{F}_{t-1})=C,\ \text{Cov}(r_{1,t}^*,r_{2,t}^*|\mathcal{F}_{t-1})=\mathbf{0}_{[k_1\times k_2]}. \tag{43}$$

[10]Note this is not the usual portfolio with weights that sum to 1 but instead obey $\omega_i'\omega_i=1$.

Therefore, there exists two sets of portfolios with return vectors $r^*_{1,t}$ and $r^*_{2,t}$ that are uncorrelated with each other, the latter portfolio consisting of k_2 assets that are uncorrelated among themselves and homoskedastic. The portfolio $r^*_{1,t}$ consists of assets that are conditionally correlated in general.

3.2 Special Case: Diagonal Factor Model (IW-F-D)

One special case of note is when V^*_t is a diagonal matrix. In this specification (32) becomes

$$V_t = \begin{pmatrix} V^*_t & 0 \\ 0 & C \end{pmatrix} \tag{44}$$

$$V^*_t = \text{diag}\{b_0\} + \sum_{j=1}^{M} \text{diag}\{b_j\} \odot \Gamma^*_{t-1,\ell_j}. \tag{45}$$

This model (IW-F-D) has the same number of parameters as the block diagonal version but imposes the strong restriction that $E(\Sigma^*_{t,ij}|\mathcal{F}_{t-1}) = 0$, for $i \neq j$. The only source of time variation is from $V^*_{t,ii}$, $i = 1, \ldots, k_1$.

Despite the restriction on this model there are two situations in which it could be useful. The first is for a very large set of assets. This model completely removes the need to compute a Cholesky decomposition since the scale matrix is diagonal and the determinant and the trace are trivial to compute in the density function. A second setting in which the model may be useful is for data that, after the transformation (27), have little to no correlation amongst themselves representing a sparse diagonal RCOV matrix. In this case the zero restriction on the off diagonal of V^*_t conforms to the data.

3.3 Model inference

To implement the transformation in (27) we apply a spectral decomposition to the sample mean of Σ_t,

$$\overline{\Sigma} = WDW', \tag{46}$$

where $\overline{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} \Sigma_t$. Given W, Σ^*_t is constructed. We sort the diagonal elements of D and hence column vectors of W according to the descending order. As discussed above asset order in forming Σ_t does not matter.

For each of the models we implement RCOV targeting for B_0. For IW-F, we set $B_0 =$

$(\iota\iota' - B_1 - \cdots - B_M) \odot \overline{\Sigma}^*_{11}$ and for IW-F-D, we set $b_{0i} = (1 - b_{1i} - \cdots - b_{Mi})\overline{\Sigma}^*_{11,ii}, i = 1, \ldots, k_1$, where $\overline{\Sigma}^*_{11}$ is the sample mean of $\Sigma^*_{t,11}$. In the same spirit, C can be targeted at its sample counterpart by letting $C = \overline{\Sigma}^*_{22}$, where $\overline{\Sigma}^*_{22}$ is the sample mean of $\Sigma^*_{t,22}$ and, by construction, is diagonal. Inference on the other parameters is based on their posterior distribution.

We discuss posterior simulation for IW-F and note that the MCMC methods for IW-F-D are virtually the same. The joint posterior distribution is proportional to

$$p(\nu)p(\Theta)f(\Sigma^*_{1:T}|\nu, C, \Theta). \tag{47}$$

The likelihood, $f(\Sigma^*_{1:T}|\nu, C, \Theta)$, is identical to the likelihood of $f(\Sigma_{1:T}|\nu, C, \Theta, W)$ which follows from (38), since Σ_t and Σ^*_t differ by an orthogonal transformation. The block diagonal structure of the scale matrix in the Wishart or inverse-Wishart transition density is greatly beneficial for reducing the computational burden of evaluating the likelihood. Take the inverse-Wishart case as an example, the conditional density of Σ^*_t is

$$\begin{aligned} f(\Sigma^*_t|\Sigma^*_{1:t-1}, \nu, C, \Theta) &= \text{Wishart}_k^{-1}(\Sigma^*_t|\nu, (\nu-k-1)V_t) \\ &= \frac{(\nu-k-1)^{\frac{k\nu}{2}}|V_t|^{\frac{\nu}{2}}|\Sigma^*_t|^{-\frac{\nu+k+1}{2}}}{2^{\frac{\nu k}{2}}\Gamma_k(\frac{\nu}{2})} \exp\left(-\frac{1}{2}tr((\nu-k-1)V_t\Sigma^{*-1}_t)\right) \\ &= \frac{(\nu-k-1)^{\frac{k\nu}{2}}|V^*_t|^{\frac{\nu}{2}}|C|^{\frac{\nu}{2}}|\Sigma^*_t|^{-\frac{\nu+k+1}{2}}}{2^{\frac{\nu k}{2}}\Gamma_k(\frac{\nu}{2})} \\ &\quad \times \exp\left(-\frac{1}{2}(\nu-k-1)tr(V^*_t Y_{t,11})\right) \times \exp\left(-\frac{1}{2}(\nu-k-1)tr(CY_{t,22})\right), \end{aligned} \tag{48}$$

where $Y_t = \begin{pmatrix} Y_{t,11} & Y_{t,12} \\ Y_{t,21} & Y_{t,22} \end{pmatrix} = \Sigma^{*-1}_t$. The last step of (48) uses the fact that the determinant of a block diagonal square matrix is equal to the products of the determinants of the diagonal blocks, and that

$$V_t\Sigma^{*-1}_t = \begin{pmatrix} V^*_t & 0 \\ 0 & C \end{pmatrix}\begin{pmatrix} Y_{t,11} & Y_{t,12} \\ Y_{t,21} & Y_{t,22} \end{pmatrix} = \begin{pmatrix} V^*_t Y_{t,11} & V^*_t Y_{t,12} \\ CY_{t,21} & CY_{t,22} \end{pmatrix}$$

hence

$$tr(V_t\Sigma^{*-1}_t) = tr(V^*_t Y_{t,11}) + tr(CY_{t,22}).$$

As a result, the likelihood function of $\Sigma^*_{1:T}$ is

$$\begin{aligned} f(\Sigma^*_{1:T}|\nu, C, \Theta) &= \prod_{t=1}^{T} f(\Sigma^*_t|\Sigma^*_{1:t-1}, \nu, C, \Theta) \\ &= \frac{(\nu-k-1)^{\frac{Tk\nu}{2}} \prod_{t=1}^{T} |V^*_t|^{\frac{\nu}{2}} \prod_{t=1}^{T} |\Sigma^*_t|^{-\frac{\nu+k+1}{2}}}{2^{\frac{T\nu k}{2}} \Gamma_k(\frac{\nu}{2})^T} \exp\left(-\frac{1}{2}(\nu-k-1) tr(\sum_{t=1}^{T} V^*_t Y_{t,11})\right) \\ &\quad \times |C|^{\frac{T\nu}{2}} \exp\left(-\frac{1}{2}(\nu-k-1) tr(C \sum_{t=1}^{T} Y_{t,22})\right). \end{aligned} \tag{49}$$

Compared with the likelihood function for the IW model in (7), (49) incurs a lower computation burden mainly due to the fact that the term $\prod_{t=1}^{T} |V_t|^{\frac{\nu}{2}}$ is decomposed into the product of two terms $\prod_{t=1}^{T} |V^*_t|^{\frac{\nu}{2}}$ and $|C|^{\frac{T\nu}{2}}$. So at each MCMC iteration, instead of computing the determinant of a $k \times k$ matrix T times, we only need to compute the determinant of a $k_1 \times k_1$ matrix T times, plus once for a $k_2 \times k_2$ matrix. When k_1 is small relative to k and/or T is large, the difference in computational cost is significant. Even though we still need to compute $\prod_{t=1}^{T} |\Sigma^*_t|$, it only needs to be computed once at the beginning of the MCMC chain and is re-used at each iteration without incurring further computation burden.

Given the posterior distribution, MH steps are used to sample ν and elements of b_j and ℓ_j. Even though we can apply RCOV targeting to C and set $C = \overline{\Sigma}^*_{22}$, the second part of (49) suggests that if we place a Wishart prior on C, its posterior also follows a Wishart distribution and can be easily sampled using a Gibbs step. Indeed, let $p(C) = \text{Wishart}_{k2}(C|\gamma_C, \frac{1}{\gamma_C} I)$, then the conditional posterior of C is

$$\begin{aligned} p(C|\Sigma^*_{1:T}, \nu, \Theta) &\propto p(C) f(\Sigma^*_{1:T}|\nu, C, \Theta) \\ &\propto \text{Wishart}_{k2}(C|\overline{\gamma}_C, \overline{Q}_C), \end{aligned} \tag{50}$$

where $\overline{\gamma}_C = \gamma_C + T\nu$ and $\overline{Q}_C = \left[(\nu-k-1)\sum_{t=1}^{T} Y_{t,22} + \gamma_C I\right]^{-1}$.

The predictive density for Σ^*_t and Σ_t given data $\Sigma_{1:t-1}$ can be estimated in the usual way by averaging over the MCMC iterations. For instance, the predictive density for Σ_t can be computed following

$$p(\Sigma_t|\Sigma_{1:t-1}) \approx \frac{1}{N} \sum_{i=1}^{N} \text{Wishart}_k^{-1}(\Sigma_t|\nu^{(i)}, (\nu^{(i)}-k-1) W V_t^{(i)} W'), \tag{51}$$

where N denotes the total number of posterior draws and $V_t^{(i)}$ is from (32) using the i-th

MCMC draw. Note that in this model the predictive distribution for different Σ_t derived from different asset orderings is the same subject to a permutation matrix. This is a result of the spectral decomposition and the orthogonal transformation of Σ_t^*. This also carries over to the predictive density of returns.

Similarly the predictive density of returns, assuming (24) and integrating Σ_t out, can be approximated as

$$p(r_t|\Sigma_{1:t-1}) \approx \frac{1}{N}\sum_{i=1}^{N} \mathrm{St}_k\left(r_t|0, \frac{\nu^{(i)}-k-1}{\nu^{(i)}-k+1} W V_t^{(i)} W', \nu^{(i)}-k+1\right). \tag{52}$$

In the next section we extend our parametric factor RCOV models to countably-infinite mixture models. Mixture models with constant weights and time-varying weights are considered.

4 Nonparametric Factor Models

4.1 Dirichlet process mixture factor model (IW-DPM-F)

Now we extend our parametric factor RCOV model to a DPM version. Again we model the dynamics of Σ_t by modeling the conditional density of Σ_t^* as

$$\begin{aligned}
f(\Sigma_t^*|\Sigma_{1:t-1}^*, \Theta, \Omega, \Phi) &= \sum_{j=1}^{\infty} \omega_j \mathrm{Wishart}_k^{-1}(\Sigma_t^*|\nu_j, (\nu_j-k-1)V_{t,j}), && (53)\\
V_{t,j} &= \begin{pmatrix} V_t^{*1/2} A_j (V_t^{*1/2})' & 0 \\ 0 & C_j \end{pmatrix}, && (54)\\
\Omega &\sim \mathbf{SBP}(\alpha), && (55)\\
(\nu_j, A_j, C_j) &\overset{iid}{\sim} G_0, \quad j=1,2,\ldots, && (56)
\end{aligned}$$

where $\Phi = \{\phi_j\}_{j=1}^{\infty} = \{(\nu_j, A_j, C_j)\}_{j=1}^{\infty}$, and V_t^* is defined the same as in the parametric factor model. We call this model IW-DPM-F. In this specification cluster dependence operates through $V_{t,j}$ and the positive definite matrices A_j and C_j each of which is of lower dimension than k. Similar to the parametric case, an immediate implication is that the conditional marginal distribution of $\Sigma_{t,11}^*$, the observed dynamic factor, follows an infinite mixture of

time-varying inverse-Wishart distributions with constant weights

$$\Sigma^*_{t,11}|\Sigma^*_{1:t-1},\Theta,\Omega,\Phi \sim \sum_{j=1}^{\infty}\omega_j \text{Wishart}^{-1}_{k_1}(\nu_j-k_2,(\nu_j-k-1)V_t^{*1/2}A_j(V_t^{*1/2})'), \tag{57}$$

while the conditional distribution of the observed static part $\Sigma^*_{t,22}$ follows an infinite mixture of time-invariant inverse-Wishart distributions

$$\Sigma^*_{t,22}|\Sigma^*_{1:t-1},\Theta,\Omega,\Phi \sim \sum_{j=1}^{\infty}\omega_j \text{Wishart}^{-1}_{k_2}(\nu_j-k_1,(\nu_j-k-1)C_j). \tag{58}$$

As in the parametric model, the conditional distribution of $\Sigma^*_{t,11}$ only depends on its own history, even though it admits a much richer functional form. For example, according to (57), the conditional expectation of $\Sigma^*_{t,11}$ is

$$E(\Sigma^*_{t,11}|\Sigma^*_{1:t-1},\Theta,\Omega,\Phi) = E(\Sigma^*_{t,11}|\Sigma^*_{1:t-1,11},\Theta,\Omega,\Phi) = \sum_{j=1}^{\infty}\omega_j V_t^{*1/2}A_j(V_t^{*1/2})', \tag{59}$$

which is a function of past $\Sigma^*_{t,11}$ (through V^*_t) only. On the other hand, the conditional distribution of $\Sigma^*_{t,22}$ is not a function of past data as in the IW-F model, with its conditional (and also unconditional) expectation being

$$E(\Sigma^*_{t,22}|\Sigma^*_{1:t-1},\Theta,\Omega,\Phi) = E(\Sigma^*_{t,22}|\Omega,\Phi) = \sum_{j=1}^{\infty}\omega_j C_j. \tag{60}$$

The conditional expectation of $\Sigma^*_{t,11}$ and $\Sigma^*_{t,22}$ can also be derived by taking conditional expectation of Σ^*_t as a whole. According to (53),

$$\begin{aligned} E[\Sigma^*_t|\Sigma^*_{1:t-1},\Theta,\Omega,\Phi] &= \sum_{j=1}^{\infty}\omega_j V_{t,j} \\ &= \begin{pmatrix} \sum_{j=1}^{\infty}\omega_j V_t^{*1/2}A_j(V_t^{*1/2})' & 0 \\ 0 & \sum_{j=1}^{\infty}\omega_j C_j \end{pmatrix}. \end{aligned} \tag{61}$$

As before, the conditional expectation of $\Sigma^*_{t,12}$ are all zero, albeit with a much more complex distribution.

Some special cases of the DPM are worth noting. First, if $\omega_j = 1$, $\omega_i = 0$ for $i \neq j$ and $A_j = I$ we have the parametric model. Second, if C_j is equal to a constant matrix C for all

j, the model becomes

$$
\begin{aligned}
f(\Sigma_t^*|\Sigma_{1:t-1}^*, \Theta, C, \Omega, \Phi) &= \sum_{j=1}^{\infty} \omega_j \text{Wishart}_k^{-1}(\Sigma_t^*|\nu_j, (\nu_j - k - 1)V_{t,j}), && (62)\\
V_{t,j} &= \begin{pmatrix} V_t^{*1/2} A_j (V_t^{*1/2})' & 0 \\ 0 & C \end{pmatrix}, && (63)\\
\Omega &\sim \mathbf{SBP}(\alpha), && (64)\\
(\nu_j, A_j) &\overset{iid}{\sim} G_0, \quad j = 1, 2, \ldots, && (65)
\end{aligned}
$$

where $\Phi = \{(\nu_j, A_j)\}_{j=1}^{\infty}$. Under this specification, the conditional distribution of $\Sigma_{t,11}^*$ still follows (57). On the other hand, even though the conditional distribution of $\Sigma_{t,22}^*$ is also an infinite mixture,

$$
\Sigma_{t,22}^*|\Sigma_{1:t-1}^*, \Theta, C, \Omega, \Phi \sim \sum_{j=1}^{\infty} \omega_j \text{Wishart}_{k_2}^{-1}(\nu_j - k_1, (\nu_j - k - 1)C), \tag{66}
$$

with different ν_j for each component j, all the component distributions now have the same mean C. As a result, both the conditional and unconditional mean of $\Sigma_{t,22}^*$ are equal to C, which can be targeted at $\overline{\Sigma}_{22}^*$ in model inference instead of being estimated. We call this special case IW-DPM-F-C. The larger the dimension of C (k_2) the greater reduction in computational costs for inference from applying RCOV targeting to C.

The conditional distribution of Σ_t under IW-DPM-F model is also an infinite mixture,

$$
f(\Sigma_t|\Sigma_{1:t-1}, \Theta, \Omega, \Phi, W) = \sum_{j=1}^{\infty} \omega_j \text{Wishart}_k^{-1}(\Sigma_t|\nu_j, (\nu_j - k - 1)WV_{t,j}W'). \tag{67}
$$

The conditional mean is

$$
E(\Sigma_t|\Sigma_{1:t-1}, \Theta, \Omega, \Phi, W) = W_1 \left[\sum_{j=1}^{\infty} \omega_j V_t^{*1/2} A_j (V_t^{*1/2})'\right] W_1' + W_2 \left[\sum_{j=1}^{\infty} \omega_j C_j\right] W_2'. \tag{68}
$$

Under (24) and (67), the conditional distribution of r_t, after integrating out Σ_t, is an infinite mixture of multivariate Student-t,

$$
f(r_t|\mathcal{F}_{t-1}, \Theta, \Omega, \Phi, W) = \sum_{j=1}^{\infty} \omega_j \text{St}_k \left(r_t \middle| 0, \frac{\nu_j - k - 1}{\nu_j - k + 1} WV_{t,j}W', \nu_j - k + 1\right), \tag{69}
$$

with each component distribution having a different scale matrix and a different degree

of freedom. This provides a very rich specification which naturally accommodates fat-tails. Similarly, r_t^*, hence $r_{t,1}^*$ and $r_{t,2}^*$, each conditionally follows a mixture Student-t distribution,[11]

$$f(r_t^*|\mathcal{F}_{t-1},\Theta,\Omega,\Phi,W) = \sum_{j=1}^{\infty}\omega_j \mathrm{St}_k\left(r_t^*\middle|0,\frac{\nu_j-k-1}{\nu_j-k+1}V_{t,j},\nu_j-k+1\right), \quad (70)$$

$$f(r_{t,1}^*|\mathcal{F}_{t-1},\Theta,\Omega,\Phi,W) = \sum_{j=1}^{\infty}\omega_j \mathrm{St}_{k_1}\left(r_{t,1}^*\middle|0,\frac{\nu_j-k-1}{\nu_j-k+1}V_t^{*1/2}A_j(V_t^{*1/2})',\nu_j-k+1\right) \quad (71)$$

$$f(r_{t,2}^*|\mathcal{F}_{t-1},\Theta,\Omega,\Phi,W) = \sum_{j=1}^{\infty}\omega_j \mathrm{St}_{k_2}\left(r_{t,2}^*\middle|0,\frac{\nu_j-k-1}{\nu_j-k+1}C_j,\nu_j-k+1\right). \quad (72)$$

Under IW-DPM-F-C, (72) becomes

$$f(r_{t,2}^*|\mathcal{F}_{t-1},\Theta,C,\Omega,\Phi,W) = \sum_{j=1}^{\infty}\omega_j \mathrm{St}_{k_2}\left(r_{t,2}^*\middle|0,\frac{\nu_j-k-1}{\nu_j-k+1}C,\nu_j-k+1\right). \quad (73)$$

In this case, even though the mixture has the same mean C for each component distribution, the scale matrix itself $\frac{\nu_j-k-1}{\nu_j-k+1}C$ is different across components.

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

$$G_0(\nu_j,A_j,C_j) \equiv \mathrm{Exp}_{\nu>k+1}(\lambda)\times \mathrm{Wishart}_{k_1}\left(\gamma_A,\frac{1}{\gamma_A}I\right)\times \mathrm{Wishart}_{k_2}\left(\gamma_C,\frac{1}{\gamma_C}I\right), \quad (74)$$

where $\gamma_A \geq k_1, \gamma_C \geq k_2$; and for the Wishart version (W-DPM-F) as:

$$\begin{aligned} G_0(\nu_j,A_j,C_j) \equiv\ & \mathrm{Exp}_{\nu>k}(\lambda)\times \mathrm{Wishart}_{k_1}^{-1}(\gamma_A,(\gamma_A-k-1)I) \\ & \times \mathrm{Wishart}_{k_2}^{-1}(\gamma_C,(\gamma_C-k-1)I), \end{aligned} \quad (75)$$

where $\gamma_A \geq k_1+1, \gamma_C \geq k_2+1$. Under G_0, ν_j, A_j and C_j are independently drawn from a truncated exponential distribution and two Wishart (inverse-Wishart) distributions, respectively. Note 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 $\Sigma_{t,11}^*$ 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,

$$\alpha \sim \mathrm{Gamma}(a_0,c_0). \quad (76)$$

[11] As in the parametric factor model $r_{t,1}^*$ and $r_{t,2}^*$ are conditionally uncorrelated.

4.1.1 Posterior inference

To sample from the posterior for the IW-DPM-F model we use slice sampling techniques introduced by Walker (2007) and extended by Kalli et al. (2011) and Papaspiliopoulos (2008).[12] 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 $\phi_j = (\nu_j, A_j, C_j)$ and in the following conditioning on $\Sigma^*_{1:t-1}$ is suppressed where the context is clear. The general model is

$$f(\Sigma^*_t|\Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \omega_j h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j), \tag{77}$$

where $h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j)$ corresponds to either the inverse-Wishart in (53) or its Wishart analogue. Introducing an auxiliary latent variable $0 < u_t < 1$, we define the joint conditional density of Σ^*_t and u_t as

$$f(\Sigma^*_t, u_t|\Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \mathbf{1}(u_t < \omega_j) h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j). \tag{78}$$

Note that integrating out u_t returns the original model (77). 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 $h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j)$. The target likelihood is now

$$\begin{aligned} f(\Sigma^*_{1:T}, u_{1:T}, s_{1:T}|\Theta, \Omega, \Phi) &= \prod_{t=1}^{T} f(\Sigma^*_t, u_t, s_t|\Theta, \Omega, \Phi) \\ &= \prod_{t=1}^{T} \mathbf{1}(u_t < \omega_{s_t}) h(\Sigma^*_t|\Theta, \nu_{s_t}, A_{s_t}, C_{s_t}), \end{aligned} \tag{79}$$

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

$$p(\Theta)p(\Omega_{\overline{K}}) \left[\prod_{i=1}^{\overline{K}} p(\nu_j, A_j, C_j)\right] \prod_{t=1}^{T} \mathbf{1}(u_t < \omega_{s_t}) h(\Sigma^*_t|\Theta, \nu_{s_t}, A_{s_t}, C_{s_t}), \tag{80}$$

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

[12] Sampling methods for the Wishart version only require minor modifications.

The posterior sampling steps are as follows.

1. $p(\phi_j|\Sigma^*_{1:T}, s_{1:T}, \Theta) \propto p(\phi_j) \prod_{\{t:s_t=j\}} h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j), j = 1, \ldots, \overline{K}$.
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. $P(s_t = j|\Sigma^*_{1:T}, \Phi, \Omega_{\overline{K}}, \Theta, u_{1:T}) \propto \mathbf{1}(u_t < \omega_j) h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j)$.
6. $p(\alpha|K) \propto p(\alpha)p(K|\alpha)$, where K is the number of active clusters in $s_{1:T}$.
7. $p(\Theta|\Sigma^*_{1:T}, s_{1:T}, \Phi) \propto p(\Theta) \prod_{t=1}^{T} h(\Sigma^*_t|\Theta, \nu_{s_t}, A_{s_t}, C_{s_t})$

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

$$p(A_j|\nu_j, C_j, \Sigma^*_{1:T}, s_{1:T}, \Theta) \propto p(A_j) \prod_{\{t:s_t=j\}} h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j). \tag{81}$$

By conjugacy, we have for IW-DPM-F model

$$A_j \sim \text{Wishart}_{k_1}(\overline{\gamma}_{A,j}, \overline{Q}_{A,j}), \tag{82}$$

where $\overline{\gamma}_{A,j} = \gamma_A + n_j\nu_j$ and $\overline{Q}_{A,j} = \left[(\nu_j - k - 1) \sum_{\{t:s_t=j\}} \left[(V_t^{*1/2}) Y_{t,11} ((V_t^{*1/2}))'\right] + \gamma_A I\right]^{-1}$, with $n_j = \#\{t : s_t = j\}$. The conditional posterior of C_j is

$$p(C_j|\nu_j, A_j, \Sigma^*_{1:T}, s_{1:T}, \Theta) \propto p(C_j) \prod_{\{t:s_t=j\}} h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j). \tag{83}$$

Again by conjugacy, we have for IW-DPM-F model

$$C_j \sim \text{Wishart}_{k_2}(\overline{\gamma}_{C,j}, \overline{Q}_{C,j}), \tag{84}$$

where $\overline{\gamma}_{C,j} = \gamma_C + n_j\nu_j$ and $\overline{Q}_{C,j} = \left[(\nu_j - k - 1) \sum_{\{t:s_t=j\}} Y_{t,22} + \gamma_C I\right]^{-1}$, The conditional posterior of ν_j is

$$p(\nu_j|A_j, C_j, \Sigma^*_{1:T}, s_{1:T}, \Theta) \propto p(\nu_j) \prod_{\{t:s_t=j\}} h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j). \tag{85}$$

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,

$$\alpha \sim p_\eta \text{Gamma}(a_0 + K, c_0 - \log\eta) + (1 - p_\eta)\text{Gamma}(a_0 + K - 1, c_0 - \log\eta), \tag{86}$$

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 parametric models, we impose the same restriction associated with RCOV targeting in the nonparametric models. That is, we set $B_0 = (\iota\iota' - B_1 - \cdots - B_M) \odot \overline{\Sigma}^*_{11}$ in estimation and reject any draws in which B_0 is not positive definite or any element of $\sum_{j=1}^M B_j$ is not less than 1 in modulus.[13]

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

$$G^{(i)} = \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} \delta_{\phi_j^{(i)}} + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) G_0. \tag{87}$$

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

Note that several of these sampling steps can exploit parallel programming. Steps 1-3, and 5 can employ parallel programming directly since the computations can be done independently. For example, in Step 1 the sampling of each $\phi_j, j = 1, \ldots, \overline{K}$ can be done simultaneously on separate CPU cores. For a large number of active clusters this can result in a significant reduction in computational time.

[13] In special cases IW-DPM-F-C and W-DPM-F-C where $C_j = C$ for all j, RCOV targeting is applied by setting $C = \overline{\Sigma}^*_{22}$. So no sampling of C_j is needed in Step 1, and other steps remain the same.

4.1.2 Predictive density

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

$$\begin{aligned} & p(\Sigma_{T+1}|\Sigma_{1:T}, G^{(i)}, W) \\ = & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}, W) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \int h(\Sigma_{T+1}|\Theta^{(i)}, \phi, W) G_0(d\phi) \quad (88) \\ \approx & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}, W) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \frac{1}{R} \sum_{l=1}^{R} h(\Sigma_{T+1}|\Theta^{(i)}, \phi^{[l]}, W), \quad (89) \end{aligned}$$

where $\phi^{[l]} \overset{iid}{\sim} G_0, l = 1, \ldots, R$. [14] For IW-DPM-F,

$$\begin{aligned} h(\Sigma_{T+1}|\Theta, \phi_j, W) &= \text{Wishart}_k^{-1}(\Sigma_{T+1}|\nu_j, (\nu_j - k - 1) W V_{T+1,j} W') \\ &= \text{Wishart}_k^{-1}(\Sigma^*_{T+1}|\nu_j, (\nu_j - k - 1) V_{T+1,j}) \\ &= h(\Sigma^*_{T+1}|\Theta, \phi_j). \end{aligned} \quad (90)$$

The second equality holds because the (inverse) Wishart distribution is closed under linear transformation and W is an orthogonal matrix. And similar results hold for W-DPM-F. In general in this framework (using Wishart families for the kernels),

$$p(\Sigma_{T+1}|\Sigma_{1:T}, G^{(i)}, W) = p(\Sigma^*_{T+1}|\Sigma^*_{1:T}, G^{(i)}). \quad (91)$$

Finally, the predictive density with all parameter and distributional uncertainty integrated out is estimated as

$$p(\Sigma_{T+1}|\Sigma_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} p(\Sigma^*_{T+1}|\Sigma^*_{1:T}, G^{(i)}). \quad (92)$$

The predictive density of r_{T+1} can be computed in a similar way. For example under

[14] In the empirical work $R = 10$ but smaller values gave similar accuracy.

IW-DPM-F specification

$$\begin{aligned} & p(r_{T+1}|\mathcal{F}_T, G^{(i)}, W) \\ = & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h_r(r_{T+1}|\Theta^{(i)}, \phi_j^{(i)}, W) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \int h_r(r_{T+1}|\Theta^{(i)}, \phi, W) G_0(d\phi), \end{aligned} \quad (93)$$

where

$$\begin{aligned} h_r(r_{T+1}|\Theta, \phi_j, W) &= \mathrm{St}_k\left(r_{T+1}\middle|0, \frac{\nu_j - k - 1}{\nu_j - k + 1} W V_{T+1,j} W', \nu_j - k + 1\right) \\ &= \mathrm{St}_k\left(r^*_{T+1}\middle|0, \frac{\nu_j - k - 1}{\nu_j - k + 1} V_{T+1,j}, \nu_j - k + 1\right). \end{aligned} \quad (94)$$

4.2 Infinite Hidden Markov Factor Model (IW-IHMM-F)

In the DPM model all time dependence occurs through the evolution of the observable V_t^*. The infinite hidden Markov model discussed in this section allows the unobserved state variable s_t to contribute to changes in the conditional distribution through time. This model is like a DPM specification with time-varying weights.

The construction of the IHMM factor model with an inverse-Wishart distribution closely follows the DPM version. Given the framework described by (16) – (20), we let Σ_t^* follow

$$\Sigma_t^* | \Sigma_{1:t-1}^*, \Theta, \Pi, \Phi, s_t \sim \mathrm{Wishart}_k^{-1}(\nu_{s_t}, (\nu_{s_t} - k - 1) V_{t,s_t}), \quad (95)$$

$$V_{t,s_t} = \begin{pmatrix} V_t^{*1/2} A_{s_t} (V_t^{*1/2})' & 0 \\ 0 & C_{s_t} \end{pmatrix}, \quad (96)$$

where $\Phi = \{\phi_j\}_{j=1}^{\infty} = \{(\nu_j, A_j, C_j)\}_{j=1}^{\infty}$, and V_t^* the same as before. We refer to this model as IW-IHMM-F. The parameters α β and κ play an important role in the number of unique clusters in the mixture as well as state persistence. Rather than setting the parameters we impose the following priors,

$$\alpha \sim \mathrm{Gamma}(a_3, c_3), \qquad \beta + \kappa \sim \mathrm{Gamma}(a_4, c_4), \qquad \rho = \frac{\kappa}{\beta + \kappa} \sim \mathrm{Beta}(a_5, c_5), \quad (97)$$

which allow for learning from the data. This prior formulation is more convenient for posterior sampling.

Under this specification, given s_{t-1}, the conditional distribution of Σ_t^* is an infinite mix-

ture with time-varying weights

$$f(\Sigma_t^*|\Sigma_{1:t-1}^*, \Theta, \Pi, \Phi, s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \text{Wishart}_k^{-1}(\Sigma_t^*|\nu_{s_t}, (\nu_{s_t} - k - 1)V_{t,s_t}). \quad (98)$$

Some key features of IW-IHMM-F are the following. Firstly, the conditional marginal distribution of $\Sigma_{t,11}^*$ follows an infinite mixture of time-varying inverse-Wishart with time-varying weights

$$\Sigma_{t,11}^*|\Sigma_{1:t-1}^*, \Theta, \Pi, \Phi, s_{t-1} \sim \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \text{Wishart}_{k_1}^{-1}(\nu_{s_t} - k_2, (\nu_{s_t} - k - 1)V_t^{*1/2} A_{s_t} (V_t^{*1/2})'). \quad (99)$$

So $\Sigma_{t,11}^*$ only depends on its own history, which is the same case as in IW-F and IW-DPM-F, but now the dependence can change over time according to the latent Markov state. In particular, the conditional mean of $\Sigma_{t,11}^*$ is

$$E(\Sigma_{t,11}^*|\Sigma_{1:t-1}^*, \Theta, \Pi, \Phi, s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} V_t^{*1/2} A_{s_t} (V_t^{*1/2})', \quad (100)$$

which is clearly time-varying.

Secondly, the conditional distribution of $\Sigma_{t,22}^*$ follows an infinite mixture of time-invariant inverse-Wishart with time-varying weights

$$\Sigma_{t,22}^*|\Sigma_{1:t-1}^*, \Theta, \Pi, \Phi, s_{t-1} \sim \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \text{Wishart}_{k_2}^{-1}(\nu_{s_t} - k_1, (\nu_{s_t} - k - 1)C_{s_t}). \quad (101)$$

Thus, the conditional distribution of $\Sigma_{t,22}^*$ is actually time-varying in contrast to the parametric case and the DPM case discussed earlier. The conditional mean of $\Sigma_{t,22}^*$ is

$$E(\Sigma_{t,22}^*|\Sigma_{1:t-1}^*, \Theta, \Pi, \Phi, s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} C_{s_t}. \quad (102)$$

In the parametric factor model we interpreted the conditional mean of $\Sigma_{t,22}^*$ as the static part of the realized covariance. This extension removes that restriction.

An interesting special case is enforcing a constant conditional (also unconditional) mean

for $\Sigma^*_{t,22}$. Specifically, let $C_j = C$ for all j and $\Phi = \{(\nu_j, A_j)\}_{j=1}^{\infty}$, then

$$\Sigma^*_{t,22}|\Sigma^*_{1:t-1}, C, \Theta, \Pi, \Phi, s_{t-1} \sim \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \text{Wishart}^{-1}_{k_2}(\nu_{s_t} - k_1, (\nu_{s_t} - k - 1)C), \tag{103}$$

which results in a constant conditional mean equal to C. Note that due to the mixing in ν_{s_t}, the conditional distribution is still time-varying. We refer to this version with constant C as IW-IHMM-F-C. With this restriction, C can be targeted at $\overline{\Sigma}^*_{22}$ in model inference instead of being estimated. This again greatly reduces computational cost, especially for large dimension where the dimension of C will be large.

Thirdly, $\Sigma^*_{t,12}$ conditionally also follows an infinite mixture of unknown form with time-varying weights, while retains an all-zero mean, both conditionally and unconditionally.

The conditional distribution of Σ_t under IW-IHMM-F is also an infinite mixture of inverse-Wishart with time-varying weights,

$$f(\Sigma_t|\Sigma_{1:t-1}, \Theta, \Pi, \Phi, W, s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \text{Wishart}^{-1}_{k}(\Sigma_t|\nu_{s_t}, (\nu_{s_t} - k - 1)WV_{t,s_t}W'). \tag{104}$$

The conditional mean becomes

$$\begin{aligned} E(\Sigma_t|\Sigma_{1:t-1}, \Theta, \Pi, \Phi, W, s_{t-1}) &= W_1 \left[\sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} V_t^{*1/2} A_{s_t} (V_t^{*1/2})'\right] W_1' \\ &\quad + W_2 \left[\sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} C_{s_t}\right] W_2'. \end{aligned} \tag{105}$$

If $W = I$ and $k_1 = k$, which means there is no factor structure and no transformation of RCOV, the IW-IHMM-F model becomes the IW-IHMM introduced by Jin & Maheu (2016).

Under (24) and (104), the conditional distribution of r_t, after integrating out Σ_t, is an infinite mixture of multivariate Student-t with time-varying weights,

$$f(r_t|\mathcal{F}_{t-1}, \Theta, \Pi, \Phi, W, s_{t-1}) = \sum_{s_t=1}^{\infty} \pi_{s_{t-1},s_t} \text{St}_k\left(r_t \middle| 0, \frac{\nu_{s_t} - k - 1}{\nu_{s_t} - k + 1} WV_{t,s_t}W', \nu_{s_t} - k + 1\right). \tag{106}$$

4.2.1 Posterior inference

Similar to the posterior sampling methods for the DPM model of Section 4.1 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. Slice sampling

introduces on an auxiliary variable that 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.

The auxiliary latent variable $0 < u_t < 1$ is introduced such that its conditional density is

$$p(u_t|s_t, s_{t-1}, \Pi) = \frac{\mathbf{1}(u_t < \pi_{s_{t-1},s_t})}{\pi_{s_{t-1},s_t}} \tag{107}$$

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

$$\begin{aligned} p(s_t|u_{1:t}, \Sigma^*_{1:t}) &\propto h(\Sigma^*_t|\phi_{s_t}) \sum_{s_{t-1}=1}^{\infty} p(u_t|s_t, s_{t-1}) p(s_t|s_{t-1}) p(s_{t-1}|\Sigma^*_{1:t-1}, u_{1:t-1}) && (108) \\ &\propto h(\Sigma^*_t|\phi_{s_t}) \sum_{s_{t-1}=1}^{\infty} \mathbf{1}(u_t < \pi_{s_{t-1},s_t}) p(s_{t-1}|u_{1:t-1}, \Sigma^*_{1:t-1}) && (109) \\ &\propto h(\Sigma^*_t|\phi_{s_t}) \sum_{s_{t-1}:u_t<\pi_{s_{t-1},s_t}} p(s_{t-1}|u_{1:t-1}, \Sigma^*_{1:t-1}). && (110) \end{aligned}$$

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

$$p(s_t|s_{t+1}, \Sigma^*_{1:T}, u_{1:T}) \propto p(s_t|u_{1:t}, \Sigma^*_{1:t}) \mathbf{1}(u_{t+1} < \pi_{s_t,s_{t+1}}). \tag{111}$$

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 (111).

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 provides full details on each of the steps.

4.2.2 Predictive density

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

$$
\begin{aligned}
& p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, \Phi^{(i)}, s^{(i)}_{1:T}, \Theta^{(i)}, W) \\
= & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j} h(\Sigma_{T+1}|\Theta^{(i)}, \phi^{(i)}_j, W) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j}\right) \int h(\Sigma_{T+1}|\Theta^{(i)}, \phi, W) G_0(d\phi) \qquad (112) \\
= & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j} h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi^{(i)}_j) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j}\right) \int h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi) G_0(d\phi) \qquad (113) \\
\approx & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j} h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi^{(i)}_j) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j}\right) \frac{1}{R} \sum_{l=1}^{R} h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi^{[l]}), \qquad (114)
\end{aligned}
$$

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

$$
p(\Sigma_{T+1}|\Sigma_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, \Phi^{(i)}, s^{(i)}_{1:T}, \Theta^{(i)}, W), \qquad (115)
$$

where the right hand side terms are from (114) which integrates out all uncertainty. Similarly, the predictive density for returns is computed as in the IW-DPM-F model with the constant weights ω_j replaced by $\pi_{s_T,j}$.

5 Benchmark Models

A multivariate vector diagonal GARCH (VDGARCH) model of Ding & Engle (2001) based on daily returns is used for comparison. The model has Student-t innovations and is

$$r_t \sim \mathrm{St}_k(0, H_t, \nu), \tag{116}$$

$$H_t = CC' + aa' \odot r_{t-1}r_{t-1}' + bb' \odot H_{t-1}, \tag{117}$$

where a and b are $k \times 1$ vectors, $\nu > 2$ is the degree of freedom and covariance targeting is implemented with $C = \mathrm{Cov}(r_t) \odot (\iota\iota'(\nu-2)/\nu - aa' - bb'(\nu-2)/\nu)$ where $\mathrm{Cov}(r_t)$ is the sample covariance of daily returns. The total number of parameters is $2k+1$.

The covariance matrix discounting model in (West & Harrison 1997, chap 16) is parsimonious and suitable for forecasting large covariance matrices of returns. The following version is used,[15]

$$H_{t+1}|r_{1:t} \sim \mathrm{Wishart}_k^{-1}(\beta n_t + k - 1, \beta n_t S_t), \tag{118}$$

$$n_t = \beta n_{t-1} + 1 \tag{119}$$

$$S_t = \frac{1}{n_t}(\beta n_{t-1} S_{t-1} + r_t r_t'). \tag{120}$$

H_{t+1} is the latent covariance matrix of r_{t+1} and its predictive distribution follows an inverse-Wishart distribution given data $r_{1:t}$. S_t can be regarded as the posterior estimate of the "true" covariance matrix at time t, given information $r_{1:t}$. Equation (118) reflects the "prior" or prediction of Σ_{t+1} given $r_{1:t}$. $\beta = 0.95$ and is the discounting factor reflecting information decay moving from time t to $t+1$. Sequential updating/re-enforcement of belief is governed by (119) through the degree of freedom parameter of the inverse-Wishart distribution. A larger value for the degree of freedom results in a tighter distribution and hence stronger belief. Sequential updating of S_t as a new observation r_t becomes available follows (120).

[15] West & Harrison (1997) use a different parameterization of the inverse-Wishart distribution (see chap. 16.4). Our notation reflects this difference.

Assuming $r_t|H_t \sim N(0, H_t)$, the predictive density of returns is:

$$r_{t+1}|r_{1:t} \sim \text{St}_k(0, S_t, \beta n_t). \tag{121}$$

We also modify the covariance matrix discounting model to what we call a RCOV discounting model. The key steps are summarized in the following equations

$$\Sigma_{t+1}|\mathcal{F}_t \sim \text{Wishart}_k^{-1}(\beta n_t + k - 1, \beta n_t S_t) \tag{122}$$

$$n_t = \beta n_{t-1} + 1 \tag{123}$$

$$S_t = \frac{1}{n_t}(\beta n_{t-1} S_{t-1} + \Sigma_t) \tag{124}$$

The model has the same interpretation as the covariance matrix discounting model except $r_t r_t'$ is replaced with Σ_t and the predictive density in (122) is for the observed RCOV. Assuming $r_t|\Sigma_t \sim N(0, \Sigma_t)$ the predictive density of returns give past data is

$$r_{t+1}|\mathcal{F}_{1:t} \sim \text{St}_k(0, S_t, \beta n_t). \tag{125}$$

Finally, a random walk (RW) that uses last period's value for all future forecasts and an exponentially weighted moving average (EWMA) with smoothing parameter 0.95 are included.

6 Empirical Applicatons

6.1 10 Asset Application

In this section we discuss the results for the 10 asset application. The benefit of this smaller dimension is that we can feasibly estimate all models including the highly parameterized non-factor nonparametric models. Factor models represent a compromise in that we can capture most of the significant structure in the data but maintain a tractable model and estimation cost. This application will allow us to measure the trade-offs.

The 10-asset RCOV daily data from Noureldin et al. (2012) are used and range from 2001/02/01 to 2009/12/31 (2092 observations).[16] The stock symbols used are: Alcoa (AA), American Express (AXP), Bank of America (BAC), Coca Cola (KO), Du Pont (DD), General Electric (GE), International Business Machines (IBM), JP Morgan (JPM), Microsoft

[16]The data were downloaded from http://realized.oxford-man.ox.ac.uk/data/download and we are grateful for the authors making them available.

(MSFT), and Exxon Mobil (XOM). The original RCOV is open-to-close, we add the outer-product of the overnight return to the original RCOV to form close-to-close RCOV in order to match daily close-to-close returns. The last 500 observations are used for out-of-sample forecast evaluation. Each model is re-estimated at each day in the out-of-sample period. The diagonal elements of Σ_t^* are displayed in Figure 1.

To evaluate model density forecasts Table 1 displays the log-predictive likelihoods for a variety of models for forecasts horizons of $h = 1, 5, 10, 20, 60$. This is computed as $\sum_{t=T_0-h}^{T-h} \log(p(\Sigma_{t+h}|\mathcal{F}_t, \mathcal{A}))$ for model $\mathcal{A}$ where T_0 is the start of the out-of-sample period. Included in the table are a Wishart (W)[17] and an inverse-Wishart (IW) specifications along with nonparametric versions with no factor structure. Following this are parametric and nonparametric factor models assuming a factor dimension from 1 to 9. Below this are two benchmark specifications discussed in Section 5.

In terms of predictive accuracy the IW-IHMM performs the best. This model strongly dominates the parametric W and IW alternatives as well as the benchmark models. For instance, the log-Bayes factor for the IW-IHMM against the IW model is 5181. The factor models all fall short of the forecast performance of the IHMM but as the dimension of the factor increases they improve.

In general for a given factor dimension the best model is the IHMM followed by the DPM and the parametric factor version. In each case, moving from the parametric factor structure to a nonparametric version results in considerable improvement. For example, the log-Bayes factor for the IW-IHMM-F with 5 factors versus the IW-F is 8070.

Finally, the Wishart version (W) is dominated by the inverse-Wishart model. This is consistent with the results in Jin & Maheu (2016).

Turning to point forecasts shown in Table 2 the IHMM factors models with 5 or more factors achieve the lowest root-mean-squared errors. None of the benchmark models are particularly competitive although the RCOV discount model and the EWMA are better as they use RCOV directly while the others use daily returns. The IHMM version is generally much better than the DPM version or parametric versions.

Capturing the complex dynamics in RCOV contributes to better density forecasts for returns as shown in Table 3. Except at $h = 1$ the IW-DPM model performs better than the IW-IHMM. However, for a given factor dimension the IHMM variant always beats the DPM factor model.

We note the following observations. The nonparametric models, particularly the IHMM

[17]Labelled as W-A(3) in Jin & Maheu (2016).

version offer large improvements in all measures of forecast accuracy. Factor models represent a compromise and diminished forecast accuracy compared to the full nonparametric models. However, a 5 factor IW-IHMM-F dominates all benchmark models with the exception of some longer horizon density forecasts for returns from the RCOV discount and VDGARCH-t models. The benefit of the factor models is reduced computation time. For instance, the approximate computing time for IW is 6m20s, for IW-IHMM is 8m23s while it is only 4m3s for IW-IHMM-F with 5 factors.[18] In larger dimensions the IW-F and IW-DPM and IW-IHMM are not practically feasible while the factors models are. We turn to a more challenging application next.

6.2 60 Asset Application

For the second dataset, we use high-frequency transaction prices of 60 liquid stocks[19] among the S&P 500 that are continuously traded over a sample period of 2265 days spanning from 2006/01/03 to 2014/12/31. The high-frequency data are obtained from the TAQ database. After cleaning the raw data[20], we follow Noureldin et al. (2012) and use 5-minute returns with subsampling to compute daily open-to-close RCOV matrices. To match the close-to-close daily return, the outer-product of the overnight return is added to the corresponding open-to-close RCOV to form close-to-close RCOV. The last 500 observations (2013/01/08 to 2014/12/31) are used for out-of-sample forecasts and model comparison.

At 60 dimensions several of the previous models, IW, IW-DPM and IW-IHMM are no longer feasible to estimate and forecast with. Instead we confine our comparison to the factor models and the benchmark specifications. Based on the results from previous applications we focus on the IHMM factor models since they generally dominated the DPM versions.

Table 4 records the log-predictive likelihood values for various out-of-sample forecasts horizons. The IW-IHMM-F is the dominant model at each forecast horizon with log-Bayes factors against alternatives in the thousands. For instance, the log-Bayes factor for the 10 factor IHMM model against the parametric (IW-F) version for $h = 1$ is 266405 while it is 213896 against the RCOV discount model. The RCOV discount model is often better than the parametric IW-F models for $h = 1, 5$ and 10.

The performance of point forecasts is found in Table 5. Here the 10 factor IW-IHMM-F

[18]The next application discusses computation time in more detail.

[19]The stock symbols are: AA, AAPL, ABT, AIG, AMGN, AMZN, APC, AXP,BA, BAC, BAX, BMY, C, CAT, CL, COF, COST, CSCO, CVS, CVX, DD, DIS, DOW, EBAY, EMR, EXC, F, GD, GE, GS, HAL, HD, HON, IBM, INTC, JNJ, JPM, KO, KR, LLY, LOW, MCD, MMM, MO, MRK, MSFT, NKE, PEP, PFE, PG, SO, UNH, UNP, UPS, USB, UTX, VZ, WFC, WMT, XOM.

[20]We follow steps in Barndorff-Nielsen et al. (2011) to clean the raw data.

model has the lowest root-mean squared forecast error at each h although the loss in accuracy in reducing the factor dimension to 5 or even 3 is minor. The most competitive benchmark models are the RCOV discount model and the EWMA. The parametric factor models with factor dimension 7 or more are generally as good or better than the RCOV discount model.

Density forecasts for daily returns are reported in Table 6. For each forecast horizon the IW-IHMM-F specification is the most accurate. As the forecast horizon h increases there is a reduction in the number of factors needed. This is consistent with the need for a more flexible model to capture the stronger short-term time-series dynamics of RCOV that are important to returns. However, there is not much loss in reducing the factor from 10 to 7 or 5 for $h = 1$.

The log-Bayes factor for the 10 factor IHMM model against the parametric (IW-F) version for $h = 1$ is 1533 while it is 492 against the VDGARCH-t model that only uses daily return data. The GARCH model is very competitive and beats each of the benchmark models as well as all the parametric factor models. Only the nonparametric IHMM factor model is better. However, set against this is a very large computational cost for the GARCH model which we discuss later.

To consider the value of these models for portfolio choice, Table 7 reports the realized variance of the global minimum variance portfolio (GMVP). The GMVP solves the following problem,

$$\min \omega_{t+h|t}^{'} \Sigma_{t+h|t} \omega_{t+h|t}, \quad \text{s.t. } \omega_{t+h|t} \iota = 1, \tag{126}$$

where ω is the portfolio weight and $\Sigma_{t+h|t} \equiv E[\Sigma_{t+h}|\mathcal{F}_t, \mathcal{A}]$ is the predictive mean of Σ_{t+h} given time t information for model $\mathcal{A}$. The optimal solution to this is

$$\hat{\omega}_{t+h|t} = \frac{\Sigma_{t+h|t}^{-1} \iota}{\iota' \Sigma_{t+h|t}^{-1} \iota}. \tag{127}$$

The ex post realized variance for model $\mathcal{A}$'s portfolio is $\frac{1}{T-T_0+1} \sum_{t=T_0-h}^{T-h} \hat{\omega}_{t+h|t}^{'} \Sigma_{t+h|t} \hat{\omega}_{t+h|t}$. Better models will produce lower ex post portfolio variances.

The 10 factor IW-IHMM-F consistently produces the smallest portfolio variance in the out-of-sample period. The difference in using the same model with less factors is fairly minor so that a 3 or 5 factor model is a good alternative. The parametric factor models are quite competitive. Most of the benchmark models produce a higher portfolio variance with the exception of the EWMA.

Full sample estimates of ℓ_2, ℓ_3 and the number of alive clusters in the mixture are in Table 8. The number of active components in the mixtures range from 14 to 16 on average. The lag length of ℓ_3 is substantially larger than ℓ_2 in all cases except for the 1 and 3 factor models.

Finally we have discussed the computational advantages of the factor model earlier. The factor model allows for a faster evaluation of the data density when the factor dimension is significantly less than the data dimension. In addition, for the infinite mixture models parallel programming is very efficient when sampling data density parameters conditional on the state indicator. These benefits are seen in Table 9. The run time for 20000 MCMC draws are all in the range of a matter of minutes. The IHMM are more expensive but no where near as prohibitive as the time to estimate the VDGARCH-t model.

In summary, the factor models provide feasible estimation times for large realized covariances. The IHMM version is not only computationally feasible but overall produces the best out-of-sample forecasts and portfolio selection.

The greatest gains are found in density forecasts of RCOV and daily returns in which the rich mixture structure captures the unknown features of RCOV. The gains in point forecasts and portfolio choice are smaller in general compared to benchmark models.

7 Conclusion

This paper introduces a new factor structure that can be used in parametric (inverse-) Wishart models as well as finite and infinite mixtures models for RCOV matrices. Mixtures models offer a tractable approach to leverage our knowledge from parametric approaches to span the complex unknown distributions of RCOV matrices. There are several computational benefits to this approach that make estimation in high dimension applications feasible. Across a range of forecast metrics and portfolio choice the infinite hidden Markov factor model performs well.

8 Appendix: Sampling details for IW-IHMM-F model

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

      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 $A_{\overline{K}+1} \sim \mathrm{Wishart}_{k_1}(\gamma_A, \frac{1}{\gamma_A} I)$, $C_{\overline{K}+1} \sim \mathrm{Wishart}_{k_2}(\gamma_C, \frac{1}{\gamma_C} I)$, $\nu_{\overline{K}+1} \sim \mathrm{Exp}_{\nu > k+1}(\lambda)$.

      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

$$p(s_t = j | u_{1:T}, \Pi, \Phi, \Theta, \Sigma^*_{1:t-1}) \propto \sum_{i=1}^{\overline{K}} \mathbf{1}(u_t < \pi_{i,j}) p(s_{t-1} = i | u_{1:T}, \Pi, \Phi, \Theta, \Sigma^*_{1:t-1}) \quad (128)$$

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

$$p(s_t = j | u_{1:T}, \Pi, \Phi, \Theta, \Sigma^*_{1:t}) \propto p(s_t = j | u_{1:T}, \Pi, \Phi, \Theta, \Sigma^*_{1:t-1}) h(\Sigma^*_t | \Sigma^*_{1:t-1}, \Theta, \phi_j) \quad (129)$$

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, C, 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

$$\boldsymbol{\pi}_0 \sim \text{Dirichlet}(\overline{m}_{\cdot 1}, \ldots, \overline{m}_{\cdot K}, \alpha). \quad (130)$$

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

$$\boldsymbol{\pi}_j \sim \text{Dirichlet}(\beta\pi_{01} + n_{j1}, \ldots, \beta\pi_{0j} + \kappa + n_{jj}, \ldots, \beta\pi_{0K} + n_{jK}, \beta\pi_{0K+1}). \quad (131)$$

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

(a) draw

$$A_j \sim \text{Wishart}_{k_1}(\overline{\gamma}_{A,j}, \overline{Q}_{A,j}), \quad (132)$$

where $\overline{\gamma}_{A,j} = \gamma_A + n_{\cdot j}\nu_j$, and $\overline{Q}_{A,j} = \left[(\nu_j - k - 1) \sum_{\{t: s_t = j\}} \left[(V_t^{1/2}) Y_{t,11} ((V_t^{1/2}))'\right] + \gamma_A I\right]^{-1}$;

(b) draw

$$C_j \sim \text{Wishart}_{k_2}(\overline{\gamma}_{C,j}, \overline{Q}_{C,j}), \tag{133}$$

where $\overline{\gamma}_{C,j} = \gamma_C + n_{\cdot j}\nu_j$, and $\overline{Q}_{C,j} = \left[(\nu_j - k - 1)\sum_{\{t:s_t=j\}} Y_{t,22} + \gamma_C I\right]^{-1}$;

(c) sample

$$\begin{aligned} \nu_j &\sim p(\nu_j|\Sigma^*_{1:T}, s_{1:T}, A_j, C_j, \Theta) \\ &\propto p(\nu_j) \prod_{\{t:s_t=j\}} h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j). \end{aligned} \tag{134}$$

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.

References

Asai, M. & McAleer, M. (2015), 'Forecasting co-volatilities via factor models with asymmetry and long memory in realized covariance', *Journal of Econometrics* **189**(2), 251–262.

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.

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. (2016), 'Forecasting comparison of long term component dynamic models for realized covariance matrices', *Annals of Economics and Statistics* (123/124), 103–134.

Bauwens, L., Braione, M. & Storti, G. (2017), 'A dynamic component model for forecasting high-dimensional realized covariance matrices', *Econometrics and Statistics* **1**, 40 – 61.

Callot, L. A. F., Kock, A. B. & Medeiros, M. C. (2017), 'Modeling and forecasting large realized covariance matrices and portfolio choice', *Journal of Applied Econometrics* **32**(1), 140–158.

Cech, F. & Barunik, J. (2017), 'On the modelling and forecasting of multivariate realized volatility: generalized heterogeneous autoregressive (ghar) model', *Journal of Applied Econometrics* **36**, 181–206.

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

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

Ding, Z. & Engle, R. (2001), Large scale conditional covariance matrix modeling, estimation and testing. NYU Working Paper No. FIN-01-029.

Engle, R., Ng, V. K. & Rothschild, M. (1990), 'Asset pricing with a factor-arch covariance structure: Empirical estimates for treasury bills', *Journal of Econometrics* **45**(1-2), 213–237.

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

Fleming, J., Kirby, C. & Ostdiek, B. (2003), 'The economic value of volatility timing using realized volatility', *Journal of Financial Economics* **67**(3), 473 – 509.

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.

Hansen, P. R., Lunde, A. & Voev, V. (2014), 'Realized beta GARCH: A multivariate GARCH model with realized measures of volatility', *Journal of Applied Econometrics* **29**(5), 774–799.

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

Jin, X. & Maheu, J. M. (2016), 'Bayesian semiparametric modeling of realized covariance matrices', *Journal of Econometrics* **192**(1), 19–39.

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

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

Noureldin, D., Shephard, N. & Sheppard, K. (2014), 'Multivariate rotated arch models', *Journal of Econometrics* **179**(1), 16 – 30.

Opschoor, A., Janus, P., Lucas, A. & Dijk, D. V. (2017), 'New heavy models for fat-tailed realized covariances and returns', *forthcoming Journal of Business & Economic Statistics* pp. 1–15.

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

Press, S. J. (2012), *Applied multivariate analysis: using Bayesian and frequentist methods of inference*, Courier Dover Publications.

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

Shen, K., Yao, J. & Li, W. K. (2015), 'Forecasting high-dimensional realized volatility matrices using a factor model', *arXiv preprint arXiv:1504.03454* .

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

Tao, M., Wang, Y., Yao, Q. & Zou, J. (2011), 'Large volatility matrix inference via combining low-frequency and high-frequency approaches', *Journal of the American Statistical Association* **106**(495), 1025–1040.

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

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

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

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

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

Table 1: Cumulative log-predictive likelihoods for RCOV: 10 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
W		−39587.22	−40777.67	−43263.63	−47009.30	−55978.90
IW		−32220.45	−34437.87	−36233.11	−39232.62	−46941.53
IW-DPM		−27451.69	**−28012.17**	−28572.06	−29680.07	−32418.61
IW-IHMM		**−27039.28**	−28023.37	**−28550.82**	**−29319.11**	**−31061.45**
IW-F	1	−46251.77	−46376.24	−46502.25	−46768.92	−47529.14
IW-DPM-F	1	−30652.93	−31039.37	−31359.83	−31859.05	−33403.69
IW-IHMM-F	1	−29395.15	−30141.83	−30668.12	−31434.20	−33418.10
IW-F	3	−40447.39	−40909.48	−41317.97	−41982.70	−43654.05
IW-DPM-F	3	−30170.77	−30601.00	−30880.66	−31530.97	−33097.95
IW-IHMM-F	3	−28987.33	−29794.05	−30262.32	−30921.79	−32651.90
IW-A-F	5	−36694.25	−37543.93	−38252.74	−39504.19	−42588.04
IW-DPM-F	5	−29713.09	−30177.69	−30482.99	−31206.87	−32989.21
IW-IHMM-F	5	−28624.16	−29589.16	−30011.87	−30726.58	−32371.98
IW-A-F	7	−34385.79	−35686.38	−36732.01	−38367.80	−42880.34
IW-DPM-F	7	−29105.27	−29706.59	−30040.01	−30652.80	−32027.02
IW-IHMM-F	7	−28455.72	−29414.00	−29900.35	−30511.38	−31664.46
IW-A-F	9	−32276.83	−34042.08	−35446.49	−37816.35	−43786.33
IW-DPM-F	9	−28104.46	−28918.04	−29411.33	−30253.61	−32261.74
IW-IHMM-F	9	−27604.53	−28664.36	−29252.26	−29929.01	−31541.95
RCOV discount		−53684.46	−51750.10	−51128.38	−52244.12	−66905.64
COV discount		−60581.88	−58634.20	−57858.83	−58412.99	−72091.22

The table reports the cumulative log-predictive likelihoods for RCOV at different forecast horizon h. Bold entries denote the maximum value in each column.

Table 2: Root mean squared error for predictive mean of RCOV: 10 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
W		85.4929	94.7907	99.6165	106.0295	113.7032
IW		85.3840	94.5600	99.0906	108.6184	122.6522
IW-DPM		85.9544	89.9484	93.7687	96.5692	**102.2756**
IW-IHMM		78.6086	86.1769	91.1179	96.3775	102.3623
IW-F	1	89.7525	94.7418	97.1521	101.5164	106.2680
IW-DPM	1	91.1326	99.5026	101.5357	103.7106	106.8961
IW-IHMM-F	1	85.8326	91.2860	94.5455	98.6635	104.8381
IW-F	3	87.6111	94.6461	99.1480	108.3969	118.5627
IW-DPM-F	3	84.3923	89.3493	93.4872	98.5372	105.5588
IW-IHMM-F	3	80.0110	87.6508	91.8337	97.7670	105.2076
IW-F	5	86.2510	93.3480	97.3201	107.0393	123.0651
IW-DPM-F	5	84.8369	89.5407	92.9232	99.6109	107.7531
IW-IHMM-F	5	78.2897	87.4064	90.9256	98.4061	105.8242
IW-F	7	85.8719	92.9670	96.8803	106.6631	123.8840
IW-DPM-F	7	85.4245	90.2819	93.6968	100.9226	109.3621
IW-IHMM-F	7	78.1382	86.7199	**90.3256**	96.9410	104.3199
IW-F	9	85.2877	92.5723	96.4360	106.2268	124.6541
IW-DPM-F	9	84.3183	88.6977	91.7675	97.0189	102.9021
IW-IHMM-F	9	**78.0443**	**85.9834**	90.5786	**95.9625**	102.4482
RCOV discount		95.9803	102.6128	109.1086	122.9308	166.6344
EWMA		89.9313	95.3414	99.8233	109.1673	126.9490
RW		104.9093	116.8036	119.6874	132.8902	150.2774
COV discount		112.2610	120.1382	127.8039	142.6002	194.9500
VDGARCH-t		105.6908	108.7505	111.7816	116.6325	128.4878

The table reports the root mean squared error for predictive mean of RCOV at different forecast horizon h. Bold entries denote the minimum value in each column.

Table 3: Cumulative log-predictive likelihoods for return: 10 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
W		−9507.10	−9695.59	−9811.81	−9958.57	−10395.33
IW		−9493.77	−9590.63	−9644.23	−9758.60	−9983.27
IW-DPM		−9512.61	**−9551.81**	**−9588.65**	**−9644.56**	**−9905.66**
IW-IHMM		**−9458.91**	−9575.71	−9661.51	−9785.24	−10097.68
IW-F	1	−10075.89	−10087.52	−10100.82	−10137.81	−10284.61
IW-DPM-F	1	−10195.24	−10226.87	−10252.75	−10302.68	−10477.34
IW-IHMM-F	1	−9687.93	−9779.76	−9874.48	−10016.97	−10354.25
IW-F	3	−9900.79	−9927.72	−9951.04	−10004.66	−10211.77
IW-DPM-F	3	−9951.60	−10020.96	−10075.63	−10190.52	−10465.70
IW-IHMM-F	3	−9609.40	−9722.48	−9803.60	−9927.11	−10328.85
IW-F	5	−9816.80	−9855.76	−9874.96	−9951.12	−10222.07
IW-DPM-F	5	−9802.46	−9866.94	−9909.53	−10043.91	−10391.41
IW-IHMM-F	5	−9548.87	−9669.23	−9749.29	−9913.57	−10253.53
IW-F	7	−9707.51	−9772.03	−9809.69	−9871.08	−10100.78
IW-DPM-F	7	−9687.82	−9760.15	−9808.84	−9917.68	−10191.76
IW-IHMM-F	7	−9501.41	−9637.25	−9699.16	−9802.82	−10090.62
IW-F	9	−9612.21	−9681.41	−9723.69	−9807.58	−10073.27
IW-DPM-F	9	−9607.70	−9690.33	−9751.33	−9862.03	−10275.72
IW-IHMM-F	9	−9475.08	−9606.33	−9686.33	−9811.01	−10177.86
RCOV discount		−9606.93	−9688.38	−9734.20	−9826.47	−10025.77
COV discount		−9803.76	−9923.18	−9960.57	−10000.86	−10239.58
VDGARCH-t		−9621.82	−9682.01	−9721.86	−9791.10	−10021.33

The table reports the cumulative log-predictive likelihoods for return data at different forecast horizon h. Bold entries denote the maximum value in each column.

Table 4: Cumulative log-predictive likelihoods for RCOV:60 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW-F	1	1017209.89	1016664.26	1015373.57	1011318.86	994026.22
IW-IHMM-F	1	1396307.46	1393526.11	1391401.42	1388192.27	1379376.08
IW-F	3	1056464.66	1054908.38	1051632.18	1043723.95	1020536.18
IW-IHMM-F	3	1398145.06	**1395433.30**	**1393253.92**	1389954.88	1381750.23
IW-F	5	1094748.58	1093189.46	1091257.50	1085498.59	1062521.66
IW-IHMM-F	5	1397132.83	1393931.81	1391398.69	1387744.92	1378840.02
IW-F	7	1114075.73	1112050.08	1109570.61	1103094.53	1077736.13
IW-IHMM-F	7	1398474.51	1394977.69	1392512.49	1389660.87	1381604.46
IW-F	10	1132220.41	1129380.22	1126290.99	1118661.56	1088156.16
IW-IHMM-F	10	**1398625.61**	1395204.87	1392819.30	**1389955.37**	**1382475.26**
RCOV discount		1184730.03	1155150.48	1128348.22	1077493.40	851783.83
COV discount		359270.91	315493.21	225129.79	7853.87	−945501.06

The table reports the cumulative log-predictive likelihoods for RCOV at different forecast horizon h. Bold entries denote the maximum value in each column.

Table 5: Root mean squared error for predictive mean of RCOV:60 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW- F	1	51.0828	52.5740	53.3587	54.0758	53.3842
IW-IHMM-F	1	45.5283	45.4724	45.3579	45.4689	46.2941
IW-F	3	47.4160	49.1300	49.9778	50.8249	50.9169
IW-IHMM-F	3	45.2321	45.2645	45.1911	45.2504	45.8372
IW-F	5	46.3333	47.9883	48.6960	49.4046	48.9379
IW-IHMM-F	5	44.9402	45.1712	45.1496	45.2417	45.6632
IW-F	7	45.8625	47.4414	48.1200	48.7621	48.3987
IW-IHMM-F	7	44.9541	45.1784	45.1306	45.1544	45.6527
IW-F	10	45.4325	46.9872	47.6369	48.1953	47.9312
IW-IHMM-F	10	**44.9076**	**45.1489**	**45.0933**	**45.1522**	**45.5994**
RCOV discount		46.0742	47.2496	48.1508	49.3480	52.8361
EWMA		47.1797	47.4266	47.5095	47.7065	47.7802
RW		62.0047	65.3585	66.7280	66.9556	64.6790
COV discount		48.4982	50.0729	51.0833	52.3614	57.4294
VDGARCH-t(returns)		87.8707	89.5497	91.5198	95.2122	107.1460

The table reports the root mean squared error for predictive mean of RCOV at different forecast horizon h. Bold entries denote the minimum value in each column.

Table 6: Cumulative log-predictive likelihoods for return:60 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW-F	1	−35770.93	−35774.67	−35795.67	−35833.23	−35877.51
IW-IHMM-F	1	−33784.05	−33853.80	−33907.43	**−33980.83**	−34098.24
IW-F	3	−35644.59	−35640.39	−35675.57	−35734.03	−35819.71
IW-IHMM-F	3	−33758.26	−33832.48	**−33907.23**	−33982.38	**−34092.31**
IW-F	5	−35500.49	−35501.91	−35523.82	−35578.79	−35652.32
IW-iHMM-F	5	−33752.66	−33833.63	−33919.43	−34014.91	−34144.16
IW-F	7	−35373.08	−35378.01	−35403.69	−35462.16	−35555.23
IW-IHMM-F	7	−33742.93	**−33817.22**	−33907.29	−34007.32	−34158.26
IW-F	10	−35266.59	−35280.06	−35307.46	−35368.11	−35479.17
IW-IHMM-F	10	**−33733.78**	−33824.80	−33921.19	−34033.74	−34175.87
RCOV discount		−34387.43	−34635.97	−34648.43	−34701.31	−34861.64
COV discount		−49411.45	−49762.37	−50359.85	−51631.38	−59408.23
VDGARCH-t		−34225.88	−34307.13	−34359.42	−34474.16	−34811.07

The table reports the cumulative log-predictive likelihoods for return data at different forecast horizon h. Bold entries denote the maximum value in each column.

Table 7: Sample mean of RV of global minimum variance portfolios: 60 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW-F	1	0.3364	0.3382	0.3434	0.3532	0.3522
IW-IHMM-F	1	0.3219	0.3221	0.3219	0.3232	0.3266
IW-F	3	0.3335	0.3329	0.3348	0.3423	0.3369
IW-IHMM-F	3	0.3225	0.3211	0.3221	0.3226	0.3241
IW-F	5	0.3312	0.3291	0.3314	0.3411	0.3349
IW-IHMM-F	5	0.3193	0.3218	0.3239	0.3254	0.3271
IW-F	7	0.3288	0.3259	0.3267	0.3377	0.3324
IW-IHMM-F	7	**0.3171**	0.3202	0.3217	0.3228	0.3248
IW-F	10	0.3255	0.3238	0.3243	0.3366	0.3316
IW-IHMM-F	10	**0.3171**	**0.3196**	**0.3211**	**0.3221**	**0.3230**
RCOV discount		0.3475	0.3586	0.3635	0.3746	0.3828
EWMA		0.3290	0.3420	0.3496	0.3590	0.3669
RW		0.3787	0.4513	0.4584	0.4669	0.4561
COV discount		0.7780	0.7617	0.7866	0.7822	0.8516
VDGARCH-t		0.3723	0.3710	0.3697	0.3688	0.3753

The table reports the sample mean of RV of global minimum variance portfolios (GMVP) against forecast horizon h for various models. Bold entries denote the minimum value in each column.

Table 8: Estimates of ℓ_2, ℓ_3 and K, 60-asset data

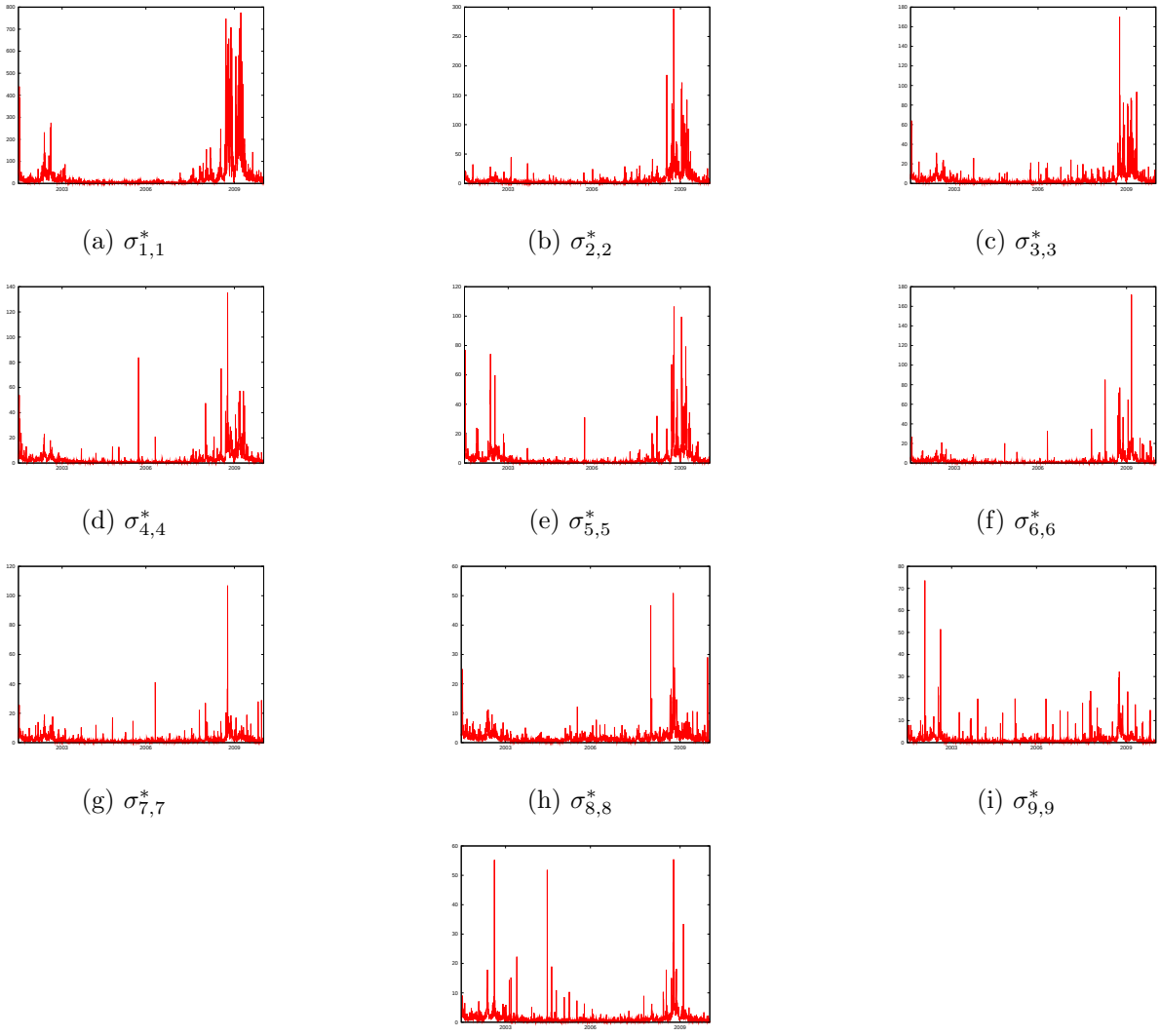
		ℓ_2		ℓ_3		K
Model	Factors	Mean	0.95DI	Mean	0.95DI	Mean
IW-F	1	2.00	$(2, 2)$	15.88	$(15, 16)$	
IW-IHMM-F	1	2.00	$(2, 2)$	14.98	$(14, 16)$	16.00
IW-F	3	2.00	$(2, 2)$	16.00	$(16, 16)$	
IW-IHMM-F	3	10.00	$(10, 10)$	80.65	$(79, 81)$	14.00
IW-F	5	11.00	$(11, 11)$	83.00	$(83, 83)$	
IW-IHMM-F	5	9.00	$(9, 9)$	66.22	$(66, 68)$	15.00
IW-F	7	11.00	$(11, 11)$	83.00	$(83, 83)$	
IW-IHMM-F	7	9.00	$(9, 9)$	43.03	$(43, 44)$	15.00
IW-F	10	11.00	$(11, 11)$	82.93	$(83, 83)$	
IW-IHMM-F	10	11.00	$(11, 11)$	92.21	$(92, 93)$	16.00

K =number of *alive* clusters in the mixture

Table 9: Model running time: 60-asset data

Parametric models	Factors	Run time	Nonparametric models	Factors	Run time
IW-F	1	3m49s	IW-IHMM-F	1	7m46s
IW-F	3	4m7s	IW-IHMM-F	3	7m37s
IW-F	5	4m57s	IW-IHMM-F	5	8m25s
IW-F	7	6m49s	IW-IHMM-F	7	9m55s
IW-F	10	8m37s	IW-IHMM-F	10	12m55s
VDGARCH-t		days			

The table records the running time of 20000 draws of MCMC simulation for each model. All models are estimated on a Linux machine with an Intel Xeon E5-2692 v2 CPU with 12 CPU cores. Parallel computing is implemented whenever possible.



(a) $\sigma^*_{1,1}$ (b) $\sigma^*_{2,2}$ (c) $\sigma^*_{3,3}$

(d) $\sigma^*_{4,4}$ (e) $\sigma^*_{5,5}$ (f) $\sigma^*_{6,6}$

(g) $\sigma^*_{7,7}$ (h) $\sigma^*_{8,8}$ (i) $\sigma^*_{9,9}$

(j) $\sigma^*_{10,10}$

Figure 1: Diagonal elements of Σ^*_t for 10-asset data