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

Xin Jin† John M. Maheu‡ Qiao Yang§

November 2018

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 the Editor, two anonymous referees, 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] Once market microstructure effects are accounted for RCOV can provide an accurate measure of ex post covariation that is observable and time-series methods can be applied directly to the data to capture their conditional distribution. However, RCOV matrices 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 fat tails is shown in Jin & Maheu (2016) and Opschoor et al. (2017).

Applications of factor 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 a 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. 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 and returns. While estimation of existing models of RCOV that provide density forecasts are not computationally feasible for large realized covariance matrices our approach is. This makes recursive estimation and forecasting practical.

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, as well as other existing models for RCOV, 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.[2] 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. Parametric factor models are discussed in Section 2 followed by their nonparametric extensions in Section 3. Application to 10 asset RCOV data and 60 asset RCOV data are in Section 4 followed by the conclusion. An online appendix collects additional results and full posterior simulation details for estimation.

2 Parametric Factor Models of RCOV

Let $\Sigma_t,\ t = 1, 2, \ldots, T$, denote a time-series of $k \times k$ realized covariance matrices and define $\Sigma_{1:t} = \{\Sigma_1, \ldots, \Sigma_t\}$. 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 (Press 2012):

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

[2]Benchmark models include rotated ARCH models (Noureldin et al. 2014) based on daily returns and extensions of the RCOV model of Yu et al. (2017) to accommodate larger dimensions.

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[3] to V gives

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

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 $W'W = WW' = I$.[4] Define the orthogonally transformed Σ_t denoted as Σ_t^* by

$$\Sigma_t^* = W'\Sigma_t W. \tag{2}$$

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^* with respect to time is the diagonal matrix D by definition:

$$E(\Sigma_t^*) = E(W'\Sigma_t W) = W'E(\Sigma_t)W = W'VW = D. \tag{3}$$

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.[5] Under this ordering scheme,

[3]Matrix decompositions are common in large dimension volatility modeling. Shirota et al. (2017) use the components of a Cholesky decomposition of RCOV in a multivariate stochastic volatility model. Factor stochastic volatility models (Kastner 2018) decompose the covariance matrix into a factor loading matrix and two diagonal matrices to reduce the number of parameters.

[4]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.

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

$$g_i = \mathrm{Var}(\sigma_{t,ii}^*) = \mathrm{Var}(w_i'\Sigma_t w_i), \tag{4}$$

which is like the variance of the realized variance of a portfolio but with weight vector w_i and condition $w_i'w_i =$

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^*. 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.

The orthogonal transformation proposed above is related to the approach taken by Noureldin et al. (2014) in the context of multivariate GARCH modeling, which applies a rotation to the raw return vector and then fits the rotated return with a multivariate GARCH specification, resulting in the rotated ARCH(RARCH) model class.[6]

2.1 Block-Diagonal Factor Model (IW-F)

This section introduces a factor model based on the inverse-Wishart distribution. The factor model applies to the Wishart distribution as well, although we will focus attention on the inverse-Wishart version. Partition Σ_t^* into blocks

$$\Sigma_t^* = \begin{pmatrix} \Sigma_{t,11}^* & \Sigma_{t,21}^{*\prime} \\ \Sigma_{t,21}^* & \Sigma_{t,22}^* \end{pmatrix}, \tag{5}$$

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 $k_1 > 0, k_2 \geq 0, k_1 + k_2 = k$. In the IW-F model the conditional distribution of Σ_t^* is specified as follows:

$$f(\Sigma_t^* | \Sigma_{1:t-1}^*, \nu, C, \Theta) = \text{Wishart}_k^{-1}(\Sigma_t^* | \nu, (\nu - k - 1)V_t), \tag{6}$$

$$V_t = \begin{pmatrix} V_t^* & 0 \\ 0 & C \end{pmatrix}, \; V_t^* = B_0 + \sum_{j=1}^{M} B_j \odot \Gamma_{t-1,\ell_j}^*, \; \Gamma_{t-1,\ell_j}^* = \frac{1}{\ell_j} \sum_{i=1}^{\ell_j} \Sigma_{t-i,11}^*. \tag{7}$$

$\text{Wishart}_k^{-1}(.|\nu, (\nu - k - 1)V_t)$ 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

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 based on forecasting results from log-predictive likelihoods for RCOV.

[6] In our model the spectral decomposition targets the unconditional mean of RCOV and not the unconditional covariance matrix of returns. Furthermore, our factor structure is introduced to achieve dimension reduction.

matrix equal to $(\nu - k - 1)V_t$. 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 $B_0, b_1, \ldots, b_M, \ell_2, \ldots, \ell_M$. Compared to the parametric model in Jin & Maheu (2016) 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.[7] B_0 is a $k_1 \times k_1$ symmetric positive-definite matrix, and b_j's are $k_1 \times 1$ 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. Following previous work we restrict attention to three components, $M = 3$, as there are no forecast gains from larger values.

By the properties of the inverse-Wishart distribution, the conditional mean of Σ_t^* is $\text{E}(\Sigma_t^*|\Sigma^*_{1:t-1}, \nu, C, \Theta) = V_t$, and the conditional second moments are (Press 2012)

$$\text{Cov}(\Sigma^*_{t,ij}, \Sigma^*_{t,lm}|\Sigma^*_{1:t-1}, \nu, C, \Theta) = \frac{2V_{t,ij}V_{t,lm} + (\nu - k - 1)(V_{t,il}V_{t,jm} + V_{t,im}V_{t,jl})}{(\nu - k)(\nu - k - 3)}, \quad (8)$$

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

V_t^* can be viewed as the set of observable dynamic factors, which contains $k_1(k_1 + 1)/2$ unique scalar elements and satisfy $E(\Sigma^*_{t,11}|\Sigma^*_{1:t-1}) = V_t^*$. Meanwhile, textbook properties of the inverse-Wishart distribution imply $\Sigma^*_{t,11}$ conditionally follows an inverse-Wishart with dimension $k_1 \times k_1$, $\Sigma^*_{t,11}|\Sigma^*_{1:t-1}, \nu, \Theta \sim \text{Wishart}^{-1}_{k_1}(\nu - k_2, (\nu - k - 1)V_t^*)$.

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,

$$\sigma^*_{t,ii}|\Sigma^*_{1:t-1}, \nu, \Theta \;\sim\; \text{Gamma}^{-1}\left(\frac{\nu - k + 1}{2}, \frac{\nu - k - 1}{2}c_{i-k_1}\right), \quad i = k_1 + 1, \ldots, k. \quad (9)$$

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 (3) 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

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

elements of $\Sigma^*_{t,11}$ have zero unconditional mean to satisfy (3). 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 (3) 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. Evaluation of the inverse-Wishart density using a Cholesky decomposition to compute the determinant of V_t has a computational complexity of $O(k^3)$ 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

$$\Sigma_t | V_t, \nu, \Theta \sim \text{Wishart}_k^{-1}(\nu, (\nu - k - 1) W V_t W^{'}). \tag{10}$$

W can be interpreted as factor loadings and imply

$$E[\Sigma_t | V_t, \nu, \Theta] \;\; = \;\; W V_t W^{'} = W_1 V_t^* W_1^{'} + W_2 C W_2^{'}, \tag{11}$$

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

$$r_t | \Sigma_t, \mathcal{F}_{t-1} \sim N(0, \Sigma_t), \tag{12}$$

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 $\text{Var}(r_t) = E(r_t r_t') = E(E(r_t r_t' | \Sigma_t)) = E(\Sigma_t) = V$. The $t-1$ conditional variance of r_t is

$$\begin{aligned} \text{Var}(r_t | \mathcal{F}_{t-1}) &= E(r_t r_t' | \mathcal{F}_{t-1}) = E(E(r_t r_t' | \mathcal{F}_{t-1}, \Sigma_t) | \mathcal{F}_{t-1}) = E(\Sigma_t | \mathcal{F}_{t-1}) \\ &= W V_t W' = W_1 V_t^* W_1' + W_2 C W_2'. \end{aligned} \tag{13}$$

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 $r_t^* \equiv W' r_t$. Then $\text{Var}(r_t^*|\Sigma_t) = W'\Sigma_t W = \Sigma_t^*$, 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 $r_t^* = (r_{t,1}^{*\prime}, r_{t,2}^{*\prime})'$, where $r_{t,1}^*$ is $k_1 \times 1$ and $r_{t,2}^*$ is $k_2 \times 1$. This model imposes the restrictions

$$\text{Var}(r_{1,t}^*|\mathcal{F}_{t-1}) = V_t^*,\ \text{Var}(r_{2,t}^*|\mathcal{F}_{t-1}) = C,\ \text{Cov}(r_{1,t}^*, r_{2,t}^*|\mathcal{F}_{t-1}) = \mathbf{0}_{[k_1 \times k_2]}. \tag{14}$$

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.

2.2 Model inference

To implement the transformation in (2) we apply a spectral decomposition to the sample mean, $\overline{\Sigma} = WDW'$, 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.[8]

For each of the models we implement RCOV targeting for B_0. For IW-F, we set $B_0 = (\iota\iota' - B_1 - \cdots - B_M) \odot \overline{\Sigma}_{11}^*$. 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.

The joint posterior distribution is proportional to $p(\nu)p(\Theta)f(\Sigma_{1:T}^*|\nu, C, \Theta)$. 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 (10), 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. For example, the

[8] Asset order in multivariate models can affect results. For a discussion of this see Chan et al. (2018) and Kastner et al. (2017).

conditional density of Σ_t^* in the IW-F model is

$$\begin{aligned} f(\Sigma_t^*|\Sigma_{1:t-1}^*,\nu,C,\Theta) &= \text{Wishart}_k^{-1}(\Sigma_t^*|\nu,(\nu-k-1)V_t) \\ &= \frac{(\nu-k-1)^{\frac{k\nu}{2}}|V_t|^{\frac{\nu}{2}}|\Sigma_t^*|^{-\frac{\nu+k+1}{2}}}{2^{\frac{\nu k}{2}}\Gamma_k(\frac{\nu}{2})}\exp\left(-\frac{1}{2}tr((\nu-k-1)V_t\Sigma_t^{*-1})\right) \\ &= \frac{(\nu-k-1)^{\frac{k\nu}{2}}|V_t^*|^{\frac{\nu}{2}}|C|^{\frac{\nu}{2}}|\Sigma_t^*|^{-\frac{\nu+k+1}{2}}}{2^{\frac{\nu k}{2}}\Gamma_k(\frac{\nu}{2})} \\ &\quad\times\exp\left(-\frac{\nu-k-1}{2}tr(V_t^*Y_{t,11})\right)\times\exp\left(-\frac{\nu-k-1}{2}tr(CY_{t,22})\right), \end{aligned} \tag{15}$$

where $Y_t = \begin{pmatrix} Y_{t,11} & Y_{t,12} \\ Y_{t,21} & Y_{t,22} \end{pmatrix} = \Sigma_t^{*-1}$. The last step of (15) uses the fact that the determinant of a block diagonal square matrix is equal to the products of the determinants of the diagonal blocks so that $tr(V_t\Sigma_t^{*-1}) = tr(V_t^*Y_{t,11}) + tr(CY_{t,22})$. As a result, the likelihood function of $\Sigma_{1:T}^*$ is

$$\begin{aligned} f(\Sigma_{1:T}^*|\nu,C,\Theta) &= \prod_{t=1}^T f(\Sigma_t^*|\Sigma_{1:t-1}^*,\nu,C,\Theta) \\ &= \frac{\prod_{t=1}^T|V_t^*|^{\frac{\nu}{2}}\prod_{t=1}^T|\Sigma_t^*|^{-\frac{\nu+k+1}{2}}}{2^{\frac{T\nu k}{2}}\Gamma_k(\frac{\nu}{2})^T}\exp\left(-\frac{\nu-k-1}{2}tr(\sum_{t=1}^T V_t^*Y_{t,11})\right) \\ &\quad\times(\nu-k-1)^{\frac{Tk\nu}{2}}|C|^{\frac{T\nu}{2}}\exp\left(-\frac{\nu-k-1}{2}tr(C\sum_{t=1}^T Y_{t,22})\right). \end{aligned} \tag{16}$$

Compared with the likelihood function for the non-factor IW model in Jin & Maheu (2016), (16) 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 (16) 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

$$p(C|\Sigma^*_{1:T}, \nu, \Theta) \quad \propto \quad p(C)f(\Sigma^*_{1:T}|\nu, C, \Theta) \propto \text{Wishart}_{k2}(C|\overline{\gamma}_C, \overline{Q}_C), \tag{17}$$

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

$$p(\Sigma_t|\Sigma_{1:t-1}) \approx \frac{1}{N}\sum_{i=1}^{N} \text{Wishart}_k^{-1}(\Sigma_t|\nu^{(i)}, (\nu^{(i)} - k - 1)WV_t^{(i)}W'), \tag{18}$$

where N denotes the total number of posterior draws and $V_t^{(i)}$ is from (7) 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 (12) and integrating Σ_t out, can be approximated as

$$p(r_t|\Sigma_{1:t-1}) \approx \frac{1}{N}\sum_{i=1}^{N} \text{St}_k\left(r_t|0, \frac{\nu^{(i)} - k - 1}{\nu^{(i)} - k + 1}WV_t^{(i)}W', \nu^{(i)} - k + 1\right). \tag{19}$$

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.

3 Nonparametric Factor Models

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

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

$$
\begin{aligned}
f(\Sigma_t^*|\Sigma_{1:t-1}^*, \Theta, \Omega, \Phi) &= \sum_{j=1}^{\infty} \omega_j \text{Wishart}_k^{-1}(\Sigma_t^*|\nu_j, (\nu_j - k - 1)V_{t,j}), && (20)\\
V_{t,j} &= \begin{pmatrix} V_t^{*1/2} A_j (V_t^{*1/2})' & 0 \\ 0 & C_j \end{pmatrix}, && (21)\\
\Omega &\sim \mathbf{SBP}(\alpha), \qquad (\nu_j, A_j, C_j) \overset{iid}{\sim} G_0, \quad j = 1, 2, \ldots, && (22)
\end{aligned}
$$

where $\Omega = \{\omega_j\}_{j=1}^{\infty}$, $\Phi = \{\phi_j\}_{j=1}^{\infty} = \{(\nu_j, A_j, C_j)\}_{j=1}^{\infty}$, and V_t^* is defined the same as in the parametric factor model. $\mathbf{SBP}(\alpha)$ denotes the stick-breaking construction of the weights: $\omega_j = v_j \prod_{l<j}(1 - v_l), v_j \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad j = 1, 2, \ldots$. As long as the context is clear we use this notation to denote a distribution with support on the natural numbers, $\Omega \sim \mathbf{SBP}(\alpha)$.

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

$$
\Sigma_{t,11}^*|\Sigma_{1:t-1}^*, \Theta, \Omega, \Phi \sim \sum_{j=1}^{\infty} \omega_j \text{Wishart}_{k_1}^{-1}(\nu_j - k_2, (\nu_j - k - 1)V_t^{*1/2} A_j (V_t^{*1/2})'), \tag{23}
$$

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

$$
\Sigma_{t,22}^*|\Sigma_{1:t-1}^*, \Theta, \Omega, \Phi \sim \sum_{j=1}^{\infty} \omega_j \text{Wishart}_{k_2}^{-1}(\nu_j - k_1, (\nu_j - k - 1)C_j). \tag{24}
$$

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, 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,

$$f(\Sigma_t|\Sigma_{1:t-1},\Theta,\Omega,\Phi,W) \;=\; \sum_{j=1}^{\infty}\omega_j \text{Wishart}_k^{-1}(\Sigma_t|\nu_j,(\nu_j-k-1)WV_{t,j}W'). \tag{25}$$

The conditional mean is

$$E(\Sigma_t|\Sigma_{1:t-1},\Theta,\Omega,\Phi,W) = W_1\left[\sum_{j=1}^{\infty}\omega_j V_t^{*1/2}A_j(V_t^{*1/2})'\right]W_1' + W_2\left[\sum_{j=1}^{\infty}\omega_j C_j\right]W_2'. \tag{26}$$

Under (12) and (25), the conditional distribution of r_t, after integrating out Σ_t, is an infinite mixture of multivariate Student-t,

$$f(r_t|\mathcal{F}_{t-1},\Theta,\Omega,\Phi,W) = \sum_{j=1}^{\infty}\omega_j \text{St}_k\left(r_t\middle|0,\frac{\nu_j-k-1}{\nu_j-k+1}WV_{t,j}W',\nu_j-k+1\right), \tag{27}$$

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.

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

$$G_0(\nu_j,A_j,C_j) \;\equiv\; \text{Exp}_{\nu>k+1}(\lambda)\times\text{Wishart}_{k_1}\left(\gamma_A,\frac{1}{\gamma_A}I\right)\times\text{Wishart}_{k_2}\left(\gamma_C,\frac{1}{\gamma_C}I\right), \tag{28}$$

where $\gamma_A \geq k_1, \gamma_C \geq k_2$; 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 independent Wishart distributions, respectively. Note that the mean of A_j satisfies $\text{E}(A_j) = I$. 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, $\alpha \sim \text{Gamma}(a_0,c_0)$.

3.1.1 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).[9] This samples from the stick-breaking representation of the infinite mixture model by introducing a slice variable

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

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

$$f(\Sigma^*_t|\Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \omega_j h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j), \tag{29}$$

where $h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j)$ corresponds to either the inverse-Wishart in (20) 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

$$f(\Sigma^*_t, u_t|\Theta, \Omega, \Phi) = \sum_{j=1}^{\infty} \mathbf{1}(u_t < \omega_j) h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j). \tag{30}$$

Note that integrating out u_t returns the original model (29). 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

$$f(\Sigma^*_{1:T}, u_{1:T}, s_{1:T}|\Theta, \Omega, \Phi) = \prod_{t=1}^{T} f(\Sigma^*_t, u_t, s_t|\Theta, \Omega, \Phi) = \prod_{t=1}^{T} \mathbf{1}(u_t < \omega_{s_t}) h(\Sigma^*_t|\Theta, \nu_{s_t}, A_{s_t}, C_{s_t}), \tag{31}$$

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

$$p(\Theta)p(\Omega_{\overline{K}}) \left[\prod_{i=1}^{\overline{K}} p(\nu_j, A_j, C_j)\right] \prod_{t=1}^{T} \mathbf{1}(u_t < \omega_{s_t}) h(\Sigma^*_t|\Theta, \nu_{s_t}, A_{s_t}, C_{s_t}), \tag{32}$$

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

The posterior sampling steps are as follows.

1. $p(\phi_j|\Sigma^*_{1:T}, s_{1:T}, \Theta) \propto p(\phi_j) \prod_{\{t:s_t=j\}} h(\Sigma^*_t|\Theta, \nu_j, A_j, C_j), j = 1, \ldots, \overline{K}$.
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. $p(\Theta|\Sigma^*_{1:T}, s_{1:T}, \Phi) \propto p(\Theta)\prod_{t=1}^{T} h(\Sigma^*_t|\Theta, \nu_{s_t}, A_{s_t}, C_{s_t})$

Each of the individual steps are detailed in the online appendix. One sweep of the sampler delivers $\{\{(\nu_j, A_j, C_j, v_j)\}_{j=1}^{\overline{K}}, \overline{K}, u_{1:T}, s_{1:T}, \alpha, \Theta\}$.

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

$$G^{(i)} = \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} \delta_{\phi_j^{(i)}} + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) G_0. \tag{33}$$

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. In this paper we use OpenMP (https://www.openmp.org/) in a shared memory setting.

3.1.2 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

$$\begin{aligned}
& p(\Sigma_{T+1}|\Sigma_{1:T}, G^{(i)}, W) \\
= & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}, W) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \int h(\Sigma_{T+1}|\Theta^{(i)}, \phi, W) G_0(d\phi) \quad (34) \\
\approx & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h(\Sigma_{T+1}|\Theta^{(i)}, \phi_j^{(i)}, W) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \frac{1}{R} \sum_{l=1}^{R} h(\Sigma_{T+1}|\Theta^{(i)}, \phi^{[l]}, W), \quad (35)
\end{aligned}$$

where $\phi^{[l]} \overset{iid}{\sim} G_0, l = 1, \ldots, R$. [10] For IW-DPM-F,

$$\begin{aligned} h(\Sigma_{T+1}|\Theta, \phi_j, W) &= \text{Wishart}_k^{-1}(\Sigma_{T+1}|\nu_j, (\nu_j - k - 1)WV_{T+1,j}W') \\ &= \text{Wishart}_k^{-1}(\Sigma^*_{T+1}|\nu_j, (\nu_j - k - 1)V_{T+1,j}) \\ &= h(\Sigma^*_{T+1}|\Theta, \phi_j). \end{aligned} \tag{36}$$

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

$$p(\Sigma_{T+1}|\Sigma_{1:T}, G^{(i)}, W) = p(\Sigma^*_{T+1}|\Sigma^*_{1:T}, G^{(i)}). \tag{37}$$

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

$$p(\Sigma_{T+1}|\Sigma_{1:T}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\Sigma^*_{T+1}|\Sigma^*_{1:T}, G^{(i)}). \tag{38}$$

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

$$\begin{aligned} & p(r_{T+1}|\mathcal{F}_T, G^{(i)}, W) \\ = & \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)} h_r(r_{T+1}|\Theta^{(i)}, \phi_j^{(i)}, W) + \left(1 - \sum_{j=1}^{\overline{K}^{(i)}} \omega_j^{(i)}\right) \int h_r(r_{T+1}|\Theta^{(i)}, \phi, W) G_0(d\phi), \end{aligned} \tag{39}$$

where

$$\begin{aligned} h_r(r_{T+1}|\Theta, \phi_j, W) &= \text{St}_k\left(r_{T+1}\middle|0, \frac{\nu_j - k - 1}{\nu_j - k + 1}WV_{T+1,j}W', \nu_j - k + 1\right) \\ &= \text{St}_k\left(r^*_{T+1}\middle|0, \frac{\nu_j - k - 1}{\nu_j - k + 1}V_{T+1,j}, \nu_j - k + 1\right). \end{aligned} \tag{40}$$

3.2 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

[10] In the empirical work $R = 10$.

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. 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). Extending Jin & Maheu (2016) we propose the following factor model (IW-IHMM-F) for Σ_t:

$$\begin{aligned}
\boldsymbol{\pi}_0|\alpha &\sim \mathbf{SBP}(\alpha), && (41)\\
\boldsymbol{\pi}_i|\boldsymbol{\pi}_0,\beta,\kappa &\sim \mathrm{DP}\left(\beta+\kappa, \frac{\beta\boldsymbol{\pi}_0+\kappa\delta_i}{\beta+\kappa}\right), && (42)\\
\phi_j &\overset{iid}{\sim} G_0, \quad j=1,2,\ldots, && (43)\\
s_t|s_{t-1}=i,\Pi &\sim \boldsymbol{\pi}_i, \quad i=1,2,\ldots, && (44)\\
\Sigma_t^*|\Sigma_{1:t-1}^*,\Theta,\Pi,\Phi,s_t &\sim \mathrm{Wishart}_k^{-1}(\nu_{s_t},(\nu_{s_t}-k-1)V_{t,s_t}), && (45)\\
V_{t,s_t} &= \begin{pmatrix} V_t^{*1/2}A_{s_t}(V_t^{*1/2})' & 0 \\ 0 & C_{s_t}\end{pmatrix}, && (46)
\end{aligned}$$

where $\Phi=\{\phi_j\}_{j=1}^{\infty}=\{(\nu_j,A_j,C_j)\}_{j=1}^{\infty}$, and V_t^* is the same as before. 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 $\mathrm{Wishart}_k^{-1}(\nu_{s_t},(\nu_{s_t}-k-1)V_{t,s_t})$ given s_t and V_{t,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. 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, $\alpha\sim\mathrm{Gamma}(a_3,c_3)$, $\beta+\kappa\sim\mathrm{Gamma}(a_4,c_4)$, $\rho=\frac{\kappa}{\beta+\kappa}\sim\mathrm{Beta}(a_5,c_5)$ which allow for learning from the data. This prior formulation is more convenient for posterior sampling.

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

$$f(\Sigma_t|\Sigma_{1:t-1},\Theta,\Pi,\Phi,W,s_{t-1})=\sum_{s_t=1}^{\infty}\pi_{s_{t-1},s_t}\mathrm{Wishart}_k^{-1}(\Sigma_t|\nu_{s_t},(\nu_{s_t}-k-1)WV_{t,s_t}W'). \quad (47)$$

The conditional mean becomes

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

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 (12) and (47), the conditional distribution of r_t, after integrating out Σ_t, is an infinite mixture of multivariate Student-t with time-varying weights,

$$f(r_t|\mathcal{F}_{t-1},\Theta,\Pi,\Phi,W,s_{t-1}) = \sum_{s_t=1}^{\infty}\pi_{s_{t-1},s_t}\mathrm{St}_k\left(r_t\middle|0,\frac{\nu_{s_t}-k-1}{\nu_{s_t}-k+1}WV_{t,s_t}W',\nu_{s_t}-k+1\right). \quad (48)$$

The online appendix discusses some of the key features of IW-IHMM-F along with some restrictions that significantly reduce computation.

3.2.1 Posterior inference

Similar to the posterior sampling methods for the DPM model of Section 3.1 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

$$p(u_t|s_t,s_{t-1},\Pi) = \frac{\mathbf{1}(u_t < \pi_{s_{t-1},s_t})}{\pi_{s_{t-1},s_t}} \quad (49)$$

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

$$\begin{aligned} p(s_t|u_{1:t},\Sigma_{1:t}^*) &\propto h(\Sigma_t^*|\phi_{s_t})\sum_{s_{t-1}=1}^{\infty}p(u_t|s_t,s_{t-1})p(s_t|s_{t-1})p(s_{t-1}|\Sigma_{1:t-1}^*,u_{1:t-1}) \quad (50)\\ &\propto h(\Sigma_t^*|\phi_{s_t})\sum_{s_{t-1}:u_t<\pi_{s_{t-1},s_t}}p(s_{t-1}|u_{1:t-1},\Sigma_{1:t-1}^*). \quad (51)\end{aligned}$$

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

$$p(s_t|s_{t+1}, \Sigma^*_{1:T}, u_{1:T}) \quad \propto \quad p(s_t|u_{1:t}, \Sigma^*_{1:t})\mathbf{1}(u_{t+1} < \pi_{s_t,s_{t+1}}). \tag{52}$$

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 (52).

It is convenient to find a finite set $\{1, \ldots, \overline{K}\}$ that includes all possible states s_t that satisfy the condition $u_t < \pi_{s_{t-1},s_t}$. Jin & Maheu (2016) give 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: $p(u_{1:T}|s_{1:T}, \Pi)$, $p(s_{1:T}|\Pi, u_{1:T}, \Phi, \Theta, \Sigma^*_{1:T})$, $p(\boldsymbol{\pi}_0|s_{1:T}, \alpha, \beta, \kappa)$, $p(\Pi|\boldsymbol{\pi}_0, s_{1:T}, \beta, \kappa)$, $p(\Phi|s_{1:T}, \Theta, \Sigma^*_{1:T})$, $p(\alpha, \beta, \kappa|s_{1:T}, \boldsymbol{\pi}_0)$, $p(\Theta|s_{1:T}, \Phi, \Sigma^*_{1:T})$. The Appendix provides full details on each of the steps.

3.2.2 Predictive density

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

$$\begin{aligned}
& p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, \Phi^{(i)}, s^{(i)}_{1:T}, \Theta^{(i)}, W) \\
= & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j} h(\Sigma_{T+1}|\Theta^{(i)}, \phi^{(i)}_j, W) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j}\right) \int h(\Sigma_{T+1}|\Theta^{(i)}, \phi, W) G_0(d\phi) \quad (53) \\
= & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j} h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi^{(i)}_j) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j}\right) \int h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi) G_0(d\phi) \quad (54) \\
\approx & \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j} h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi^{(i)}_j) + \left(1 - \sum_{j=1}^{K^{(i)}} \pi^{(i)}_{s^{(i)}_T,j}\right) \frac{1}{R} \sum_{l=1}^{R} h(\Sigma^*_{T+1}|\Theta^{(i)}, \phi^{[l]}), \quad (55)
\end{aligned}$$

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

$$p(\Sigma_{T+1}|\Sigma_{1:T}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\Sigma_{T+1}|\Sigma_{1:T}, \Pi^{(i)}, \Phi^{(i)}, s_{1:T}^{(i)}, \Theta^{(i)}, W), \tag{56}$$

where the right hand side terms are from (55) 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 $\pi_{s_T,j}$.

4 Empirical Applications

4.1 10 Asset Application

In this section we discuss the results for a 10 asset application. The benefit of this smaller dimension example is that we can feasibly estimate different models including the highly parameterized non-factor nonparametric models from Jin & Maheu (2016) and other likelihood based benchmark RCOV 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 used is from Noureldin et al. (2012) and constructed from subsampling based in 5-minute retruns. The data range from 2001/02/01 to 2009/12/31 (2092 observations). The last 500 observations are used for out-of-sample forecast evaluation.

We include the generalized conditional autoregressive Wishart model (GCAW) for RCOV proposed by Yu et al. (2017) as a benchmark. The GCAW model can be seen as a generalization of the Wishart autoregressive model (WAR) of Gourieroux et al. (2009) and the conditional Wishart autoregressive model (CAW) of Golosnoy et al. (2012). It uses a noncentral Wishart distribution with both the noncentrality matrix and the scale matrix driven by the past values of RCOV matrices, and is shown to have superior forecasting performance than both WAR and CAW. The general GCAW(p, q, r) model can be described as:

$$f(\Sigma_t|\Sigma_{1:t-1}, \Theta) = \text{NCW}_k(\Sigma_t|\nu, V_t/\nu, \Lambda_t) \tag{57}$$

$$\Lambda_t = \sum_{i=1}^{r} M_i \Sigma_{t-i} M_i', \quad V_t = LL' + \sum_{i=1}^{p} J_i V_{t-i} J_i' + \sum_{i=1}^{q} U_i \Sigma_{t-i} U_i'. \tag{58}$$

$\text{NCW}_k(.|\nu, V_t/\nu, \Lambda_t)$ denotes a non-central Wishart density over positive definite matrices of dimension k and ν is the real-valued degree of freedom. V_t/ν and Λ_t are the scale matrix and

the noncentrality matrix, respectively. L is a $k \times k$ lower triangular matrix and J_i, U_i, M_i are $k \times k$. Thus, the likelihood function is a product of the non-central Wishart densities,

$$p(\Sigma_{1:T}|\Theta) = \prod_{t=1}^{T} \text{NCW}_k(\Sigma_t|\nu, V_t/\nu, \Lambda_t) = \prod_{t=1}^{T} \frac{\nu^{\frac{k\nu}{2}}|\Sigma_t|^{\frac{\nu-k-1}{2}}|V_t|^{-\frac{\nu}{2}}}{2^{\frac{\nu k}{2}}\Gamma_k(\frac{\nu}{2})} \exp\Big(-\frac{\nu}{2}Tr[V_t^{-1}(\Sigma_t + \Lambda_t)]\Big) \times {}_0F_1(\nu;\ \frac{\nu^2}{4}V_t^{-1}\Lambda_t V_t^{-1}\Sigma_t), \tag{59}$$

where ${}_0F_1$ is the hypergeometric function of matrix argument. Note that in addition to $|V_t|$ which is $\mathcal{O}(k^3)$ in computation, evaluating ${}_0F_1(.;.)$ requires singular value decomposition of multiple matrix multiplication products, which is also of at least $\mathcal{O}(k^3)$ computations even with block-diagonalized V_t and Λ_t. This makes GCAW not only impractical for large dimensions but also unsuitable for adopting the factor structure proposed in this paper. We consider a GCAW$(2, 2, 1)$ model with diagonal J_i, U_i, M_i.

A special case of GCAW with $\Lambda_t = 0$ makes ${}_0F_1$ vanish and reduces the non-cental Wishart density to a Wishart, so the GCAW model becomes the CAW model. We also include a CAW$(2, 2)$ model in the 10 asset application.

Model evaluations of density forecasts in terms of predictive likelihoods and point forecasts in terms of root-mean squared forecast error (RMSFE) are carried out over the out-of-sample data for different forecast horizons h. In particular, the cumulative log-predictive likelihoods 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. Each model is re-estimated at each day in the out-of-sample period. Parametric and nonparametric factor models with factor dimensions from 1 to 9 are compared against non-factor models including the benchmarks. Results are reported in Tables 1 and 2.[11]

For density forecast the IW-IHMM performs the best. This model strongly dominates all the parametric 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.

Meanwhile, the benchmarks GCAW and CAW perform very poorly, not only being the

[11]The IW is a non-factor inverse-Wishart model, the IW-DPM and IW-IHMM are non-factor nonparametric models. See the online appendix for more details.

worst among non-factor models, but also beaten by all nonparametric factor models and parametric factor models with more than 3 factors.

Turning to point forecasts based on the predictive mean, the IHMM factor models with 5 or more factors achieve the lowest RMSFE. The IHMM version is generally much better than the DPM version or parametric versions. The GCAW and CAW are more competitive in point forecasts with lower RMSFE than other parametric models, but the IW-IHMM-F with 3 or more factors prevails.

We note the following observations. The nonparametric models, particularly the IHMM version offer large improvements in both measures of forecast accuracy. Factor models represent a compromise and diminished forecast accuracy compared to the full nonparametric models. However, even a 3 factor IW-IHMM-F dominates all benchmark 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.

In larger dimensions the IW-F and IW-DPM and IW-IHMM are not practically feasible while the factor models are. We turn to a more challenging application next.

4.2 60 Asset Application

For the second dataset, we use high-frequency transaction prices of 60 liquid stocks[12] 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 according to Barndorff-Nielsen et al. (2011), 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, IW, IW-DPM, IW-IHMM and the benchmark GCAW are no longer feasible to estimate and forecast with. This is also the case with CAW in its original form due to the high computation cost for $|V_t|$. One simple solution is to make V_t a diagonal matrix in the CAW model, thus computing $|V_t|$ becomes $\mathcal{O}(k)$. But this assumes both the

[12]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.

conditional and the unconditional mean of the off-diagonal elements of Σ_t are zero, which obviously is too unrealistic. As a remedy, we propose to again first transform the original Σ_t into Σ_t^* and then fit Σ_t^* using the CAW model with diagonal V_t, which will at least match the unconditional moment condition. We call this model transformed diagonal CAW (TD-CAW) and include it as a benchmark.

The RARCH models introduced by Noureldin et al. (2014) are easy to estimate with as covariance targeting can be trivially implemented by setting the target as the identity matrix, hence they are suitable for relatively large dimensions while allowing for rich dynamics. We include as benchmark a diagonal rotated DCC (RDCC) model which achieves the best performance in Noureldin et al. (2014). In addition to the original RDCC which assumes the Normal distribution, we also include an extended version using the Student-t distribution (RDCC-t).

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,[13] $H_{t+1}|r_{1:t} \sim \text{Wishart}_k^{-1}(\beta n_t + k - 1, \beta n_t S_t)$, $n_t = \beta n_{t-1} + 1$ and $S_t = \frac{1}{n_t}(\beta n_{t-1} S_{t-1} + r_t r_t')$. H_{t+1} is the latent covariance matrix of r_{t+1} and its predictive distribution follows an inverse-Wishart distribution given data $r_{1:t}$. $\beta = 0.95$ and is the discounting factor reflecting information decay moving from time t to $t+1$.[14] Assuming $r_t|H_t \sim N(0, H_t)$, the predictive density of returns is $r_{t+1}|r_{1:t} \sim \text{St}_k(0, S_t, \beta n_t)$.

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 $\Sigma_{t+1}|\mathcal{F}_t \sim \text{Wishart}_k^{-1}(\beta n_t + k - 1, \beta n_t S_t)$ and $S_t = \frac{1}{n_t}(\beta n_{t-1} S_{t-1} + \Sigma_t)$. 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 is for the observed RCOV. Assuming $r_t|\Sigma_t \sim N(0, \Sigma_t)$ the predictive density of returns given past data is $r_{t+1}|\mathcal{F}_{1:t} \sim \text{St}_k(0, S_t, \beta n_t)$.

Finally, a random walk (RW) that uses last period's value for all future forecasts, an exponentially weighted moving average (EWMA) with smoothing parameter 0.99, and a simple moving average (SMA) with a window of 500 days are included.

We focus 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.

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

[14] 0.95 is typically used in other empirical work but other parameter values (including 0.99) give similar or worse results for the covariance matrix discounting model

Table 3 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 TD-CAW performs the worst for all h.

The performance of point forecasts is found in Table 4. Here the 10 factor IW-IHMM-F 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 TD-CAW, 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 TD-CAW.

Density forecast performance for daily returns is reported in Table 5. For EWMA and SMA which only produce point forecasts of RCOV, we assume a Student-t distribution with 10 degrees of freedom for the return conditional on RCOV, and use the predictive mean of RCOV as a plug-in estimate to compute pseudo predictive likelihoods. For $h = 1, 5, 60$, 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 8 against the RDCC-t model. The RDCC-t is very competitive and beats most of the benchmark models including its normal counterpart RDCC, as well as all the parametric factor models. However, set against this is a very large computational cost for the GARCH model which we discuss later. The SMA model with Student-t assumption achieves the best long term results for $h = 20$ and $h = 60$.

To consider the value of these models for portfolio choice, Table 6 reports the realized variance of the global minimum variance portfolio (GMVP). The GMVP solves the following problem,

$$\min \omega_{t+h|t}^{'} \Sigma_{t+h|t} \omega_{t+h|t}, \quad \text{s.t. } \omega_{t+h|t} \iota = 1, \tag{60}$$

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

$$\hat{\omega}_{t+h|t} = \frac{\Sigma_{t+h|t}^{-1}\iota}{\iota'\Sigma_{t+h|t}^{-1}\iota}. \tag{61}$$

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. This is consistent with the best point forecasts of Σ_{t+h} from Table 4. 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 TD-CAW. Hautsch & Voigt (2017) point out that transactions costs, which we do not consider and shrinkage, can affect these results and model rankings.

Full sample estimates of ℓ_2, ℓ_3 and the number of alive clusters in the mixture are in Table 7. 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 8. 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 RDCC-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.

5 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.

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 Forecasting* **36**, 181–206.

Chan, J., Leon-Gonzalez, R. & Strachan, R. W. (2018), 'Invariant inference and efficient computation in the static factor model', *Journal of the American Statistical Association* **113**(522), 819–828.

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.

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.

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.

Hautsch, N. & Voigt, S. (2017), Large-scale portfolio allocation under transaction costs and model uncertainty. https://arxiv.org/abs/1709.06296.

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.

Kastner, G. (2018), Sparse Bayesian time-varying covariance estimation in many dimensions. forthcoming Journal of Econometrics.

Kastner, G., Fruhwirth-Schnatter, S. & Lopes, H. F. (2017), 'Efficient bayesian inference for multivariate factor stochastic volatility models', *Journal of Computational and Graphical Statistics* **26**(4), 905–917.

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.

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

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.

Shirota, S., Omori, Y., Lopes, H. F. & Piao, H. (2017), 'Cholesky realized stochastic volatility model', *Econometrics and Statistics* **3**, 34 – 59.

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	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
GCAW		−39052	−39305	−39460	−39602	−40982
CAW		−39121	−39659	−40445	−41805	−43505
IW		−32220	−34437	−36233	−39232	−46941
IW-DPM		−27451	**−28012**	−28572	−29680	−32418
IW-IHMM		**−27039**	−28023	**−28550**	**−29319**	**−31061**
IW-F	1	−46251	−46376	−46502	−46768	−47529
IW-DPM-F	1	−30652	−31039	−31359	−31859	−33403
IW-IHMM-F	1	−29395	−30141	−30668	−31434	−33418
IW-F	3	−40447	−40909	−41317	−41982	−43654
IW-DPM-F	3	−30170	−30601	−30880	−31530	−33097
IW-IHMM-F	3	−28987	−29794	−30262	−30921	−32651
IW-F	5	−36694	−37543	−38252	−39504	−42588
IW-DPM-F	5	−29713	−30177	−30482	−31206	−32989
IW-IHMM-F	5	−28624	−29589	−30011	−30726	−32371
IW-F	7	−34385	−35686	−36732	−38367	−42880
IW-DPM-F	7	−29105	−29706	−30040	−30652	−32027
IW-IHMM-F	7	−28455	−29414	−29900	−30511	−31664
IW-F	9	−32276	−34042	−35446	−37816	−43786
IW-DPM-F	9	−28104	−28918	−29411	−30253	−32261
IW-IHMM-F	9	−27604	−28664	−29252	−29929	−31541

The table reports the cumulative log-predictive likelihoods for RCOV at different forecast horizon h. Bold entries denote the maximum value in each column. Forecasts are for the last 500 observations (2008/1/9-2009/12/31).

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

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
GCAW		82.03	89.10	92.43	99.87	108.51
CAW		82.43	89.51	92.26	99.88	109.50
IW		85.38	94.56	99.09	108.61	122.65
IW-DPM		85.95	89.94	93.76	96.56	**102.27**
IW-IHMM		78.60	86.17	91.11	96.37	102.36
IW-F	1	89.75	94.74	97.15	101.51	106.26
IW-DPM	1	91.13	99.50	101.53	103.71	106.89
IW-IHMM-F	1	85.83	91.28	94.54	98.66	104.83
IW-F	3	87.61	94.64	99.14	108.39	118.56
IW-DPM-F	3	84.39	89.34	93.48	98.53	105.55
IW-IHMM-F	3	80.01	87.65	91.83	97.76	105.20
IW-F	5	86.25	93.34	97.32	107.03	123.06
IW-DPM-F	5	84.83	89.54	92.92	99.61	107.75
IW-IHMM-F	5	78.28	87.40	90.92	98.40	105.82
IW-F	7	85.87	92.96	96.88	106.66	123.88
IW-DPM-F	7	85.42	90.28	93.69	100.92	109.36
IW-IHMM-F	7	78.13	86.71	**90.32**	96.94	104.31
IW-F	9	85.28	92.57	96.43	106.22	124.65
IW-DPM-F	9	84.31	88.69	91.76	97.01	102.90
IW-IHMM-F	9	**78.04**	**85.98**	90.57	**95.96**	102.44

The table reports the root mean squared forecast error for predictive mean of RCOV at different forecast horizon h. Bold entries denote the minimum value in each column. Forecasts are for the last 500 observations (2008/1/9-2009/12/31).

Table 3: Cumulative log-predictive likelihoods for RCOV:60 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW-F	1	1017209	1016664	1015373	1011318	994026
IW-IHMM-F	1	1396307	1393526	1391401	1388192	1379376
IW-F	3	1056464	1054908	1051632	1043723	1020536
IW-IHMM-F	3	1398145	**1395433**	**1393253**	1389954	1381750
IW-F	5	1094748	1093189	1091257	1085498	1062521
IW-IHMM-F	5	1397132	1393931	1391398	1387744	1378840
IW-F	7	1114075	1112050	1109570	1103094	1077736
IW-IHMM-F	7	1398474	1394977	1392512	1389660	1381604
IW-F	10	1132220	1129380	1126290	1118661	1088156
IW-IHMM-F	10	**1398625**	1395204	1392819	**1389955**	**1382475**
TD-CAW		787651	766960	740936	708430	644569
RCOV discount		1184730	1155150	1128348	1077493	851783

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 4: Root mean squared forecast error for predictive mean of RCOV:60 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW- F	1	51.08	52.57	53.35	54.07	53.38
IW-IHMM-F	1	45.52	45.47	45.35	45.46	46.29
IW-F	3	47.41	49.13	49.97	50.82	50.91
IW-IHMM-F	3	45.23	45.26	45.19	45.25	45.83
IW-F	5	46.33	47.98	48.69	49.40	48.93
IW-IHMM-F	5	44.94	45.17	45.14	45.24	45.66
IW-F	7	45.86	47.44	48.12	48.76	48.39
IW-IHMM-F	7	44.95	45.17	45.13	45.15	45.65
IW-F	10	45.43	46.98	47.63	48.19	47.93
IW-IHMM-F	10	**44.90**	**45.14**	**45.09**	**45.15**	**45.59**
TD-CAW		45.72	48.47	51.13	56.39	75.59
RDCC		62.44	65.64	68.96	74.30	89.42
RDCC-t		60.63	63.75	66.95	72.41	87.58
RCOV discount		46.07	47.24	48.15	49.34	52.83
EWMA		46.09	46.39	46.57	46.77	47.55
SMA		48.39	48.67	48.99	49.77	53.45
RW		62.00	65.35	66.72	66.95	64.67

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

Table 5: Cumulative log-predictive likelihoods for return:60 assets

Model	Factors	$h = 1$	$h = 5$	$h = 10$	$h = 20$	$h = 60$
IW-F	1	−35770	−35774	−35795	−35833	−35877
IW-IHMM-F	1	−33784	−33853	−33907	−33980	−34098
IW-F	3	−35644	−35640	−35675	−35734	−35819
IW-IHMM-F	3	−33758	−33832	−33907	−33982	**−34092**
IW-F	5	−35500	−35501	−35523	−35578	−35652
IW-IHMM-F	5	−33752	−33833	−33919	−34014	−34144
IW-F	7	−35373	−35378	−35403	−35462	−35555
IW-IHMM-F	7	−33742	**−33817**	−33907	−34007	−34158
IW-F	10	−35266	−35280	−35307	−35368	−35479
IW-IHMM-F	10	**−33733**	−33824	−33921	−34033	−34175
TD-CAW		−49649	−52911	−55384	−58026	−61337
RDCC		−34823	−34944	−35136	−35454	−36097
RDCC-t		−33741	−33832	**−33893**	−34046	−34530
RCOV discount		−34387	−34635	−34648	−34701	−34861
COV discount		−49411	−49762	−50359	−51631	−59408
EWMA-t		−34924	−35037	−35098	−35180	−35295
SMA-t		−33854	−33872	−33887	**−33934**	**−34092**

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. Since EWMA and SMA only produce point forecasts of RCOV, we assume a Student-t distribution with 10 degrees of freedom for the return conditional on RCOV, and use the predictive mean of RCOV as a plug-in estimate to compute pseudo predictive likelihoods, hence EWMA-t and SMA-t.

Table 6: 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**
TD-CAW		0.3322	0.3320	0.3342	0.3424	0.3652
RDCC		0.3596	0.3644	0.3646	0.3573	0.4147
RDCC-t		0.3524	0.3541	0.3532	0.3523	0.4146
RCOV discount		0.3475	0.3586	0.3635	0.3746	0.3828
EWMA		0.3373	0.3423	0.3498	0.3562	0.3909
SMA		0.3458	0.3510	0.3548	0.3656	0.3902
RW		0.3787	0.4513	0.4584	0.4669	0.4561

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 7: 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 8: 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
RDCC		22h	RDCC-t		24h
TD-CAW		22m34s			

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 using OpenMP, (https://www.openmp.org/).