Noname manuscript No.
(will be inserted by the editor)

Sparse Kalman Filtering Approaches to Realized Covariance Estimation from High Frequency Financial Data

Michael Ho · **Jack Xin**

Received: date / Accepted: date

Abstract Estimation of the covariance matrix of asset returns from high frequency data is complicated by asynchronous returns, market microstructure noise and jumps. One technique for addressing both asynchronous returns and market microstructure is the Kalman-Expectation-Maximization (KEM) algorithm. However the KEM approach assumes log-normal prices and does not address jumps in the return process which can corrupt estimation of the covariance matrix.

In this paper we extend the KEM algorithm to price models that include jumps. We propose a sparse Kalman filtering approach to this problem. In particular we develop a Kalman Expectation Conditional Maximization (KECM) algorithm to determine the unknown covariance as well as detecting the jumps. In order to promote a sparse estimate of the jumps ,we consider both Laplace and the spike and slab jump priors. Numerical results using simulated data show that each of these approaches provide for improved covariance estimation relative to the KEM method in a variety of settings where jumps occur.

Keywords Spike and Slab · ECM · Kalman Filtering · ℓ_1 regularization

Mathematics Subject Classification (2000) 90C26 · 62P05

The work was partially supported by NSF grants DMS-1522383 and IIS-1632935

Michael Ho
Department of Mathematics, University of California at Irvine, Irvine, CA 92697, USA.
E-mail: mtho1@uci.edu

Jack Xin
Department of Mathematics, University of California at Irvine, Irvine, CA 92697, USA.
Tel.: +1-949-824-5309
Fax: +1-949-824-7993
E-mail: jxin@math.uci.edu

1 Introduction

The covariance matrix of asset returns is an integral element of many financial optimization problems such as portfolio design. For example, in minimum variance portfolio optimization the criterion for selecting the portfolio weights ,w, can be written as

$$\begin{aligned} &\min_{w} w^T \Gamma w \\ \text{s.t.} &\sum_i w_i = 1 \end{aligned}$$

where Γ is the covariance matrix of the asset returns. Since the covariance matrix is usually unknown, the above criterion cannot be implemented exactly. Instead an estimate of the covariance matrix, $\hat{\Gamma}$, is obtained and substituted into the portfolio optimization criterion.

A simple and intuitive approach to estimating the covariance matrix is to form a sample average of the covariance matrix using from past return data. However when a finite number of samples are used, covariance estimation errors will be present. These errors can result in portfolio performance that departs significantly from the optimal performance under known statistics [16, 6, 23]. Thus for portfolio optimization to be effective an accurate estimate of the covariance matrix is paramount.

Appealing to the law of large numbers covariance estimation errors can be reduced by using more data in the sample average estimate. One approach to obtain more data is to simply increase the time window size when forming the sample covariance (e.g. use 1 year of data vs 3 months of data). In order for this approach to be effective the additional data used in covariance estimation should be nearly identically distributed to future data. If the data statistics are non-stationary then increasing the window's size to obtain more data may not improve portfolio performance as the additional data used in the covariance estimation may not be relevant to future returns.

Another approach to obtaining more data is to sample at a higher frequency [3] (e.g. 1 second update rate vs 1 day update rate) and maintain the sampling window size. This approach is less vulnerable to non-stationary statistics but presents additional challenges unique to high frequency data. For example, high frequency data is subject to market microstructure noise [12] such as bid-ask bounce which can corrupt volatility and covariance estimates. At higher frequencies the variance of the market microstructure noise can mask the true volatility of the asset returns if it is not accounted for [4, 3]. Asynchronous trading of assets observed at higher frequencies [26] further complicates covariance estimation as the standard sample average estimate assumes return data is available at each time instance.

Many approaches have been proposed for estimating covariance matrices from high frequency data in the presence of asynchronous trading and microstructure noise. For example, the refresh-time approach proposed in [5] addresses asynchronous trading by attempting to synchronize the return data by waiting for all assets to trade at least one time prior to forming a asset price vector used in covariance estimation. One disadvantage of this approach is that much of the data is ignored while waiting for all assets to trade. The pairwise refresh approach [18] uses more data by refreshing the covariance matrix element by element. This allows for more data to be used but the resulting sample covariance matrix is not guaranteed to be positive semi-definite without applying additional corrections such as a projection method [18]. Another approach is the previous tick method employed in [38] where a fixed sampling grid is defined and trade prices are approximated on that grid as the nearest previous trade price.

To address both micro-structure noise and asynchronous returns, quasi-maximum likelihood estimators were proposed in [2,25] that utilize pairwise refresh. A two scale realized covariance (TSCV) approach was developed in [38] where covariance estimates are obtained using both low frequency and high frequency sampling. An approach based on Kalman filtering and the Expectation Maximization (EM) algorithm [15], models the true unobserved log-price process and observed prices as a discrete linear normal dynamical system. Here the unobserved synchronous true price is treated as latent data and the EM algorithm is used to determine a maximum-likelihood estimate of the covariance. A Bayesian version of the Kalman-EM approach where the posterior distribution of the covariance is approximated via an augmented Gibbs sampler is proposed in [31]. This technique generates an estimate of the posterior distribution of the covariance which can then be used to obtain to a point estimate.

Each of the above techniques addressing micro-structure noise and asynchronous returns utilize a log-normal price model. However, empirical return data often exhibits heavy tails that are better explained by a jump diffusion or stochastic volatility models. Under these conditions the approaches which assume log-normal returns will yield sub-optimal results. Techniques for addressing jumps have been proposed in the literature. In [19] the authors propose wavelet techniques for detecting jumps with an application to volatility estimation. The jumps estimated using this approach are then removed from the observed data prior to volatility estimation. In [10] a jump detector is employed to selectively remove data that contain jumps from the covariance estimation samples prior to TSCV. Another technique proposed in [9] is also robust to jumps but does not address market microstructure noise.

In this paper we extend the Kalman-EM approach in [15] to discretized jump diffusion models by introducing two Kalman-ECM (KECM) approaches. In our first KECM approach we model the jumps as Laplace distributed random variables. Although the Laplace prior may seem to be an unnatural model for a jump process, we will see that the prior promotes a sparse posterior mode for the jumps by inducing an ℓ_1 norm penalty on the jumps into the complete log-likelihood function. Conditioned on other variables determining the posterior mode for the jumps is a convex ℓ_1 norm penalized quadratic program which can be solved with a variety of fast techniques [22,11,7]. In our second KECM approach we consider a more natural, but non-convex, spike and slab model for the jump process. The remainder of this paper is organized as follows. In section 2 we introduce the models which form the basis for our covariance estimation approaches. In section 3 we describe numerical algorithms for computing the covariance estimate with both the Laplace and spike and slab prior. A performance evaluation of our proposed approaches are presented in section 4 using simulated high frequency data. A summary and conclusion are presented in section 5.

2 High Frequency Return Modeling

Suppose that we have N assets where the true (or efficient) log price of the n^{th} asset at time t is $X_n(t)$. Let $X(t)$ denote the $N \times 1$ vector of log prices for each asset at time t and let T denote the total number of time samples. Here $X_n(t)$ can be viewed as the fundamental value of the asset in an efficient market without friction [32].

We model the dynamics of the log prices using a discrete time jump diffusion model with a drift D

$$X_i(t) = X_i(t-1) + V_i(t) + \tilde{J}_i(t)Z_i(t) + D_i. \tag{1}$$

This model is similar to the model proposed in [15] except we consider jumps as an additional component of an asset's return. In the above model the asset return from time $t-1$ to t consists of three components

- $V_i(t)$ represents a random return which occurs at every time step
- D_i is the mean return for a single time step
- $\tilde{J}_i(t)Z_i(t)$ represents a rarely occurring jump in the asset prices which may occur at any time.

In this paper we assume the following:

- $V(t)$ is multivariate normally distributed with mean 0 and covariance Γ
- $\tilde{J}_i(t)$ is normally distributed with zero mean and variance $\sigma^2_{j,i}(t)$
- $Z_i(t)$ is Bernoulli distributed, with $Pr(Z_i(t) = 0) = \zeta$
- $\tilde{J}_m(t) \perp\!\!\!\perp \tilde{J}_n(s), Z_m(t) \perp\!\!\!\perp Z_n(s) \quad, m \neq n$ and all t, s
- $\tilde{J}, Z, V$ are jointly independent.

To simplify notation we denote the jump component as

$$J(t) = \tilde{J}(t)Z(t). \tag{2}$$

In many markets trading of distinct assets does not occur simultaneously. When trades occur asynchronously, current pricing data for all assets will not be observed. For prices that are observed, market microstructure noise needs to be addressed. Here transaction costs due to order processing expenses, inventory costs and adverse selection costs [12] add noise to the true efficient price. Thus the true efficient price is not directly observed.

Both asynchronous returns and microstructure noise can be captured in the following observation model

$$Y(t) = \tilde{I}(t)X(t) + W(t) \tag{3}$$

where

- $\tilde{I}(t)$ is a"partial" identity matrix where the rows corresponding to missing asset prices at time t are removed
- $W(t)$ is normal distributed market microstructure noise with zero mean and covariance $\Sigma_o(t) = \tilde{I}(t)\Sigma'_o\tilde{I}(t)^T$.

Here Σ'_o is a diagonal matrix $\mathrm{diag}(\sigma^2_{o,1}, \ldots, \sigma^2_{o,N})$. In this section we shall assume that $\{W(t), X(t)\}_{t=1}^T$ are jointly independent. In section 4 we will test our algorithms on simulated data where the microstructure noise and price innovation are statistically dependent.

2.1 Conditional Distributions of Observations and Log-Prices

Now we examine the joint probability distribution of $X(1:T)$,$Y(1:T)$,$J(2:T)$. Here the notation $X(m:n)$ refers to the set $\{X(m), X(m+1), \ldots, X(n)\}$. We consider the case of when the parameters $D, \Gamma, \sigma^2_{o,i}, \zeta$ and $\sigma^2_{j,i}$ are random variables with known prior distributions. Details on our assumed priors are given in section 2.2.

To determine the probability distribution we first note that the following conditional independence properties hold

$$\begin{aligned} Y(t) \perp\!\!\!\perp J(s)|X(t) \;\; \forall s \\ Y(t) \perp\!\!\!\perp X(s)|X(t) \;\; \forall s \neq t \\ X(t) \perp\!\!\!\perp X(s)|(X(t-1), J(t)) \;\; \forall s < t-1 \\ X(t) \perp\!\!\!\perp J(s)|(X(t-1), J(t)) \;\; \forall s \neq t. \end{aligned}$$

From the conditional independence we have that the probability distribution conditioned on the parameter values may be fully characterized as follows

$$\begin{aligned} p(y(t)|x(1:T), \Sigma^2_o(t)) &\sim \mathcal{N}(\tilde{I}(t)x(t), \Sigma_o(t)) \\ p(x(t+1)|x(1:t), j(2:t+1), d, \Gamma) &\sim \mathcal{N}(x(t) + j(t+1) + d, \Gamma) \\ p(x(1)) &\sim \mathcal{N}(\mu, K) \\ p(j(t)|\zeta, \sigma^2_j(t)) &\sim \prod_{i=1}^{N} f(j_i(t)). \end{aligned}$$

Here f is the spike and slab prior

$$f(j_i(t)) = \zeta\delta_0(j_i(t)) + \frac{1-\zeta}{\sqrt{2\pi}\sigma_{j,i}(t)} \exp\left(-\frac{j_i(t)^2}{2\sigma^2_{j,i}(t)}\right) \tag{4}$$

with δ_0 being a point mass distribution at 0. The initial time parameters, μ and K can be chosen based on prior stock return data and will be treated as known values. Note that since joint estimation of $\sigma_{j,i}$, ζ ,and j_i is ill-posed we impose regularization in the form of priors on both $\sigma_{j,i}$, and ζ. Details are given in Section 2.2.

2.2 Prior Distribution of Parameters

To allow for more flexible modeling we shall impose prior distributions on the parameters $D, \Gamma, \sigma^2_{o,i}$ as well as the jump parameters ζ and $\sigma^2_{j,i}$. Here we take a commonly used approach of using conjugate prior distributions which facilitate calculation of conditional maximum a posteriori (MAP) parameter estimates. These priors will play an essential part in the proofs of convergence for the ECM algorithm presented in Section 3.

The drift parameter D is modeled as normally distributed with mean $\bar{D}$ and covariance $\sigma^2_D I$

$$D \sim \mathcal{N}(\bar{D}, \sigma^2_D I),$$

which is conjugate to the multivariate normal distribution given above. For the covariance matrix prior we use an inverse Wishart prior (which is also conjugate to the multivariate normal) with $\eta > N-1$ degrees of freedom and positive definite scale matrix W_o

$$\Gamma \sim \mathcal{W}^{-1}(W_o, \eta).$$

In the observation noise variance,$\sigma^2_{o,i}$, we impose a inverse gamma distribution with shape parameter $\alpha_o > 0$ and scale $\beta_o > 0$

$$\sigma^2_{o,i} \sim IG(\alpha_o, \beta_o).$$

Finally for the jump parameters ζ and σ^2_j we use the beta distribution and inverse gamma distribution as priors

$$\zeta \sim \text{Beta}(\alpha_\zeta, \beta_\zeta)$$

$$\sigma^2_{j,i}(t) \sim IG(\alpha_j, \beta_j).$$

We assume that ζ and $\sigma^2_{j,i}(t)$ are independent and that the parameters in each of the prior distributions is known.

2.3 Laplace Prior Approximation

The jump model is equivalent to a switching state space model. Inference in switching state space models becomes intractable as the number of states increase [21]. In this section we approximate the distribution of J using a Laplace distribution. We denote the Laplace distribution for J as $g(j)$

$$p(j) \approx g(j|\lambda) \doteq \prod_{i,t} \frac{\lambda_i(t)}{2} \exp\left(-\lambda_i(t)|j_i(t)|\right)$$

where $\lambda_i(t) > 0$.

There are two advantages to taking this approximation. First the log-likelihood of a Laplace distribution is concave in its parameter. This aids in conditional MAP estimation of J. Secondly, the Laplace distribution is desired in that it promotes sparse MAP estimates of J [33,29,1] making it a good approximation to infrequent jumps. To make the model more robust we will not assume that each $\lambda_i(t)$ is known. Instead we will estimate $\lambda_i(t)$ from the data. Since the problem of estimating both $J_i(t)$ and $\lambda_i(t)$ is ill-posed we regularize it by introducing a prior distribution on each $\lambda_i(t)$ which we denote as $q(\lambda)$.

We wish to design the prior distribution q such that it induces a similar level of sparseness that is induced by the spike and slab prior f. To develop a criterion for designing q we first define a notion of similarity between $g(j|\lambda)$ and $f(j|, \zeta, \sigma_j^2)$.

Definition 1 *Let V be a zero-mean normal random variable with variance σ_v^2 and let $J_1 \sim Laplace(\lambda')$ and $J_2 \sim SpikeSlab(\zeta', \sigma_j^{2'})$ which are independent of V. Define*

$$Y_1 = J_1 + V$$
$$Y_2 = J_2 + V.$$

*Then $Laplace(\lambda')$ is σ_v^2-**equivalent** to $SpikeSlab(\zeta', \sigma_j^{2'})$ (denoted $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$) if*

$$\mathbb{E}_{p(y_2|J_2=0)} Pr(J_2 = 0|Y_2) = \mathbb{E}_{p(y_1|J_1=0)} Pr(\bar{J}_1 = 0) \qquad (5)$$

where $\bar{J}_1$ is the mode of $p(j_1|Y_1)$.

To interpret the above definition assume that a jump has not occurred. Then $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$ if the probability of falsely declaring a jump under the Laplace model (with MAP criterion) equals the average posterior probability of a jump under the spike and slab prior with parameters ζ' and $\sigma_j^{2'}$. Here σ_v^2 can be interpreted as the squared volatility of the diffusion component of the asset returns. Note that for each triplet $(\sigma_v^2, \zeta', \sigma_j^{2'})$ there is a unique λ' such that $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$.

Since $(\sigma_v^2, \zeta', \sigma_j^{2'})$ are random and unobserved we cannot directly select a λ' such that $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$. However the distribution of $(\sigma_v^2, \zeta_o', \sigma_j^{2'})$ induces a distribution on λ through the mapping $\sim_{\sigma_v^2}$. The resulting distribution can then be used as a prior $q(\lambda)$. An example on how to construct an inverse gamma approximation of distribution for λ is shown in Appendix C.

3 KECM Approach to estimation of Γ

Maximum a posteriori (MAP) estimation of Γ with Kalman ECM (KECM) techniques is investigated in this section. The first ECM approach is an approximate technique where the prior distribution on the jumps is modeled as a Laplace distribution. The advantage of this approximation is that the conditional maximization steps in the ECM approach result in global (conditional) optimal solutions can be obtained. The disadvantage is that we are approximating the true spike and slab jump model. The second approach uses the spike and slab model for jumps, which is a true representation of the model presented in Section 2. However we will see that using the spike and slab jump model results in a non-convex optimization problem in the conditional M-step for J.

3.1 KECM algorithm for Laplace Distribution

First we consider a KECM approach to estimating Γ when J_i is approximated by a Laplace distributed random variable. We define

$$\Theta = [\Theta_1, \Theta_2, \Theta_3, \Theta_4, \Theta_5]$$

where

$$\begin{aligned}
&\Theta_1 = D \\
&\Theta_2 = \Gamma \\
&\Theta_3 = \sigma_{o,i}^2, 1 \leq i \leq N \\
&\Theta_4 = J(2:T) \\
&\Theta_5 = \{\lambda_i(t)^{-1}\}_{1\leq i\leq N, 2\leq t\leq T}
\end{aligned}$$

as our vector of unknown parameters and $X(1:T)$ as the latent variables.

The KECM approach is an iterative algorithm that can be applied to the following problem

$$\Theta^* = \arg\max_{\theta} L(\theta)$$

where $L(\theta)$ is the log posterior of Θ. In the KECM algorithm we iterate over E-steps and conditional M-steps to arrive at an estimate of Θ.

The E-step in the KECM algorithm involves computing the expected value of

$$\log p(X(1:T), y(1:T)|\theta)p(\theta)$$

with respect to $p(x(1:T)|y, \Theta^{(k)})$

$$\mathcal{G}(\theta, \Theta^{(k)}) = \mathbb{E}_{p(x|y,\Theta^{(k)})} \log p(X(1:T), y(1:T)|\theta) + \log(p(\theta))$$

where $\Theta^{(k)}$ is an estimate of Θ at the k^{th} iteration and where $p(\theta)$ is the prior distribution of parameters

$$p(\theta) = p(\theta_1)p(\theta_2)p(\theta_3)g(\theta_4, |\lambda)q_{inv}(\lambda^{-1}).$$

Here the complete log-likelihood is

$$\log p(x, y|\theta) = -0.5 \sum_{t=1}^{T} \log(|\Sigma_o(t)|) - \frac{1}{2} \sum_{t=1}^{T} ||y(t) - \tilde{I}(t)x(t)||^2_{\text{diag}(\Sigma_o(t)^{-1}),\ell_2}$$
$$- \frac{T-1}{2} \log(|\Gamma|) - \frac{1}{2} \sum_{t=2}^{T} r(t)^T \Gamma^{-1} r(t) + const$$

where

$$r(t) = x(t) - x(t-1) - d - j(t).$$

and where

$$||q||^2_{\beta,\ell_2} = \sum_i \beta_i q_i^2.$$

It is well known that the function $\mathcal{G}(\theta, \Theta^{(k)})$ serves as a lower bound to $\log p(\theta, y)$ and that $\log p(\Theta^{(k)}, y) = \mathcal{G}(\Theta^{(k)}, \Theta^{(k)})$ [17].

The EM approach prescribes a joint maximization of $\mathcal{G}(\theta, \Theta^{(k)})$ with respect to θ. This is difficult due to the coupling of variables and the non-concavity of the problem. Conditional maximization of each parameter in turn is more tractable. Thus we apply conditional maximization as in the ECM [30] algorithm. The conditional M-steps involves a coordinate-wise maximization of $\mathcal{G}$. Here the conditional M-steps are

$$\Theta_i^{(k+1)} = \arg\max_{\theta_i} \mathcal{G}\left(\left[\Theta_1^{(k+1)}, \ldots \Theta_{i-1}^{(k+1)}, \theta_i, \Theta_{i+1}^{(k)}, \ldots, \Theta_5^{(k)}\right], \Theta^{(k)}\right) \quad (6)$$

Each of these problems can be readily solved as we will show later.

3.1.1 E-step of KECM

The posterior $p(x|y, \Theta^{(k)})$ needed to perform the E-step is normal and can be computed using a Kalman smoother [35]. By normality and the Markov property the posterior is completely defined by the following posterior moments for $m = T$

$$\bar{X}(t|m) \doteq \mathbb{E}(X(t)|y(1:m))$$

$$P(t|m) \doteq cov(X(t), X(t)|y(1:m))$$

$$P(t, t-1|m) \doteq cov(X(t), X(t-1)|y(1:m))$$

where $cov(:,:)$ refers to the covariance function. Equations for these quantities are derived in [34] and given in Appendix A. The expected value of log-posterior distribution with respect to the posterior of $X(1:T)$ can be shown to be

$$\begin{aligned}\mathcal{G}(\theta,\Theta^{(k)}) &= \mathbb{E}_{p(x|y,\Theta^{(k)})}\log p(X(1:T),y(1:T)|\theta)+\log(p(\theta))\\ &= -\frac{T-1}{2}\log(|\Gamma|)-\frac{1}{2}\mathrm{tr}(\Gamma^{-1}(C-B-B^T+A))\\ &\quad -\frac{1}{2}\sum_{t=1}^{T}||y(t)-\tilde{I}(t)\bar{X}(t)||^2_{\mathrm{diag}(\Sigma_o(t)^{-1}),\ell_2}+\mathrm{tr}(P(t|T)\tilde{I}(t)^T\Sigma_o(t)^{-1}\tilde{I}(t))\\ &\quad -0.5\sum_{t=1}^{T}\log(|\Sigma_o(t)|)+\log(p(\theta))+const \end{aligned}\tag{7}$$

where

$$A=\sum_{t=2}^{T}\left(P(t-1|T)+\bar{X}(t-1|T)\bar{X}(t-1|T)^T\right)$$

$$B=\sum_{t=2}^{T}\left(P(t,t-1|T)+(\bar{X}(t|T)-D^{(k)}-J^{(k)}(t))\bar{X}(t-1|T)^T\right)$$

$$C=\sum_{t=2}^{T}\left(P(t|T)+(\bar{X}(t|T)-D^{(k)}-J^{(k)}(t))(\bar{X}(t|T)-D^{(k)}-J^{(k)}(t))^T\right).$$

For notational convenience the dependence of $P(t|m)$ and $P(t,t-1|m)$ on the iteration number has been dropped.

3.1.2 Conditional M-steps of KECM

For the conditional M-step it can be shown using standard conjugate prior relationships [20] that

$$D^{(k+1)}=F\left(\frac{1}{\sigma_D^2}\bar{D}+\Gamma^{(k)^{-1}}\sum_{t=2}^{T}\bar{X}(t|T)-\bar{X}(t-1|T)-J^{(k)}(t)\right)\tag{8}$$

and

$$\Gamma^{(k+1)}=\frac{1}{T-1+\eta}\left(A+C^{(k)}-B^{(k)}-B^{(k)T}\right)+\frac{1}{T-1+\eta}W\tag{9}$$

where

$$F=\left((T-1)\Gamma^{(k)^{-1}}+\sigma_D^{-2}I\right)^{-1}$$

$$B^{(k)}=\sum_{t=2}^{T}\left(P(t,t-1|T)+(\bar{X}(t|T)-D^{(k+1)}-J(t)^{(k)})\bar{X}(t-1|T)^T\right)$$

and

$$\begin{aligned} C^{(k)} &= \sum_{t=2}^{T} P(t|T) \\ &+ \sum_{t=2}^{T} (\bar{X}(t|T) - D^{(k+1)} - J(t)^{(k)})(\bar{X}(t|T) - D^{(k+1)} - J(t)^{(k)})^T. \end{aligned}$$

The conditional M-step for the observation noise variance is

$$\sigma_{o,i}^{2,(k+1)} = \frac{2\beta_o + \sum_{t \in \mathcal{T}_i} (y(t) - \tilde{I}(t)\bar{X}(t|T))^2_{\eta(i,t)} + (P(t|T))_{i,i}}{2\alpha_o + 2 + M_i}. \tag{10}$$

Here $\mathcal{T}_i$ is the set of times where the price of asset i is observed and M_i is the total number of prices observed for asset i. The subscript $\eta(i,t)$ is the row number of $\tilde{I}(t)$ such that $\tilde{I}(t)_{\eta(i,t),i} = 1$.

For each conditional M-step $P(t,T)$, $P(t, t-1|T)$ and $\bar{X}(t|T)$ are evaluated with respect to $p(X(1:T)|Y, \Theta^{(k)})$.

To compute the conditional M-step for J we denote

$$Q(j) \doteq \mathcal{G}([\Gamma^{(k+1)}, D^{(k+1)}, \{\sigma_{o,i}^2\}_{1 \le i \le N}, j, \{\lambda_i(t)^{-1}\}^{(k)}_{1 \le i \le N, 2 \le t \le T}], \Theta^{(k)}).$$

Then up to a constant not depending on j

$$\begin{aligned} Q(j) = &-\frac{1}{2} \sum_{t=2}^{T} j(t)^T (\Gamma^{(k+1)})^{-1} j(t) \\ &+ \sum_{t=2}^{T} (\bar{X}(t|T) - D^{(j+1)} - \bar{X}(t-1|T))^T (\Gamma^{(k+1)})^{-1} j(t) \\ &- \sum_{t=2}^{T} ||j(t)||_{\lambda(t),\ell_1} + const. \end{aligned}$$

where $||j(t)||_{\lambda(t),\ell_1} = \sum_{n=1}^{N} \lambda_n(t)|j_n(t)|$.

Referring to equation (6) we see that $J^{(k+1)}(t)$ is the solution of the following ℓ_1 penalized quadratic program

$$J^{(k+1)}(t) = \arg\min_j \frac{1}{2} j^T (\Gamma^{(k+1)})^{-1} j - j^T (\Gamma^{(k+1)})^{-1} \Delta^{(k+1)} + ||j(t)||_{\lambda(t),\ell_1} \tag{11}$$

where

$$\Delta^{(k)}(t) = \bar{X}(t|T) - D^{(k)} - \bar{X}(t-1|T). \tag{12}$$

This problem can be solved with a variety of fast algorithms such as ADMM [11] and FISTA [7].

Now we determine $\{\lambda_i(t)^{-1}\}_{1\leq i\leq N, 2\leq t\leq T}$ which depends only on $q_{inv}(\lambda^{-1})$ and $p(j|\lambda)$. Using conjugate prior relationships we have $p(\lambda_i(t)^{-1}|j_i(t))$ is inverse gamma distributed with shape $\alpha_\lambda + 1$ and scale $\beta_\lambda + |j_i(t)|$. Thus the conditional MAP estimate is

$$\lambda_i(t)^{-1} = \frac{|J_i^{(k+1)}(t)| + \beta_\lambda}{\alpha_\lambda + 2}. \tag{13}$$

which implies that

$$\lambda_i(t) = \frac{\alpha_\lambda + 2}{|J_i^{(k+1)}(t)| + \beta_\lambda}. \tag{14}$$

An outline of the KECM algorithm for Laplace jump models is given below.

Algorithm 1 KECM Algorithm for estimation of Γ under Laplace Prior

```
**Initialize:** $\Theta^{(0)}, k = 0$
**while** not converged **do**
    Compute $\bar{X}(t|T), P(t|T), P(t, t-1|T)$ using Kalman smoothing equations for $\Theta^{(k)}$ using equations (24)-(29)
    Compute $D^{(k+1)}, \Gamma^{(k+1)}$, and $\sigma_{o,i}^{2,(k+1)}$ using equations (8),(9), and (10) respectively
    Compute $J^{(k+1)}$ by solving (12)
    Compute $\{\lambda_i(t)\}_{1\leq i\leq N, 2\leq t\leq T}$ by solving (14)
    $k = k + 1$
**end while**
```

Here the convergence criterion could be set in a number of ways such as reducing the ℓ_2 norm of the difference in estimates, $\Theta^{(k+1)} - \Theta^{(k)}$ below a predefined threshold. Convergence results for this algorithm are given in Appendix B.

Remark 1 Since the value of $\{\lambda_i(t)\}_{1\leq i\leq N, 2\leq t\leq T}$ changes with each iteration we see that we effectively reweight the ℓ_1 penalty in (11) after each iteration. Reweighting of the ℓ_1 norm has been proposed in several papers and has been shown to have improved performance in compressive sensing problems versus a fixed set of weights [13].

3.2 KECM approach for the Spike and Slab Jump Prior

Now we present a KECM for the spike and slab jump prior. As with the Laplace prior we treat X as a latent variable. Let us denote the unknown parameters as Φ where

$$\begin{aligned}
\Phi_1 &= D\\
\Phi_2 &= \Gamma\\
\Phi_3 &= \sigma_{o,i}^2\\
\Phi_4 &= Z(2:T), \tilde{J}(2:T)\\
\Phi_5 &= \zeta\\
\Phi_6 &= \{\sigma_{j,i}^2(t)\}_{i=1,\ldots,N,t=1,\ldots,T}\,.
\end{aligned}$$

Here we allow for distinct σ_j^2 values for each time and asset.

The E-step as well as the conditional M-steps for Φ_1, Φ_2, Φ_3 are identical to the KECM algorithm for Laplace priors. The differences for this section are in the conditional M-steps for J, ζ and σ_j^2.

First we address the conditional M-step for $J^{(k+1)}$. Here we need to solve

$$\begin{aligned}
[J^{(k+1)}(t), Z^{(k+1)}(t), \tilde{J}^{(k+1)}(t)] = \arg\min_{j,z,\tilde{j}}\ & \frac{1}{2} j^T (\Gamma^{(k+1)})^{-1} j\\
& - j^T (\Gamma^{(k+1)})^{-1} \Delta^{(k+1)} - \sum_{i=1}^{N} \log(f(\tilde{j}_i, z_i))\\
\text{s.t. } & j_i = \tilde{j}_i z_i \qquad (15)
\end{aligned}$$

where

$$\log f(\tilde{j}_i, z_i) = \log\left(\zeta 1_{z_i=0} + (1-\zeta) 1_{z_i=1}\right) + \log\left(\frac{1}{\sqrt{2\pi\sigma_{j,i}^2}} \exp\left(-\frac{\tilde{j}_i^2}{2\sigma_{j,i}^2}\right)\right). \qquad (16)$$

Here we dropped the notation for time dependence. When restricted to $j_i = \tilde{j}_i z_i$, $-\log(\tilde{j}_i, z_i)$ induces a penalty on j_i. which is a weighted sum of an ℓ_0 and squared ℓ^2 norm. A plot of this penalty is shown in Figure 1.

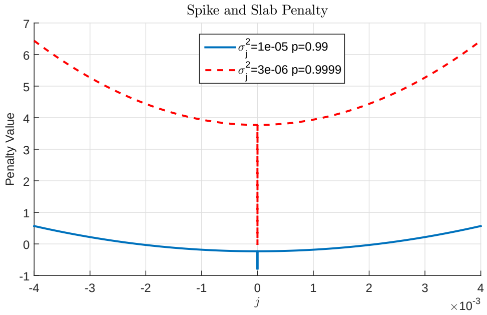

Fig. 1 Spike and slab penalty function for various parameter values. Here we see that the penalty is a weighted sum of ℓ_0 and squared ℓ_2 norms.

The term $-\log(\tilde{j}_i, z_i)$ is non-convex and complicates the conditional M-step (15). Hence we seek an approximate maximization through coordinate descent. Here we divide the problem into tractable 1-dimensional optimization problems with respect to one asset at a time. The method and equations for implementing coordinate descent are described below. For ease of notation we drop the notation denoting dependence on k.

Let us define the following conditional mean and variance

$$a(i) = \varDelta_i(t) + \varGamma_{i,-i}\varGamma_{-i,-i}^{-1}(j_{-i}(t) - \varDelta_{-i}(t)) \tag{17}$$

and

$$b^2(i) = \varGamma_{i,i} - \varGamma_{i,-i}\varGamma_{-i,-i}^{-1}\varGamma_{-i,i} \tag{18}$$

where the subscript $-i$ is to be interpreted as all indices except i. Then the following rule determines the MAP optimal value of $z_i(t)$ conditioned on $j_{-i}(t)$

$$z_{i|-i}(t) = \begin{cases} 0 \text{ if } \frac{\zeta}{1-\zeta}\mathcal{N}(0, a(i), b^2(i)) > \mathcal{N}(0, a(i), b^2(i) + \sigma_{j,i}^2(t)) \\ 1 \text{ else} \end{cases} \tag{19}$$

where $\mathcal{N}(0, a(i), b^2(i))$ is the normal PDF with mean $a(i)$ and variance $b^2(i)$ evaluated at 0. An optimal value of $\tilde{J}_{i|-i}(t)$ is then given as

$$\tilde{J}_{i|-i}(t) = \begin{cases} \frac{a}{1+b^2\sigma_{j,i}^{-2}(t)} \text{ if } z_{i|-i}(t) \neq 0 \\ 0 \text{ else} \end{cases} . \tag{20}$$

The mapping defined by equations (19) and (20) is a combination of a thresholding step followed by a shrinkage operation

$$\begin{aligned} J_{i|-i}(t) &= SpikeSlabShrink(a, b^2) \\ &\doteq \begin{cases} 0 & \text{if } \frac{\zeta\mathcal{N}(0,a(i),b^2(i))}{(1-\zeta)\mathcal{N}(0,a(i),b^2(i)+\sigma_{j,i}^2(t))} > 1 \\ \frac{a(i)}{1+b^2(i)\sigma_{j,i}^{-2}(t)} & \text{else} \end{cases} . \end{aligned} \tag{21}$$

This spike and slab shrinkage is illustrated in Figure 2. As the plots indicate the shrinkage is discontinuous and large values are shrunk more than smaller values.

Equation (21) is cycled through all $i = 1 \ldots N$. Multiple cycles may also be performed to obtain an improved estimate of J. A summary of the algorithm for the conditional M-step for J is given below in Algorithm 2.

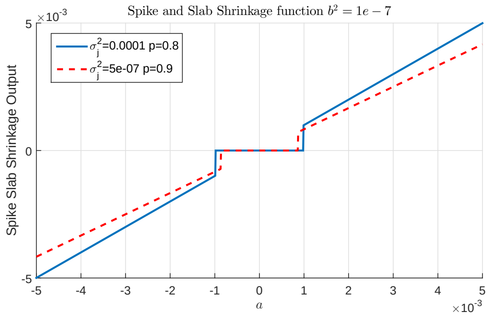

Fig. 2 Spike and slab shrinkage function for various parameter values

Algorithm 2 Coordinate Descent for Determination of $Z^{(k+1)}(t)$,$\tilde{J}^{(k+1)}(t)$, and $J^{(k+1)}(t)$

```
Initialize: Set $J^{(k+1)}(t) = J^{(k)}(t)$, $it$=0, $L > 0$
while $it \leq L$ do
  $it = it + 1$
  $i = 0$
  while $i < N$ do
    $i = i + 1$
    Compute $Z_i^{(k+1)}(t)$ using equations (17), (18), and (19)
    Compute $\tilde{J}_i^{(k+1)}(t)$ using equations (17), (18), and (20)
    Set $J_i^{(k+1)}(t) = Z_i^{(k+1)}(t)\tilde{J}_i^{(k+1)}(t)$
  end while
end while
return $J^{(k+1)}(t)$
```

Although this method is not guaranteed to solve (15) it will not increase the value of the objective function compared with $J^k(t)$.

Once $J^{(k+1)}$ is obtained, values for $\zeta^{(k+1)}$ and $\sigma_j^{2,(k+1)}$ are easily computed through conjugate prior relationships. First let N_Z be number of zero values in $J(2:T)^{(k+1)}$. Then by conjugate prior relationships the conditional M-steps for ζ and σ_j^2 are

$$\zeta^{(k+1)} = \frac{\alpha_\zeta + N_Z}{N(T-1) + \beta_\zeta + \alpha_\zeta} \tag{22}$$

and

$$\sigma_{j,i}^{2,(k+1)}(t) = \frac{\beta_j + 0.5(J_i(t))^2}{\alpha_j + 1 + 0.5(Z_i(t))}. \tag{23}$$

The KECM algorithm for spike and slab models is summarized in Algorithm 3.

Algorithm 3 KECM Algorithm for estimation of Γ under Spike and Slab Prior

```
Initialize: $\Phi^{(0)}, k = 0$
while not converged do
    Compute $\bar{X}(t|T), P(t|T), P(t, t-1|T)$ using Kalman smoothing equations for $\Theta^{(k)}$ using equations (24)-(29)
    Compute $D^{(k+1)}, \Gamma^{(k+1)}$, and $\sigma_{o,i}^{2,(k+1)}$ using equations (8), (9), and (10) respectively
    For all $t$, compute $\tilde{J}^{(k+1)}(t)$, $Z^{(k+1)}(t)$ using Algorithm 2
    Set $J_i^{(k+1)}(t) = Z_i^{(k+1)}(t)\tilde{J}_i^{(k+1)}(t)$
    Compute $\zeta^{(k+1)}$ using equation (22)
    Compute $\sigma_{j,i}^{2,(k+1)}(t)$ using equation (23) for all $i, t$
    $k = k + 1$
end while
```

Similar to Algorithm 1 the convergence criterion could be set in a number of ways such as reducing the ℓ_2 norm of the difference in estimates, $\Theta^{(k+1)} - \Theta^{(k)}$ below a predefined threshold.

Note that although $J(t)$ is only approximately maximized in each conditional M-step this is still an ECM algorithm. To see this we can simply redefine Φ as

$$\left[D, \Gamma, \sigma_o^2, J_1(2), \dots, J_N(2), \dots, J_1(T), \dots, J_N(T), \zeta, \sigma_j^2\right].$$

Then the above algorithm is an ECM algorithm for the redefined parameter vector. The convergence of Algorithm 3 is similar to the proof of the convergence of Algorithm 1 in Appendix B.

Remark 2 A comparison of the spike and slab shrinkage function with the shrinkage function of the $b^2 - \textit{equivalent}$ Laplace prior is shown in Figure 3. The Laplace shrinkage function (with parameter λ) is defined as

$$\textit{LaplaceShrink}(a, b^2) \doteq \begin{cases} a - \lambda b^2 & \text{if } a > \lambda b^2 \\ a + \lambda b^2 & \text{if } a < -\lambda b^2 \\ 0 & \text{else} \end{cases}.$$

The graphs illustrate advantages and disadvantages of the Laplace prior. One notable disadvantage is that for large σ_j^2 the Laplace prior has a large bias relative to spike and slab priors. However for small σ_j^2 and large values of a we see that the Laplace prior is less biased than the spike and slab. This can be attributed to the quadratic penalty induced by the spike and slab prior which penalizes large jumps more heavily than the Laplace prior.

Remark 3 The use of Laplace priors and ℓ_1 penalties has been applied in context of robust Kalman filtering and smoothing in [29, 1]. Here the authors considered the problem of non-gaussian heavy tailed observation noise rather than process noise.

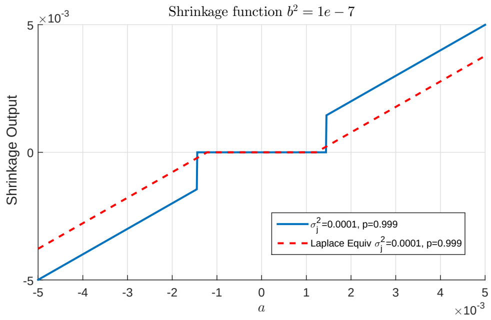

Fig. 3 Shrinkage Functions of the spike and slab and the corresponding $b^2 - \mathit{equivalent}$ Laplace prior

4 Numerical Results

In this section we evaluate the performance of the following algorithms

1. KEM [15]
2. KECM Laplace(section 3)
3. KECM Spike and Slab (section 3)
4. Pairwise refresh with TSCV [18,38]
5. Pairwise refresh with TSCV and jump correction [10]

for determining a covariance matrix from high frequency data. The KEM and pairwise refresh algorithms are included in the study to serve as benchmarks for scenarios where jumps are not present, small, or infrequent. The pairwise refresh with jump correction algorithm was included as a benchmark for cases where jumps are present. The performance of each algorithm is evaluated using a Monte Carlo approach with simulated high frequency return data.

4.1 Performance Assessment Methodology

We track two performance measures for the covariance estimate, $\hat{\Gamma}$, in this study. For the first performance measure we compute the minimum variance portfolio

$$\begin{aligned}\tilde{w} &= \arg\min_{w} w^T \hat{\Gamma} w \\ \text{s.t.} &\sum_i w_i = 1.\end{aligned}$$

The variance of this portfolio's return is then computed as a figure of merit. The variance of the portfolio return is given below

$$\tilde{w}^T \Gamma \tilde{w}.$$

For the second performance measure we compute the relative Frobenius norm of the error between the true and estimated covariance

$$\frac{\sqrt{\sum_{i,j} |\Gamma_{i,j} - \hat{\Gamma}_{i,j}|^2}}{\sqrt{\sum_{i,j} |\Gamma_{i,j}|^2}}.$$

4.2 Algorithm Initialization and other considerations

In each study we initialize the algorithms in the same way. The hyper-parameters for the prior distribution are listed in Table 1. For the KEM and KECM algorithms the initial covariance estimate is computed using the time refresh method in [5]. The initialization of drift and jump estimate of each algorithm is set to zero.

In the KECM algorithms we employ one additional initialization step to avoid being trapped in an over-smoothed local solution. This step involves using a forward Kalman filter rather than a smoother to approximate the posterior distribution of $X(t)$ in the first 10 iteration of the KECM algorithms. After 10 iterations we revert to the approaches described in Section 3 which use the Kalman smoother.

The stability of the covariance estimate forms the basis for a stopping criterion in the KECM algorithms. The KECM algorithms are terminated at iteration n when the relative difference between the current and previous covariance estimate is less than 0.001

$$\frac{\sqrt{\sum_{i,j} |\hat{\Gamma}^{(n)}_{i,j} - \hat{\Gamma}^{(n-1)}_{i,j}|^2}}{\sqrt{\sum_{i,j} |\hat{\Gamma}^{(n-1)}_{i,j}|^2}} < 0.001.$$

Since jumps cannot be predicted an ambiguity occurs if there is no observation of the price at the time the jump occurs. Thus to prevent ambiguity we assume jumps in the i^{th} asset price can only occur if an observation of the i^{th} price is made. We believe that this is a mild assumption given that in many markets jumps in the efficient price will be traded upon almost immediately. This assumption is built into the KECM approach by setting $\lambda = \infty$ and $\zeta = 1$ when an observation does not occur.

4.3 Simulated Data Jump Model

For the data study we simulated 30 minutes of data from 20 assets according to equations (1) and (3) at 1 second intervals. Here 50 data sets were generated to test our algorithms. Taking motivation from factor models for U.S. stock returns we set our covariance Γ according to the following 5 factor model

$$\Gamma = \sum_{i=1}^{5} \beta_{v_i} v_i v_i^T + \epsilon I.$$

Table 1 Parameters used in KEM, and KECM algorithms

	Value	Comment
α_ζ	10×0.995	
β_ζ	$10 - \alpha_\zeta$	prior mean of ζ is 0.995
α_j	10	
β_j	$0.01^2(\alpha_j + 1)$	prior mode of σ_j^2 is 1e-4
α_o	5	
β_o	$(\alpha_o + 1) \times 0.0001^2$	prior mode of σ_o^2 is 1e-8
η	$N + 5$	
W_o	$\frac{0.02^2(\eta+N+1)}{23400} I$	Corresponds to 0.02% daily volatility
α_λ	5.6	Obtained using method in Section 2.3
β_λ	5e-04	Obtained using method in Section 2.3

Here we compute a new covariance for each Monte Carlo data set. We draw v_1 from a multivariate normal distribution with mean $\frac{1}{\sqrt{2}}$ and covariance $0.5I$. For $i > 1$, we draw v_i from a multivariate normal distribution with zero mean and covariance I. The factor variance β_{v_i} is modeled as gamma distributed with shape 2 and mean $\frac{0.7*0.02^2}{23400}$ for $i = 1$ and mean $\frac{0.3/4*0.02^2}{23400}$ for $i \neq 1$. The ϵ term is defined to be $\frac{0.02^2}{23400*100}$. With these settings each simulated asset will on average have a daily return volatility of approximately 2 percent.

For the D parameter we use a random number generator for each data set. The value for D was drawn from a multi-variate normal distribution with mean 0 and covariance $\left(\frac{0.01}{23400}\right)^2 I$. The observation noise variance of each asset was set to a random number drawn from a gamma distribution with shape 2 and mean 0.0002^2. For a stock price of \$25 this corresponds to a mean noise standard deviation of about \$0.005. The jump parameters ζ and σ_j^2 were varied parametrical over several values.

The KECM algorithm require hyperparameters to be specified for the prior distributions. For these experiments we choose hyperparameters which would result in diffuse priors in order to minimize bias. For the hyperparameters of the Laplace prior in the KECM algorithm we used the technique described in Appendix C. A listing of all the hyperparameters used in the algorithms are shown in Table 1.

The probability that any given price is observed is set to be commensurate with the price innovation. This is consistent with empirical observations that trading volume can be positively correlated with volatility [24]. To model this association the probability that the m^{th} asset price will be observed at time t is simulated as

$$p_{obs,m}(t) = \frac{|X_m(t) - X_{m-1}(t) - D_m|}{|X_m(t) - X_{m-1}(t) - D_m| + \nu}$$

where

$$\nu = \frac{\sqrt{2\Gamma_{m,m}}}{\pi} \left(\frac{1}{p_{Obs}} - 1 \right).$$

Table 2 Portfolio variance for jump model

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	1.2e-10	1.3e-10	1.3e-10	1.8e-10	2.3e-10
0.9999	6.25e-06	1.5e-10	1.4e-10	1.4e-10	1.8e-10	3.3e-10
0.9999	0.0001	1.6e-10	1.4e-10	1.4e-10	2.6e-10	3.5e-10
0.9995	6.25e-06	1.6e-10	1.3e-10	1.3e-10	2.4e-10	3.5e-10
0.9995	0.0001	3e-10	1.3e-10	1.2e-10	7.9e-10	4.4e-10
0.999	6.25e-06	2.4e-10	1.6e-10	1.6e-10	4.7e-10	4.1e-10
0.999	2.5e-05	4.5e-10	1.7e-10	1.7e-10	9.8e-10	4.3e-10
0.999	0.0001	8.2e-10	1.6e-10	1.6e-10	1.7e-09	6.4e-10

Table 3 Average covariance error for jump model

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	0.2	0.2	0.2	0.48	0.51
0.9999	6.25e-06	0.22	0.22	0.22	0.47	0.52
0.9999	0.0001	0.73	0.21	0.21	0.89	0.56
0.9995	6.25e-06	0.29	0.21	0.21	0.55	0.58
0.9995	0.0001	3.5	0.18	0.18	2.9	0.57
0.999	6.25e-06	0.36	0.21	0.21	0.67	0.63
0.999	2.5e-05	1.1	0.21	0.21	1.5	0.68
0.999	0.0001	4.8	0.2	0.2	4.6	0.73

This choice of ν ensures that when the innovation achieves its mean absolute value ,$\sqrt{\frac{2\Gamma_{m,m}}{\pi}}$, the probability of an observation will be p_{Obs}. We set $p_{Obs} = 0.3$ in our numerical experiments.

The performance results for different values of the jump parameters are shown Tables 2 and 3. For the majority of cases we see that the KECM approaches outperform the other methods when jumps are present. In Figure 4 we show the Kalman estimate of the true price for KEM and KECM-Laplace. The figure highlights the disadvantage of the KEM algorithm in the presence of jumps, namely that it over smoothes prices near jumps.

4.4 Simulated Data from GARCH(1,1)-jump model

In addition to the jump diffusion model we also evaluate the algorithms against a multivariate GARCH(1,1)-jump pricing model [14,28,8], where the effect of jumps persists in the price volatility. Using the GARCH(1,1)-jump model the log- price data is generated as

$$X_i(t) = X_i(t-1) + \sqrt{h_i}V_i(t) + J_i(t)Z_i(t) + D$$

$$h_i(t+1) = b_i h_i(t) + a_i(X_i(t) - X_i(t-1) - D)^2 + c_i$$

$$h_i(0) = \Gamma_{i,i}$$

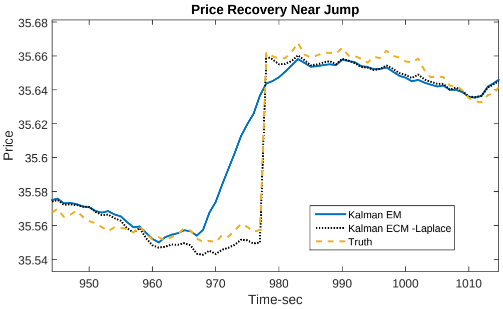

Fig. 4 Price estimate example from the KEM and KECM-Laplace algorithms. This is an example of the KEM algorithm over-smoothing near a small jump in price

Table 4 Portfolio variance for GARCH(1,1)-jump model.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	1.3e-10	1.3e-10	1.3e-10	2.5e-10	4e-10
0.9999	6.25e-06	1.6e-10	1.6e-10	1.5e-10	2.4e-10	3.1e-10
0.9999	0.0001	1.6e-10	1.3e-10	1.3e-10	4.4e-10	4.2e-10
0.9995	6.25e-06	2e-10	1.5e-10	1.5e-10	4.4e-10	4.6e-10
0.9995	0.0001	3.7e-10	1.3e-10	1.4e-10	1e-09	3.9e-10
0.999	6.25e-06	2.6e-10	1.4e-10	1.4e-10	5.8e-10	4.5e-10
0.999	2.5e-05	5.5e-10	1.5e-10	1.7e-10	1.4e-09	7e-10
0.999	0.0001	1.1e-09	1.6e-10	1.5e-10	2e-09	6.8e-10

where a_i, b_i, c_i are non-negative with $b_i + a_i < 1$ and $c_i = \Gamma_{i,i}(1 - a_i - b_i)$. Here $V(t)$ is modeled as multivariate normal with

- $V_i(t) \sim \mathcal{N}(0,1)$
- $\mathbb{E}V_i(t)V_j(t) = \frac{\Gamma_{i,j}}{\sqrt{\Gamma_{i,i}\Gamma_{j,j}}}$
- $\mathbb{E}V_i(t_1)V_j(t_2) = 0$ for $t_1 \neq t_2$.

The value of c ensures that in the absence of jumps, the long term average volatility for the i^{th} asset will be $\sqrt{\Gamma_{i,i}}$. We also see that the correlation coefficient between any two assets is constant [8].

In these experiments $a_i = 0.3$ and $b_i = 0.5$. This allows for volatility clustering which has been observed in many empirical stock return data. All other parameters such as the covariance matrix are identical to the previous experiment.

The results for the GARCH(1,1)-jump model are shown in Tables 4 - 5. From these tables we see that both KECM algorithms are robust to the volatility clustering exhibited in GARCH models.

Table 5 Average covariance error for GARCH(1,1)-jump model.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	0.37	0.37	0.38	0.5	0.52
0.9999	6.25e-06	0.43	0.37	0.38	0.58	0.55
0.9999	0.0001	3.3	0.39	0.4	1.7	0.55
0.9995	6.25e-06	0.88	0.42	0.43	0.81	0.62
0.9995	0.0001	18	0.65	0.49	8.5	0.61
0.999	6.25e-06	1.4	0.48	0.51	1.2	0.64
0.999	2.5e-05	7.7	0.64	0.62	4.5	0.69
0.999	0.0001	36	1.4	0.67	16	0.71

4.5 Simulated Data from GARCH(1,1)-jump Model and stochastic microstructure variance

In this section we test our algorithms under a GARCH(1,1)-jump model with stochastic microstructure variance. This microstructure noise model accounts for a positive correlation between the bid-ask spread and the squared innovation. This models an empirical phenomena that has been observed in many markets [39]. Here we assume the same efficient price innovation as the GARCH(1,1)-jump model but now we allow for time-varying variance in the microstructure noise. In this model the variance of the microstructure noise at time t for i^{th} asset is

$$\left(0.1\frac{(X_i(t)-X_i(t-1)-D)^2}{\Gamma_{i,i}}+0.9\right)\tilde{\sigma}_{o,i}^2$$

which is the sum of fixed variance and time varying term which is dependent on the efficient price innovation. Here we see that when the squared innovation equals the variance then the observation noise variance equals $\tilde{\sigma}_{o,i}^2$. As in the previous simulations, $\tilde{\sigma}_{o,i}^2$ is chosen to be a realization of a gamma distributed random variable with shape 2 and mean 0.0002^2.

The results for this model are shown in Tables 6 and 7. From a comparison with prior tables we see that the covariance errors are larger for the nonstationary microstructure noise model. Here the KECM-Laplace model is especially sensitive to the stochastic microstructure noise variance for $\sigma_j^2 = 1e-4$. In some cases the covariance error increases by about a factor of 10. The KECM-spike and slab approach is not as sensitive to the stochastic noise variance.

4.6 Numerical Results Summary

The following are key observations from the numerical simulation results:

1. Both KECM approaches outperform KEM in the presence of jumps.
2. Laplace prior underperforms spike and slab models for large jumps.

Table 6 Portfolio variance for GARCH(1,1)-jump model with stochastic microstructure noise variance.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	1.5e-10	1.6e-10	1.6e-10	2.2e-10	4.7e-10
0.9999	6.25e-06	1.6e-10	1.6e-10	1.6e-10	3e-10	2.8e-10
0.9999	0.0001	2e-10	1.6e-10	1.6e-10	5.1e-10	4.6e-10
0.9995	6.25e-06	2.6e-10	1.9e-10	1.9e-10	4.6e-10	5.1e-10
0.9995	0.0001	5.1e-10	1.8e-10	1.8e-10	1.5e-09	7.4e-10
0.999	6.25e-06	2.3e-10	1.5e-10	1.5e-10	5e-10	5.4e-10
0.999	2.5e-05	5.6e-10	1.7e-10	1.7e-10	1.4e-09	1e-09
0.999	0.0001	9e-10	2e-10	1.6e-10	2.3e-09	7.4e-10

Table 7 Average covariance error for GARCH(1,1)-jump model with stochastic microstructure noise variance.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	0.37	0.37	0.38	0.57	0.6
0.9999	6.25e-06	0.51	0.42	0.42	0.55	0.56
0.9999	0.0001	21	1.5	0.38	2.6	0.56
0.9995	6.25e-06	0.78	0.4	0.41	0.82	0.55
0.9995	0.0001	75	3.3	0.47	9.5	0.67
0.999	6.25e-06	1.2	0.41	0.44	1	0.6
0.999	2.5e-05	13	0.48	0.51	3.5	0.79
0.999	0.0001	1.3e+02	13	2.7	13	2.2

3. Spike and slab models are more robust to stochastic microstructure noise variance than the Laplace prior model.
4. Pairwise refresh with jump correction performed worse than pairwise refresh without jump correction for scenarios where jumps were infrequent and small.
5. Both KECM methods perform similarly to the KEM approach when jumps are not present,infrequent or small.

The first observation is not surprising since both KECM approaches explicitly account for jumps. The second and third observations may be the result of a large jump estimation bias that can occur when using the Laplace prior for large σ_j^2. Observation 4 could be due to overcompensation or incorrect detection of jumps in the pairwise refresh with jump correction approach. This is in contrast to the KECM approaches which performed well for small and infrequent jumps as well as large and frequent jumps.

5 Conclusion

This work has introduced two jump robust KECM methods for estimating asset return covariance from high-frequency data. The methods address 3 features found in high frequency data: 1) asynchronous returns, 2) market microstructure noise, and 3) jumps. Jumps were addressed using both Laplace and spike and slab distributed models.

Both proposed techniques improve covariance estimation performance versus existing methods when jumps are present. When comparing the spike and slab and Laplace jump models, the spike and slab approach demonstrated more robustness especially to larger jumps and stochastic microstructure noise variance.

Many problems remain to be investigated. As future work other jump models besides spike and slab and Laplace can be considered. For example both the spike and slab and Laplace priors create a bias in the jump estimates. The use of other penalties which induce less penalty for large jumps may reduce this bias and improve estimation performance. Another area that can be addressed is global convergence of the KECM algorithms. Since the KECM does not necessarily converge to a globally optimal solution additional performance gains may be achievable by attempting multiple initializations or other approaches. Further questions to study include convergence rates of KECM type algorithms, asymptotic/non-asymptotic properties of our estimators, and applications to complex real-world financial data.

References

1. A. Aravkin, B. Bell, J. Burke, and G. Pilonetto. An ℓ_1 Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 56(12), Dec. 2011.
2. Y. Aït-Sahalia, J. Fan, and D. Xiu. High-Frequency Covariance Estimates With Noisy and Asynchronous Financial Data. *Journal of the American Statistical Association*, 105(492):1504–1517, 2010.
3. Y. Aït-Sahalia, P. Myklank, and L. Zhang. How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *Review of Financial Studies*, 100:1394–1411, 2005.
4. F. Bandi and J. Russell. Separating microstructure noise from volatility. *Journal of Financial Economics*, 79:655–692, 2006.
5. O. Barndorff-Nielsen, P. Hansen, A. Lunde, and N. Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162, 2011.
6. C.B. Barry. Portfolio Analysis under Uncertain Means,Variances and Covariances. *Journal of Finance*, 29:515–522, May 1974.
7. A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Imaging Scienses*, 2(1):183–202, 2009.
8. T. Bollerslev. Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach. *Review of Economics and Statistics*, 72:498–505, 1990.
9. K. Boudt, C. Croux, and S. Laurent. Outlyingness weighted covariation. *Journal of Financial Econometrics*, 9(4), 2011.
10. K. Boudt and J. Zhang. Jump robust two time scale covariance estimation and realized volatility budgets. *Quantitative Finance*, 15(6), 2015.
11. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
12. J. Campbell, A. Lo, and A.C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1996.

13. E. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimzation. *Journal of Fourier Anal. Appl.*, 14:877–905, 2008.
14. W. Chan and J. Maheu. Conditional jump dynamics in stock market returns. *Journal of Business and Economic Statistics*, 20(3):377–389, 2002.
15. F. Corsi, S. Peluso, and F. Audrino. Missing in Asynchronicity: A Kalman-EM Approach for Multivariate Realized Covariance Estimation. *Journal of Applied Econometrics*, 30(3), 2015.
16. V. DeMiguel, L. Garlappi, and R. Uppal. Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.
17. A. Dempster, N Laird, and D. Rubin. Maximum likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.
18. J. Fan, Y. Li, and K. Yu. Vast Volatility Matrix Estimation Using High Frequency Data for Portfolio Selection. *Journal of the American Statistical Association*, 107(497):412–428, 2012.
19. J. Fan and Y. Wang. Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102(480):1349–1362, 2007.
20. D. Fink. A compendium of conjugate priors. Technical report, Montana State Univeristy, 1997.
21. Z. Ghahramani and G.E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 2000.
22. T Goldstein and S. Osher. The Split Bregman Method for ℓ_1 Regularized problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009.
23. J.D. Jobson and B. Korkie. Estimation for Markowitz Efficient Portfolios. *Journal of the American Statistical Association*, 75:544–554, Sept 1980.
24. J. Karpoff. The relation between price changes and trading volume: A survey. *The Journal of Financial and Quantitative Analysis*, 22:109–126, Mar. 1987.
25. C. Liu and C.Y. Tang. A quasi-maximum likelihood approach for integrated covariance matrix estimation with high frequency data. *Journal of Econometrics*, 180, 2014.
26. A. Lo and A.C. MacKinlay. An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45:181–211, 1990.
27. D. Lunenberger and Y. Ye. *Linear and Nonlinear Programming.* New York,NY, 2008.
28. J.M. Maheu and T.H. McCurdy. News arrival, jump dynamics, and volatility components for individual stock returns. *Journal of Finance*, 59(2):755–793, 2004.
29. J. Mattingley and S. Boyd. Real-Time Convex Optimization in Signal Processing. *IEEE Signal Processing Magazine*, May 2010.
30. X.L Meng and D. Rubin. Maximum likelihood estimation via the ECM algorithm:A general framework. *Biometrika*, 80:267–278, 1993.
31. S. Peluso, F. Corsi, and A. Mira. A Bayesian High-Frequency Estimator of the Multivariate Covariance of Noisy and Asynchronous Returns. *Journal of Financial Econometrics*, 13(3):665–697, 2015.
32. R. Roll. A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market. *The Journal of Finance*, 39(4):1127–1139, 1984.
33. M.W. Seeger. Bayesian Inference and Optimal Design for the Sparse Linear Model. *Journal of Machine Learning Research*, 9:759–813, 2008.
34. R. Shumway and D. Stoffer. *Time Series Analysis and its Applications with R Examples.* Springer, 2011.
35. R.H. Shumway and D.S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.
36. C.F. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.
37. W. Zangwill. *Nonlinear Programming: A Unified Approach.* Prentice-Hall, 1969.
38. L. Zhang. Estimating covariation: Epps effect and microstructure noises. *Journal of Economics*, 160(1):33–47, 2011.
39. M. Zhang, J. Russel, and R. Tsay. Determinants of bid and ask quotes and implications for the cost of trading. *Journal of Empirical Finance*, 15(4):656–678, Sept. 2008.

Appendices

A Kalman Smoothing Equations

The Kalman smoother can be used to compute the posterior distribution of $X(t)$ given Y and an estimate of $\Theta = [D, \Gamma, \Sigma_o', J]$. From [35] the posterior distribution is normal and is completely characterized by the following quantities for $m = T$

$$\bar{X}(t|m) = \mathbb{E}(X(t)|y(1:m))$$

$$P(t|m) = cov(X(t), X(t)|y(1:m))$$
$$P(t, t-1|m) = cov(X(t), X(t-1)|y(1:m)).$$

These values can be computed efficiently using a set of well known forward and backward recursions [34] known as the Rauch-Tung-Striebel (RTS) smoother. The forward recursions are

$$\begin{aligned}
\bar{X}(t|t-1) &= \bar{X}(t-1|t-1) + D + J(t) && (24)\\
P(t|t-1) &= P(t-1|t-1) + \Gamma && (25)\\
G(t) &= P(t|t-1)I(t)^T \left(I(t)P(t|t-1)I(t)^T + \Sigma_o^2(t) \right)^{-1} && (26)\\
\bar{X}(t|t) &= \bar{X}(t|t-1) + G(t)(y(t) - I(t)\bar{X}(t|t-1)) && (27)\\
P(t|t) &= P(t|t-1) - G(t)I(t)P(t|t-1) && (28)
\end{aligned}$$

with $\bar{X}(0|0) = \mu$ and $P(0|0) = K$.

The backward equations are given by

$$\begin{aligned}
H(t-1) &= P(t-1|t-1)P(t|t-1)^{-1}\\
\bar{X}(t-1|T) &= \bar{X}(t-1|t-1) + H(t-1)(\bar{X}(t|T) - \bar{X}(t|t-1))\\
P(t-1|T) &= P(t-1|t-1)\\
&\quad + H(t-1)(P(t|T) - P(t|t-1))H(t-1)^T.
\end{aligned}$$

A backward recursion for computing $P(t, t-1|T)$ is

$$\begin{aligned}
P(t-1, t-2|T) &= P(t-1|t-1)H(t-2)^T\\
&\quad + H(t-1)\left(P(t, t-1|T) - P(t-1|t-1)\right)H(t-2)^T
\end{aligned}$$

where

$$P(T, T-1|T) = (I - G(T)I(T))P(T-1|T-1). \quad (29)$$

B Convergence of KECM Algorithms

Convergence of the EM and ECM algorithms in general is considered in [36] and [30] respectively. It is shown in [30] that the ECM algorithm converges to stationary point of the log posterior under the following mild regularity conditions

1. Any sequence $\Theta^{(k)}$ obtained using the ECM algorithm lies in a compact subset of the parameter space, Ω. For our case we need to restrict the parameter space such that $\sigma_o^2 \neq 0$ and Γ is positive definite.
2. $\mathcal{G}(\Theta, \Theta')$ is continuous in both Θ and Θ'.
3. The log posterior $L(\Theta)$ is continuous in Ω and differentiable in the interior of Ω.

B.1 Algorithm 1

Since the Laplace prior on J is not differentiable condition 3 is not satisfied and the results in [30] are not directly applicable. However the proofs and solution set in [30] can be modified to handle this irregularity.

Before addressing condition 3 we first verify condition 1. We start by examining the sequence of covariance estimates $\Gamma^{(k)}$.

Lemma 1 *Assume a noisy asset price is observed at least one time for each asset for $t > 1$ and that $\tilde{I}(t) \neq 0$ for all t. Let $\Gamma^{(k)}$ be a sequence of solutions obtained with Algorithm 1, where $\Gamma^{(0)}$ is positive definite. Then sequences $\Gamma^{(k)}$ and $\frac{1}{s^{(k)}}$ are bounded where $s^{(k)}$ is the minimum eigenvalue of $\Gamma^{(k)}$. In addition the sequence $\sigma_{o,i}^{2,(k)}$ is bounded below and above by positive values for all i.*

Proof Since W_o is positive definite we have from equation (9) that $s^{(k)}$ is bounded below by a positive constant which implies $\frac{1}{s^{(k)}}$ is bounded. Similarly by equation (10) we have $\sigma_{o,i}^{2,(k)}$ is bounded below by a positive constant. To prove that $\Gamma^{(k)}$ is bounded we note that the posterior may be written as

$$\begin{aligned} p(\theta|y) &= C_1 p(y|\theta)p(\theta) \\ &= C_1 p(y(1)|\theta)p(\theta)\prod_{t=2}^{T} p(y(t)|y(1:t-1),\theta) \\ &\leq C_2 p(y(1)|\theta)\prod_{t=2}^{T} p(y(t)|y(1:t-1),\theta) \end{aligned}$$

where C_1 is a constant not dependent on θ and where $C_2 = C_1 \sup_\theta p(\theta)$. Note that $C_2 < \infty$.

For $t > 1$ each of the conditional distributions $p(y(t)|y(1:t-1),\theta)$ is a normal distribution with covariance

$$Q(t) = \tilde{I}(t)P(t|t-1)\tilde{I}(t)^T + \sigma_o^2 I$$

where for notational simplicity we suppress the dependence of $Q(t)$ and $P(t|t-1)$ on k. Since $\sigma_{o,i}^2$ is bounded below by a positive value, it follows that $\frac{1}{|Q(t)|}$ is bounded.

Now suppose that $\Gamma^{(k)}$ is unbounded. Then since

$$P(t|t-1) = P(t-1|t-1) + \Gamma$$

$P(t|t-1)$ is unbounded as k goes to ∞. Since an observation of each asset's price occurs at least once for $t > 1$ it follows that $Q(\tau)$ is unbounded (as $k \to \infty$) for some $\tau > 1$. Then since the smallest eigenvalue of $Q(\tau)$ is bounded below by a positive constant, the determinant of $Q(\tau)$ is unbounded. Thus a subsequence of $p(y(\tau)|y(1:\tau-1),\Theta^{(k)})$ will approach 0. Since $\frac{1}{|Q(t)|}$ is bounded, $p(y(t)|y(1:t-1),\Theta^{(k)})$ will remain bounded above for all t. Then using (30) we have

$$\begin{aligned} p(\theta|y) &\leq C_2 p(y(1)|\theta)\prod_{t=2}^{T} p(y(t)|y(1:t-1),\theta) \\ &= C_2 p(y(\tau)|y(1:\tau-1),\theta)\prod_{t\neq\tau}^{T} p(y(t)|y(1:t-1),\theta) \end{aligned}$$

which implies a subsequence of $p(\Theta^{(k)}|y)$ will converge to 0. This contradicts the monotonicity of the ECM algorithm [30]. The proof that the sequence $\sigma_{o,i}^{2,(k)}$ is bounded above for all i is similar.

Lemma 2 *Assume the conditions of Lemma 1. Let $\lambda(t)^{-1,(k)}$ be a sequence of solutions obtained with Algorithm 1 where $\Gamma^{(0)}$ is positive definite. Then there exist finite positive numbers a, b where $a \leq \lambda_i(t)^{(k)} \leq b$ for all t, k and i.*

Proof By the update equation (14) we may set $b = \frac{\alpha_\lambda + 2}{\beta_\lambda}$ which is positive and finite. By way of contradiction assume the lower bound does not hold. Then for some i and t there exists a subsequence $\lambda_i(t)^{(k_n)}$ such that $\lim_{n\to\infty} \lambda_i(t)^{-1,(k_n)} = \infty$. Since each $\lambda_i(t)^{-1}$ is the mode of an inverse gamma distribution it follows that the posterior scale parameter,$(\beta_\lambda + |j^{(k_n)}|)$ goes to infinity . This implies that $p(\lambda_i(t)^{-1,(k_n)}, j_i(t)^{(k_n)}) \to 0$. Since each prior density function is bounded as $\lambda_i(t) \to 0$ this implies that $p(\theta)$ goes to zero, contradicting the monotonicity of the ECM algorithm. Thus there exists an $a > 0$ such that $\lambda_i(t)^{(k)} > a$ for all t, k and i.

Now we prove that the sequences $J^{(k)}$ and $D^{(k)}$ are also well behaved.

Lemma 3 *Assume the conditions of Lemma 1. Let $J^{(k)}$ and $D^{(k)}$ be sequences of solutions obtained with Algorithm 1 where $\Gamma^{(0)}$ is positive definite. Then sequences $J^{(k)}$ and $D^{(k)}$ are bounded.*

Proof From Lemma 1 the likelihood $p(y|\theta)$ is bounded above. Recall from the previous lemma that there exists an $a > 0$ such that for all k, $\lambda_i(t)^{(k)} \geq a$. Since the prior density function is bounded above for each parameter it follows that $\lim_{j\to\infty} p(\theta) = 0$. This implies $J^{(k)}$ is bounded by the monotonicity of the ECM algorithm. Since $\lim_{d\to\infty} p(\theta) = 0$ it also follows that $D^{(k)}$ is bounded.

The above lemmas imply the following corollary.

Corollary 1 *The sequence $\Theta^{(k)}$ is bounded and all limit points are feasible (e.g. variance non-zero, positive definite covariance).*

Now we derive some additional properties of the limit points of $\Theta^{(k)}$. To do this we shall refer to Zangwill's convergence theorem [37]. To use Zangwill's theorem, we first define $\mathcal{A}$ to be a point to set mapping defined by the ECM algorithm i.e. $\Theta^{(k+1)} \in \mathcal{A}(\Theta^{(k)})$. Let us define a solution set, $\mathcal{S}$, as the set of θ such that

$$\begin{aligned}
\theta_1 &= \arg\max_v \mathcal{G}\left([v, \theta_2, \theta_3, \theta_4, \theta_5], \theta\right) \\
\theta_2 &= \arg\max_v \mathcal{G}\left([\theta_1, v, \theta_3, \theta_4, \theta_5], \theta\right) \\
\theta_3 &= \arg\max_v \mathcal{G}\left([\theta_1, \theta_2, v, \theta_4, \theta_5], \theta\right) \\
\theta_4 &= \arg\max_v \mathcal{G}\left([\theta_1, \theta_2, \theta_3, v, \theta_5], \theta\right) \\
\theta_5 &= \arg\max_v \mathcal{G}\left([\theta_1, \theta_2, \theta_3, \theta_4, v], \theta\right).
\end{aligned}$$

By definition $\theta \in \mathcal{A}(\theta)$ for all $\theta \in \mathcal{S}$. This along with the monotonicity of the ECM algorithm implies that $L(\theta)$ is an ascent function, i.e.

$$\begin{aligned}
L(\theta') > L(\theta) &\text{ for all } \theta \notin \mathcal{S}, \theta' \in \mathcal{A}(\theta) \\
L(\theta') \geq L(\theta) &\text{ for all } \theta \in \mathcal{S}, \theta' \in \mathcal{A}(\theta).
\end{aligned}$$

Since $\mathcal{G}(\theta, \theta')$ is continuous in both θ and θ' we have that $\mathcal{A}$ is a closed mapping. Thus we have the following theorem.

Theorem 1 *All limit points of $\Theta^{(k)}$ belong to $\mathcal{S}$.*

Proof This is a direct consequence of Zangwill's convergence theorem [37] (also known as the Global convergence theorem [27]). To invoke the theorem we must meet the following conditions

- $\Theta^{(k)}$ belongs to a compact subset of the feasible solutions
- $\mathcal{A}$ is closed
- There exists a continuous ascent function

All three of these conditions were shown above, thus the theorem follows from Zangwill's convergence theorem.

Now we show that if $\theta' \in \mathcal{S}$ then θ' is in some sense a "stationary" point of the log posterior $L(\theta) = \log p(\theta|y)$.

Theorem 2 *Let $\theta' \in \mathcal{S}$. Then*

$$\nabla_{\theta_i} L(\theta)_{|\theta=\theta'} = 0 \text{ for } i \in 1, 2, 3, 5$$

and

$$0 \in \partial_{\theta_4} L(\theta)_{|\theta=\theta'}.$$

Proof To show this we first note that $L(\theta)$ can be written as [30]

$$L(\theta|y) = \mathcal{G}(\theta, \theta') - H(\theta, \theta')$$

where

$$H(\theta, \theta') = \mathbb{E}_{p(x|y,\theta')} \log p(X|y, \theta).$$

From the information inequality we have that $H(\theta', \theta') \geq H(\theta, \theta')$ for all feasible θ. Since $H(\theta, \theta')$ is differentiable with respect to θ it follows that

$$\nabla_\theta H(\theta, \theta')_{|\theta=\theta'} = 0.$$

Since $\nabla_{\theta_i} \mathcal{G}(\theta, \theta')_{|\theta=\theta'} = 0$ for $i \in 1, 2, 3, 5$ it follows that

$$\nabla_{\theta_i} L(\theta)_{|\theta=\theta'} = 0 \text{ for } i \in 1, 2, 3$$

Also since $\mathcal{G}(\theta, \theta')$ and $H(\theta, \theta')$ are convex in j, and $\theta' \in \mathcal{S}$, it follows that

$$0 \in \partial_{\theta_4} \mathcal{G}(\theta, \theta')$$

which implies

$$0 \in \partial_{\theta_4} L(\theta, \theta').$$

B.2 Algorithm 3

Analogous results to Corollary 1 and Theorem 1 may proven for Algorithm 3 using same arguments as Algorithm 1. The following result is analogous to Theorem 2.

Theorem 3 *Let $\theta' \in \mathcal{S}$ where $\mathcal{S}$ is the set of fixed points of the Algorithm 3. Then*

$$\nabla_{\theta_i} L(\theta)_{|\theta=\theta'} = 0 \text{ for } i \in 1, 2, 3, 5, 6.$$

The proof of this result is the same as Theorem 2.

C Procedure for selecting $q(\lambda)$

In this section we outline the method for selecting the distribution $q(\lambda)$ for a special case of when the prior distribution of volatility of each asset is identical. Suppose the squared volatility of each asset return is inverse gamma distributed with scale c and shape η. Let σ_v^2 be distributed as $IG(c, \eta)$ and be statistically independent of ζ' and σ_j^2.

To determine an appropriate prior distribution of λ we first obtain samples of λ,

$$\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{M_\lambda}$$

by the performing the following steps

1. For $k = 1, \ldots, M_\lambda$
2. Draw independent samples from the distribution of $(\sigma_v^{2'}, \zeta', \sigma_j^{2'})$. This is relatively straight forward using standard statistical functions due to the independence assumptions.
3. Determine a λ' such that $\lambda' \sim_{\sigma_v^{2'}} (\zeta', \sigma_j^{2'})$. This can be done via Monte Carlo integration as shown below.
   - For a large number L draw a sample $v_1 \ldots v_L$ from the distribution $\mathcal{N}(0, \sigma_v^2)$.
   - Compute $P_i = Pr(J = 0 | J + V = v_i)$, where $J \sim SpikeSlab(\zeta', \sigma_j^{2'})$. The value of P_i is
   $$\frac{\frac{\zeta'}{\sqrt{\sigma_v^{2'}}} \exp(-v_i^2/(2\sigma_v^2))}{\frac{\zeta'}{\sqrt{\sigma_v^{2'}}} \exp(-v_i^2/(2\sigma_v^{2'})) + \frac{1-\zeta'}{\sqrt{\sigma_v^{2'}+\sigma_j^{2'}}} \exp(-v_i^2/(2(\sigma_v^{2'} + \sigma_j^{2'})))}.$$
   - Compute the simulated empirical mean $\bar{P} = \frac{1}{L}\sum_{i=1}^{L} P_i$.
   - Choose λ' such that (5) is satisfied with $\mathbb{E}_{p(y_2|J_2=0)} Pr(J_2 = 0|Y_2)$ approximated as $\bar{P}$. This value is given below
   $$\lambda' = \frac{\operatorname{erf}^{-1}(\bar{P})\sqrt{2\sigma_v^{2'}}}{\sigma_v^{2'}}$$
   where $\operatorname{erf}^{-1}()$ is the inverse error function.
4. Set $\tilde{\lambda}_k = \lambda'$
5. Goto step 1

Examples of histograms of samples obtained using the above procedures are shown in Figure 5. Once we obtain samples of λ we fit a smooth distribution to the sampled data. Since the gamma distribution is a conjugate prior to the Laplace distribution a gamma distribution is a convenient choice for $q(\lambda)$. Furthermore examination of Figure 5 indicates that a gamma distribution is a reasonable approximation. Thus we choose

$$q(\lambda) = \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma_f(\alpha_\lambda)} \lambda^{\alpha_\lambda - 1} \exp\left(-\lambda\beta_\lambda\right)$$

where $\Gamma_f()$ is the gamma function. Here α_λ and β_λ can be selected using maximum likelihood or method of moments.

Since $q(\lambda)$ develops a singularity near zero for large values of β_λ we shall impose a prior of λ^{-1} rather than λ. We denote this prior as $q_{inv}(\lambda^{-1})$. Since λ is gamma distributed with shape α_λ and rate β_λ it follows that $q_{inv}(\lambda^{-1})$ is the inverse gamma distribution with shape α_λ and scale β_λ

$$q_{inv}(\lambda^{-1}) = \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma_f(\alpha_\lambda)} (\lambda^{-1})^{-\alpha_\lambda - 1} \exp\left(-\frac{\beta_\lambda}{\lambda^{-1}}\right).$$

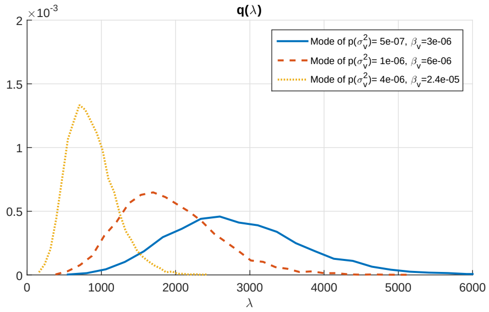

Fig. 5 Normalized histograms of λ samples. In all experiments $\sigma_j^2 \sim IG(10, 0.0011)$,$\zeta \sim$ Beta$(5, 1.0201)$, $\sigma_v^2 \sim IG(5, \beta_v)$.