A fast and accurate variational inference for a large dimensional Markov Switching model [*]

Xuan Vu[†]

November 13, 2023

Abstract

The multivariate Markov switching model identifies bull and bear markets from disaggregated indices and allows for asset allocation decisions. However, the Monte Carlo Markov chain algorithm's computational cost increases significantly with its dimension and quickly reaches a prohibited computational time. I propose a new variational inference (VI) algorithm to estimate a large-dimensional Markov switching model fast and accurately in the bull and bear markets. While taking substantially less time to compute, this method achieves comparable in-sample and out-of-sample results to its MCMC counterpart. In addition, this inference allows for the inclusion of important restrictions to identify hidden market states. The forward filtering backward smoothing algorithm of my novel VI is also common in economic literature. My simulation studies emphasize the accuracy and timely benefit of the new technique in, for example, identifying the bull and bear states, detecting regime switching, and providing forecasts for investment strategies. I investigate the empirical applications of three sets of stock returns that are listed in the S&P 500 and one set of industry portfolios and find similar insights.

Keywords: Variational inference, returns, multivariate, Markov switching model
JEL Classification: C11, C32, C55, G11, G17

[*] I thank seminar participants at the Melbourne Bayesian Econometrics Workshop 2023, Vietnam Economist Annual Meeting 2023 for useful feedback. I acknowledge the financial support of University of Melbourne's FBE Ph.D. Scholarship for this research.

[†] Department of Economics, University of Melbourne and Department of Quantitative Analysis, Vietnam National University of Agriculture. Email: khacv@student.unimelb.edu.au

1 Introduction

The Markov Switching (MS) model plays a crucial role in identifying bull and bear states, which typically represent markets with positive and negative mean returns, respectively. The advantages of this model include endogenously allocating data into regimes, capturing nonlinear temporal structures, and considering the variance of mean returns. Moreover, identification and forecasting are feasible in a single process. These advantages are extensively discussed in Kole and Van Dijk (2017) and Maheu and McCurdy (2000). In this paper, our focus is on providing a feasible and more computationally efficient inference for the multivariate Markov switching model, especially in the context of a large dimension.

The univariate MS model was first introduced by Hamilton (1989) and Turner et al. (1989) and has since found widespread application in economics. This model utilizes a Markov chain of a latent discrete variable to capture the behavior of a time series. Most bull and bear applications are under the univariate setting, as seen in works such as Rydén et al. (1998), Maheu and McCurdy (2000), Haase and Neuenkirch (2023), and Chen (2009).

Large-dimensional models have gained importance in economic literature, even though their dimension remains limited. These models can exploit a larger amount of information essential in economic analysis and forecasting. For instance, in asset allocation, understanding the structure of financial time series correlation is crucial, and the variation of correlation is beneficial for forecasting future correlation. The multivariate setting also enables the use of disaggregated asset returns. Additionally, the recent availability of more data supports applications of these models. Typical examples include Chan et al. (2011), Chevallier (2012), Guidolin and Timmermann (2007), Guidolin and Timmermann (2008), and Liu and Maheu (2018). Unfortunately, all of these applications involve less than ten time series in their specification. Moreover, the purpose of these papers is not focused on identifying the bull and bear state.

The challenge of a large-dimensional MS model lies in its computational cost. The frequentist approach, maximum likelihood estimation, is feasible if regularity conditions are satisfied. However, as the number of series and states increases, the model parameters explode, rendering its log-likelihood function unbounded. Consequently, maximum likelihood becomes infeasible. The Bayesian approach resolves these dimensional problems, providing exact inference and good predictive performance. Hence, the Monte Carlo Markov chain (MCMC) algorithm is a popular solution for MS models (Frühwirth-Schnatter, 2006). However, this method has a disadvantage that is a significant increase in estimation cost when the model's dimension increases.

Variational inference (VI) is popular in Bayesian econometrics because it can approximate a high-dimensional model with a large dataset using less computational time (Blei et al., 2017, Ormerod and Wand, 2010). This serves as a substitute for MCMC and is useful in dealing with large-scale problems or models with high complexity.

However, variational inference has shortcomings in its application to the large MS model. Firstly, studies on variational inference for the MS model do not impose economic restrictions. For example, typical works on the univariate version (Beal, 2003, McGrory and Titterington, 2009) and several on the multivariate version (Foti et al., 2014, Gruhl and Sick, 2016, Ji et al., 2006) do not possess any restrictions. While these restrictions are crucial in identifying bull and bear states, including them would violate the conjugacy assumption of VI. Specifically, an economic restriction will form a truncated multivariate distribution of the mean. Consequently, this truncation invalidates the assumption of the exponential family, which is essential in VI.

Secondly, these previous studies employ an unfamiliar approach to economists. They address the use of the Baum–Welch forward-backward algorithm, originating from (Baum and Petrie, 1966, Baum et al., 1970), which is not common in economic literature Frühwirth-Schnatter (2006). Both algorithms share the idea of expectation-maximization.

This paper proposes a variational inference approach to address the aforementioned problems. The proposed inference method estimates a large-dimensional Markov switching (MS) model much faster than the commonly used Monte Carlo Markov chain algorithm while maintaining the same level of result accuracy in-sample and out-of-sample. This method incorporates economic identification restrictions to detect hidden states without invalidating the variational inference assumptions. Additionally, it introduces a new variational forward filtering-backward smoothing algorithm, originally proposed by Chib (1996), which is well-established in economic literature and familiar to economists. I present two versions in which the data can follow either a multivariate Normal or a multivariate Student t distribution.

I conducted a simulation study to measure the in-sample and out-of-sample performance and the computational cost of this VI algorithm. The benchmark is the result of models with and without restrictions estimated by the MCMC approach. These include models under Normal and Student t distributed innovations and models with a large number of assets in time series. This paper provides an alternative approach to estimating the model under the Student t distribution, as well as the method in literature such as in Christmas and Everson (2010). I used the computing system SPARTAN with an Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz running MATLAB 2021b.

Regarding in-sample performance, I find the estimated variational parameters are almost the same as the posterior parameters. The estimated variational probability is highly correlated with the conditional posterior probability of each state, with a correlation coefficient of almost 1. I obtain the average values of Kullback-Leibler divergence to the true distribution of the mean, log score, and mean squared errors (Chan and Yu, 2022) from 100 different samples under the same data generation process. There is no discrepancy in results between the two algorithms with a small dimension. This gap is slightly larger once I use 30 series. However, its magnitude is relatively small compared to the value of the corresponding measure.

Regarding out-of-sample performance, I find the results from the VI approach comparable to the ones from the MCMC method. This result is consistent with observations from Frazier et al. (2022), Quiroz et al. (2022). I use the average predictive log score and mean squared forecasting errors (Gefang et al., 2022) to evaluate. The difference between results is negligible with a forecasting window of 100 periods. Although the discrepancy is larger when there are more series, its relative magnitude is not large.

I find the economic value of this algorithm by applying it to investment strategies. I conduct one-period-ahead forecasts for use in market timing portfolios and mean-variance portfolios. For market timing, I compare three strategies: buying an equally weighted portfolio and holding; using the probability of the bull state to signal the time to invest in a risk-free asset; and using the predicted probability to allocate the invested weight. For mean variance, I focus on minimizing the global variance portfolio using the predicted variance-covariance matrix and maximizing the Sharpe ratio. My results show minor gaps between the two algorithms, with a slight favor for the VI approach.

The significant gain from this new algorithm is its computational time. The time complexity in each iteration is similar between the VI and MCMC methods. However, by quickly converging, the VI approach significantly improves computational time. In my simulation with 30 series, my VI approach yields results after a few seconds, while the comparable MCMC needs more than 1,000 seconds.

I have applied the VI algorithm to the monthly stock returns from the S&P 500 and industry portfolios from the Kenneth French data library. Both sets of data contain multiple time series, from 34 to 103. It also spans a long horizon, from January 1926 to late 2022. My application results are consistent with my simulation results. I gain a significant advantage on the computational side. With the modest application of the 34 series, VI runs for 20 seconds, while MCMC needs more than 2 hours. In the most extreme case in my application, 103 series, my VI method costs only approximately 3 minutes, compared to almost 11 hours required by the MCMC algorithm. So VI can reduce computing time by 99

My algorithm has its weaknesses. Although the solution for my restricted VI algorithm is tractable, values in the solution can only be obtained through sampling. I use Botev (2017)'s sampling method to sample the truncated distribution because of its reliability.

The out-of-sample performance of the two approaches is similar in most cases. The log predictive score and mean squared forecasting errors are closed. The average global minimum variance and maximum Sharpe ratio are closed. The market-timing portfolio shows almost no gap between the two methods.

The paper is organised as follows: Section 2 presents the multivariate Markov switching model. Section 3 shows inference methods for the multivariate MS model. Section 4 presents methods to compare results from algorithms. Section 5 provides the empirical application using the S&P500 and Kenneth library's statistics, and Section 6 concludes.

2 Markov switching model

In this section, I focus on introducing the specification of the multivariate Markov switching (MS) model. I first introduce the model, and then I present identification restrictions.

I have a data set of returns on N assets at each time t with $t = 1, 2, \cdots, T$. The vector $y_t = \{y_{1,t}, y_{2,t}, \ldots, y_{N,t}\}'$ includes the returns of each index at time t. Suppose the MS model has K states. Then, I have the model as

$$y_t \mid s_t = k \sim \mathbb{N}(M_k, \Sigma_k) \tag{1}$$

$$P(s_{t+1} = j \mid s_t = i) = p_{ij} \tag{2}$$

$S = \{s_t\}_{t=1}^T$ is the latent variable denoting the latent state. $M = \{M_k\}_{k=1}^K$ and $V = \{V_k\}_{k=1}^K = \Sigma = \{\Sigma_k\}_{k=1}^K$ are the mean and the variance of the regime k, respectively. $M_k = (\mu_{1k}, \mu_{2k}, \ldots, \mu_{Nk})'$ is a $N \times 1$ vector while $V_k = \Sigma_k$ is a $N \times N$ variance covariance matrix. $\mathbb{N}$ denotes a multivariate normal distribution[1]. And P is a $K \times K$ transition probability matrix.

Note that, at time 1, there is no transition from the previous period. Hence, the latent state s_1 does not depend on the transition matrix P. Instead, I suppose that s_1 follows a discrete distribution where each state shares the same probability.

I assume two regimes ($K = 2$): regime 1 as the bear market and regime 2 as the bull market. Additionally, I adopt an equally-weighted portfolio as my target market portfolio. In the bear market, the equally-weighted portfolio from y_t has a negative mean, while it has a positive mean in the bull market. Let ι_N as a $N \times 1$ vector of ones. The restriction regarding the means is equivalent to

$$\iota_N^{'} M_1 < 0 \qquad \iota_N^{'} M_2 > 0$$

I extend the model with a multivariate Student t distribution. Financial time series often include information about financial and economic shocks. Hence, it is common to observe distributions of these series with thick tails. Student t distribution is more appropriate than the Normal distribution in the sense that the former may possess thick tails. Therefore, I also propose using a multivariate Student t distribution[2] in Equation 1 as follows

$$y_t \mid s_t = k \sim \mathbb{T}(M_k, \Sigma_k, \nu_k) \tag{3}$$

[1] Suppose $y = \{y_i\}_{i=1}^N$ follow a multivariate Normal distribution with the mean M and variance Σ. The multivariate Normal distribution has a probability density function

$$p(y \mid M, V) = (2\pi)^{-N/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - M)'\Sigma^{-1}(y - M)\right)$$

[2] Suppose $y = \{y_i\}_{i=1}^N$ follow a multivariate Student t distribution with the mean parameter M, the scale parameter Σ

where M_k is the mean, Σ_k is the scale parameter, ν_k is the degrees of freedom for component k. The variance covariance matrix of component k is $V_k = \frac{\nu_k}{\nu_k - 2}\Sigma_k$ provided that $\nu_k > 2$.

Because a Student t distribution does not have an exponential conjugate prior, its inference is not straightforward. One notices that this distribution is an infinite mixture of a scaled Normal distribution (see Appendix A). I then rewrite the returns equation as

$$y_t \mid s_t = k, \tau_t \sim \mathbb{N}(M_k, \tau_t^{-1}\Sigma_k)$$
$$\tau_t \mid \nu_{s_t} \sim \mathbb{G}\left(\frac{\nu_{s_t}}{2}, \frac{\nu_{s_t}}{2}\right)$$

where $\mathbb{G}$ denote the Gamma distribution[3]. Parameters of the gamma distribution are chosen in relation to degrees of freedom ν_{s_t}. Mean of the auxiliary variables τ_t is $\frac{\nu_{s_t}}{\nu_{s_t}} = 1$ and its variance is $\frac{2\nu_{s_t}}{\nu_{s_t}^2} = \frac{2}{\nu_{s_t}}$. This implies that the higher the value of ν_{s_t}, the mean of τ_t conditional on ν_{s_t} approaches 1. This reflects a property of a multivariate Student t distribution: when its degree of freedom gets larger, this distribution approaches a multivariate normal distribution.

2.1 Prior

The priors for the Normal MS model are

$$\Sigma \sim \mathbb{IW}(\Psi, n)$$
$$M \mid \Sigma \sim \mathbb{N}(\mu, h^{-1}\Sigma)$$
$$p(P) \sim \prod_{k=1}^{K} \mathbb{D}ir(\alpha_k)$$

and the degrees of freedom ν. The multivariate Student t distribution has the probability density function as

$$p(y \mid M, \Sigma, \nu) = \frac{\Gamma((\nu + N)/2)}{\Gamma(\nu/2)\nu^{N/2}\pi^{N/2}|\Sigma|^{1/2}}\left[1 + \frac{1}{\nu}(y - M)'\Sigma^{-1}(y - M)\right]^{-(\nu+N)/2}$$

[3] Suppose τ_i follow a Gamma distribution with the degrees of freedom v and the scale parameter s. The Gamma distribution has the probability density function as

$$p(\tau_i \mid v, s) = \frac{s^v}{\Gamma(v)}\tau_i^{v-1}e^{-s\tau_i}$$

where $\mathbb{IW}$ means Inverse Wishart distribution[4], and $\mathbb{D}ir$ means Dirichlet distribution[5]. Regarding the mean M, μ and h are its prior mean and scale parameters. Regarding the variance covariance Σ, the scale matrix is $\Psi = cov(y_{1:T})(n - N - 1)$ and the degrees of freedom is $n = N + 2$, so the prior mean of Σ is the sample covariance of y_t. Assume that there is no correlation between them, so I have Σ is an $N \times N$ identity matrix. The Dirichlet prior of the transition matrix favours the persistence of the states.

Similarly, priors for the multivariate Student t distribution are

$$\begin{aligned} M \mid \Sigma &\sim \mathbb{N}(\mu, h^{-1}\Sigma) & \Sigma &\sim \mathbb{IW}\left(\Psi, n\right) \\ \tau \mid \nu &\sim \mathbb{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) & \nu &\sim \mathbb{G}(v_\nu, s_\nu) \end{aligned}$$

At time 1, I assume a uniform probability for each latent state as

$$p(s_1 = k) = 1/K$$

for any k. Let Φ denote the set of time invariant parameters, M, Σ, P under normal assumption and M, Σ, P, ν with τ under Student t assumption.

2.2 Conditional posterior

The conditional posterior distribution of the Normal MS model is

$$\begin{aligned} p(\Phi, S \mid Y) &\propto p(M, \Sigma)p(P)p(S \mid P, \pi)p(Y \mid S, M, \Sigma) \\ &\propto \left[\prod_{k=1}^{K} p(M_k, \Sigma_k)\right] \left[\prod_{k=1}^{K} p(p_{k\cdot})\right] p(s_1)p(y_1 \mid M_{s_1}, \Sigma_{s_1}) \prod_{t=2}^{T} p(s_t \mid s_{t-1}, P)p(y_t \mid M_{s_t}, \Sigma_{s_t}) \end{aligned}$$

The conditional posterior distribution of the Student t MS model is

$$p(\Phi, S \mid Y) \propto p(M, \Sigma)p(P)p(S \mid P, \pi)p(Y \mid S, M, \Sigma, \tau)p(\tau \mid \nu)p(\nu)$$

[4] Suppose Σ follow an Inverse Wishart distribution with the degrees of freedom n and the scale matrix Ψ. The Inverse Wishart distribution has the probability density function as

$$p(\Sigma \mid n, \Psi) = \frac{|\Psi|^{n/2}}{2^{nN/2}\Gamma_N\left(\frac{n}{2}\right)} |\Sigma|^{-(n+N+1)/2} e^{-\frac{1}{2}tr(\Psi\Sigma^{-1})}$$

[5] The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \alpha_2, \ldots, \alpha_K > 0$ has a probability density function

$$p(y_1, \ldots, y_K \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} y_k^{\alpha_k - 1}$$

$$\propto \left[\prod_{k=1}^{K} p(M_k, \Sigma_k)\right] \left[\prod_{k=1}^{K} p(p_{k\cdot})\right] p(s_1) p(y_1 \mid M_{s_1}, \Sigma_{s_1}, \tau_{s_1})$$

$$\times \prod_{t=2}^{T} p(s_t \mid s_{t-1}, P) p(y_t \mid M_{s_t}, \Sigma_{s_t}, \tau_{s_t}) \left[\prod_{t=1}^{T} p(\tau_t \mid \nu_{s_t})\right] \left[\prod_{k=1}^{K} p(\nu_k)\right]$$

3 Estimation, inference and forecasting

3.1 MCMC

The most common approach to estimating this model is to use the MCMC algorithm. Because all time-invariant parameters have exponential conjugate priors, their estimation is straightforward. The difficulty is in the estimation of the conditional posterior $p(S \mid R, \Phi)$. The problem arises from the intertemporal dependence among the latent states.

Several algorithms to solve this problem are discussed in Frühwirth-Schnatter (2006). In general, these approaches use the forward filtering and backward smoothing algorithm to find the conditional posterior of s_t, $p(s_t \mid y_{1:T}, \Phi)$. The filtered pass is similar among those approaches in order to find $p(s_t \mid y_{1:t}, \Phi)$. There are several backward smoothing methods and they end up with a marginal density of $p(s_t \mid y_{1:T}, \Phi)$. In this paper, I choose the most common MCMC algorithm in economic literature, Chib (1996)'s algorithm.

The posterior parameters are calculated through G iterations after several burn-ins are removed in order to remove the impact of the initial choice. In particular, I want to compute the parameter, M_1, which is the mean of latent state 1 in the multivariate Markov switching model. It is given by

$$\widehat{M_1} = \frac{1}{G} \sum_{g=1}^{G} M_1^{(g)}$$

where the superscript (g) is the g-th iteration after the burn-ins. The precision of the conditional posterior estimation is better when the sample G is larger.

3.1.1 $\Phi \mid S, Y$ with multivariate Normal distribution

1. $P \mid S$

   For each row in P, $p_{k\cdot}$ has a Dirichlet prior. The likelihood is a discrete distribution. Hence, the conditional posterior density is

   $$p(p_{k\cdot} \mid S) \propto \prod_{j=1}^{K} p_{kj}^{\alpha_{kj} + \sum_{t=2}^{T} \mathbf{1}(s_t = k, s_{t-1} = j) - 1}$$

   that is a kernel of a Dirichlet distribution, $\mathbb{D}ir\left(\alpha_{kj} + \sum_{t=2}^{T} \mathbf{1}(s_t = k, s_{t-1} = j)\right)$.

2. $M, \Sigma \mid S, Y$

   For M, Σ in regime k, it has a multivariate Normal - Inverse Wishart prior with the density function as

$$
\begin{aligned}
p(M_k, \Sigma_k \mid h, \mu, n, \Psi) = & \left(\frac{h}{2\pi}\right)^{N/2} |\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(M_k - \mu)' h \Sigma^{-1} (M_k - \mu)\right\} \\
& \times \frac{|\Psi|^{n/2}}{2^{nN/2} \pi^{N(N-1)/4} \prod_{i=1} \Gamma\left(\frac{n+1-i}{2}\right)} |\Sigma_k|^{-\dfrac{n+N+1}{2}} \\
& \times \exp\left\{-\frac{1}{2} tr(\Psi \Sigma_k^{-1})\right\}
\end{aligned}
$$

   The likelihood is a multivariate normal distribution

$$
p(y_t \mid M_k, \Sigma_k, s_t = k) = \left(\frac{1}{2\pi}\right)^{N/2} |\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(y_t - M_k)' \Sigma_k^{-1} (y_t - M_k)\right\}
$$

   Hence, the conditional posterior density is

$$
p(M_k, \Sigma_k \mid Y, S) \sim \mathbb{N} - \mathbb{IW}(\mu_k^{(p)}, h_k^{(p)}, n_k^{(p)}, \Psi_k^{(p)})
$$

   where

$$
\begin{aligned}
h_k^{(p)} &= h + \sum_{t=1}^{T} \mathbf{1}(s_t = k) \\
\mu_k^{(p)} &= \frac{1}{h_k^{(p)}} \left(h\mu + \sum_{t=1}^{T} \mathbf{1}(s_t = k) y_t\right) \\
n_k^{(p)} &= n + \sum_{t=1}^{T} \mathbf{1}(s_t = k) \\
\Psi_k^{(p)} &= \Psi + \sum_{t=1}^{T} \mathbf{1}(s_t = k) y_t y_t' - h_k^{(p)} \mu_k^{(p)} (\mu_k^{(p)})' + h\mu\mu'
\end{aligned}
$$

   with the superscript (p) denoting the posterior.

3.1.2 $\Phi \mid Y, S$ with multivariate Student t distribution

There is no change to the posterior of $P \mid S$ because the distribution of Y is not involved.

1. $M, \Sigma \mid S, Y, \tau$

   The prior of M and Σ is Normal - Inverse Wishart. I rewrite the likelihood functional form of the multivariate Student t distribution as the multivariate Normal distribution

$$
p(y_t \mid M_{s_t}, \Sigma_{s_t}, \tau_t, s_t) = \left(\frac{\tau_t}{2\pi}\right)^{N/2} |\Sigma_{s_t}|^{-1/2} \exp\left\{-\frac{\tau_t}{2}(y_t - M_{s_t})' \Sigma_{s_t}^{-1} (y_t - M_{s_t})\right\} \tag{4}
$$

Hence, the conditional posterior density is

$$M_k, \Sigma_k \mid Y, S, \tau \sim \mathbb{N} - \mathbb{IW}(\mu_k^{(p)}, h_k^{(p)}, n_k^{(p)}, \Psi_k^{(p)})$$

where

$$\begin{aligned}
h_k^{(p)} &= h + \sum_{t=1}^{T} \mathbf{1}(s_t = k)\tau_{s_t} \\
\mu_k^{(p)} &= \frac{1}{h_k^{(p)}} \left(h\mu + \sum_{t=1}^{T} \mathbf{1}(s_t = k)\tau_t y_t \right) \\
n_k^{(p)} &= n + \sum_{t=1}^{T} \mathbf{1}(s_t = k) \\
\Psi_k^{(p)} &= \Psi + \sum_{t=1}^{T} \mathbf{1}(s_t = k)\tau_t y_t y_t' - h_k^{(p)} \mu_k^{(p)} (\mu_k^{(p)})' + h\mu\mu'
\end{aligned}$$

2. τ

The prior of τ_t is a Gamma distribution as

$$p\left(\tau_t \mid \frac{\nu_{s_t}}{2}, \frac{\nu_{s_t}}{2}\right) = \Gamma\left(\frac{\nu_{s_t}}{2}\right)^{-1} \left(\frac{\nu_{s_t}}{2}\right)^{\frac{\nu_{s_t}}{2}} \tau_t^{\frac{\nu_{s_t}}{2}-1} \exp\left\{-\frac{\nu_{s_t}}{2}\tau_t\right\}$$

The likelihood is a multivariate Normal distribution as in Equation 4. Therefore, the conditional posterior distribution is

$$\begin{aligned}
p(\tau_t \mid M, \Sigma, \nu, y_t, s_t) &\propto \tau_t^{\frac{\nu_{s_t}}{2}-1} \exp\left\{-\frac{\nu_{s_t}}{2}\tau_t\right\} \tau_t^{N/2} \exp\left\{-\frac{\tau_t}{2}(y_t - M_{s_t})' {\Sigma_{s_t}}^{-1} (y_t - M_{s_t})\right\} \\
&= f_G\left(\tau_t \mid \frac{\nu_{s_t}}{2} + \frac{N}{2}, \frac{\nu_{s_t}}{2} + (y_t - M_{s_t})' {\Sigma_{s_t}}^{-1} (y_t - M_{s_t})\right)
\end{aligned}$$

3. ν

There are several approaches to estimating this parameter.

First, I use the variance formula of the Student t distribution to estimate the degrees of freedom that is

$$V_k = \frac{\nu_k}{\nu_k - 2}\Sigma_k$$

In particular, I find the sample variance covariance matrix, $\hat{V}_k$, the scale matrix, Σ_k, and then recover the degrees of freedom. We find ν_k to minimise the distance between two sides. Details are in Appendix B.2.

For each draw of S, I re-calculate the degrees of freedom of each component. Only a value that is greater than 2 is accepted since it permits the existence of variance. Because the degrees of freedom are calibrated conditionally on the other parameters, we denote the MCMC algorithm using this method as $MCMC - CC$.

Second, I put a hierarchical prior on ν. The choice of the prior on ν varies from an exponential distribution(Geweke, 1993), an uniform distribution (Chib et al., 2002) to a gamma distribution (Nakajima and Omori, 2012). Unfortunately, all are not conjugate prior. In this study, I choose a Gamma prior as follows

$$\nu \sim \mathbb{G}(v_\nu, s_\nu)$$

Hence, I use Metropolis-Hasting as rejection method. In particular, I propose a new ν from this distribution

$$\nu \mid \nu' \sim \mathbb{G}\left(\xi, \frac{\xi}{\nu'}\right)$$

with ν' is the value of ν from the last iteration, ξ is the degree of freedom, and ξ/ν' is the scale parameter. The acceptance probability is

$$\left\{1, \frac{p(\nu \mid \tau)p(\nu' \mid \xi, \xi/\nu)}{p(\nu' \mid \tau)p(\nu \mid \xi, \xi/\nu')}\right\}$$

This method is denoted as $MCMC$.

Third, I make an attempt to use exponential conjugacy. Inspired by Christmas and Everson (2010), I use Stirling's approximation to re-write the complete likelihood in order to generate conjugacy that gives a Gamma posterior distribution (see Appendix B.3). In particular, the approximated conditional posterior follows a Gamma distribution

$$\mathbb{G}\left(v_\nu + \frac{\sum_{t=1}^{T} \mathbf{1}(s_t = k)}{2}, s_\nu + \frac{\sum_{t=1}^{T} \mathbf{1}(s_t = k)(\tau_{s_t} - \log \tau_{s_t})}{2} - \frac{\sum_{t=1}^{T} \mathbf{1}(s_t = k)}{2}\right)$$

Because we have the approximation of the complete likelihood, we denote this method as $MCMC-A$. This approximation is purely for a comparison to Christmas and Everson (2010)'s VI in the literature. However, in Appendix B.3, I show that this approximation cannot work well in the case of small degrees of freedom, which is the case of interest.

3.1.3 $S \mid Y, \Phi$

Since S follows a Dirichlet distribution and the likelihood follows a multivariate Normal distribution, there exists an exponential conjugacy. The conditional posterior then belongs to the exponential family. The complete likelihood of S and Y is

$$p(S_{1:T}, y_{1:T} \mid \Phi) = p(s_1)p(y_1 \mid s_1, \Phi) \prod_{t=2}^{T} p(s_t \mid s_{t-1}, P)p(y_t \mid s_t, \Phi) \tag{5}$$

where

$$p(s_1) \propto const; \qquad p(s_t \mid s_{t-1}, P) \propto \prod_{j=1}^{K} \prod_{k=1}^{K} p_{jk}^{1(s_t=k, s_{t-1}=j)};$$

$$p(y_t \mid s_t, \Phi) \propto \prod_{k=1}^{K} p(y_t \mid M_k, \Sigma_k)^{\mathbf{1}(s_t=k)}$$

The forward filtering of states at time t from 1 to T is calculated as

$$F(s_t) = p(s_t \mid y_{1:t}, \Phi) = \frac{p(y_t \mid s_t, M, \Sigma)p(s_t \mid y_{1:t-1}, \Phi)}{\sum_{s_t} p(y_t \mid s_t, M, \Sigma)p(s_t \mid y_{1:t-1}, \Phi)}$$

with $p(s_t \mid y_{1:t-1}, \Phi) = \sum\limits_{s_{t-1}} F(s_{t-1})p(s_t \mid s_{t-1}, P)$. The derivation for this filtered pass can be found in Appendix B.4.1.

Then I apply the backward smoother proposed by Chib (1996) to find

$$p(s_t \mid y_{1:T}, \Phi) = \sum_{s_{t+1}} p(s_t, s_{t+1} \mid y_{1:T}, \Phi) \tag{6}$$

from time $T-1$ to time 1. I start with the joint marginal density of s_t and s_{t+1}.

$$p(s_t, s_{t+1} \mid y_{1:T}, \Phi) = p(s_{t+1} \mid y_{1:T}, \Phi)\frac{p(s_{t+1} \mid s_t, \Phi)F(s_t)}{\sum_{s_t} p(s_{t+1} \mid s_t, \Phi)F(s_t)}$$

Hence, I deduce that

$$p(s_t \mid y_{1:T}, \Phi) = \sum_{s_{t+1}} p(s_{t+1} \mid y_{1:T}, \Phi)\frac{p(s_{t+1} \mid s_t, \Phi)F(s_t)}{\sum_{s_t} p(s_{t+1} \mid s_t, \Phi)F(s_t)} \tag{7}$$

Details are discussed in Appendix B.4.2.

3.1.4 Identification restrictions

Identification restrictions are important to label the states. In my paper, these restrictions are on the mean of the states. I can impose either multiple restrictions or one restriction. In this subsection, I employ Botev (2017)'s algorithm to find the conditional expectation with respect to q of the restricted parameter. This algorithm helps simulate the truncated Normal distribution in high dimensions. Hence, I find a tractable solution for VI when applying identification restrictions.

The identification of two states, bull and bear, requires restrictions. In a simple case, the prior distribution for M_k should be

$$p(M_1 \mid \Sigma_1) = f_N(M_1 \mid \mu, h, \Sigma_1)I(\iota' M_1 < 0) \tag{8}$$

$$p(M_2 \mid \Sigma_2) = f_N(M_2 \mid \mu, h, \Sigma_2)I(\iota' M_2 > 0) \tag{9}$$

The change in their functional form is in the normalising constant and the support. Therefore, these restrictions enter the conditional posterior distributions as follows:

$$p(M_1^{(p)} \mid \Sigma_1^{(p)}, \cdot) = f_N(M_1^{(p)} \mid \mu_1^{(p)}, h_1^{(p)}, \Sigma_1^{(p)})I(\iota' M_1^{(p)} < 0)$$
$$p(M_2^{(p)} \mid \Sigma_2^{(p)}, \cdot) = f_N(M_2^{(p)} \mid \mu_2^{(p)}, h_2^{(p)}, \Sigma_2^{(p)})I(\iota' M_2^{(p)} > 0)$$

where f_N is the probability density function of a Normal distribution.

Similarly, the conditional posterior distributions in the Student t distribution are

$$p(M_1^{(p)} \mid \Sigma_1^{(p)}, \tau_1^{(p)}, \cdot) = f_N(M_1^{(p)} \mid \mu_1^{(p)}, h_1^{(p)}, \Sigma_1^{(p)}, \tau_1^{(p)}) I(\iota' M_1^{(p)} < 0)$$
$$p(M_2^{(p)} \mid \Sigma_2^{(p)}, \tau_2^{(p)}, \cdot) = f_N(M_2^{(p)} \mid \mu_2^{(p)}, h_2^{(p)}, \Sigma_2^{(p)}, \tau_2^{(p)}) I(\iota' M_2^{(p)} > 0)$$

with the priors of

$$p(M_1 \mid \Sigma_1, \tau_1, \cdot) = f_N(M_1 \mid \mu_1, h_1, \Sigma_1, \tau_1) I(\iota' M_1 < 0) \tag{10}$$

$$p(M_2 \mid \Sigma_2, \tau_2, \cdot) = f_N(M_2 \mid \mu_2, h_2, \Sigma_2, \tau_2,) I(\iota' M_2 > 0) \tag{11}$$

In the MCMC algorithm, I sample M for each iteration under these restrictions. Sampling from these conditional posterior distributions is straightforward. One naive method is to keep drawing samples until restrictions are satisfied. However, as the dimension of the data increases, this restricted problem becomes more complex and may cause difficulties in finding a correct sample. In this paper, I choose the solution from Botev (2017)'s algorithm to achieve a valid draw faster.

Sampling is necessary when the closed form solution is not available. In this paper, I need to have the expected mean and variance of the truncated multivariate distribution. I use Botev (2017)'s algorithm to sampling these values. This technique is an extension of the separation of variables (SOV) method. The separation-of-variables (SOV) reparameterizes each element x_i conditionally on the values of the preceding terms $x_1, \ldots, x_{i-1}$ using the LQ transformation. After the transformation, the cumulative density probability of the multivariate normal distribution is under the transformed region. The innermost integral of this probability depends on all variables, while the outermost integral depends on x_1. The computation of this probability is crucial. Botev (2017) suggests using the exponentially tilted technique in simulation to compute reliably this probability by using the tilted version of each integral.

3.2 Variational Inference

A review of the VI method can be found in Blei et al. (2017). The idea of VI is to minimise the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the complete likelihood of the model, $p(\Phi, S, Y)$ and a variational distribution, $q(\Phi, S)$. Instead of minimising this divergence directly, I solve an equivalent problem that is to maximise the evidence lower bound ($\mathcal{L}$) where

$$\mathcal{L} = E[\log p(\Phi, S, Y)] - E[\log q(\Phi, S)] = \log p(Y) - KL[q(\Phi, S)||p(\Phi, S \mid Y)]$$

where E implies $E_{q(\Phi,S)}$. Since $\log p(Y)$ is a constant, minimising KL divergence is the same as maximising the evidence lower bound. The problem now is

$$q^*(\Phi, S) = \arg\max_{q \in Q} E[\log p(\Phi, S, Y)] - E[\log q(\Phi, S)]$$

The common choice of the family, Q, is a mean field family. This family simplifies the optimisation problem by assuming an independent structure among factors. In my estimation, I also choose a mean field distribution as

$$q(\Phi, S) = \prod_{k=1}^{K} q(M_k, \Sigma_k) \prod_{k=1}^{K} q(p_{k\cdot}) q(S_{1:T})$$

for the Normal distribution, and

$$q(\Phi, S) = \prod_{k=1}^{K} q(M_k, \Sigma_k) \prod_{k=1}^{K} \left[\prod_{t=1}^{T} q(\tau_{tk}) \right] q(\nu_k) \prod_{k=1}^{K} q(p_{k\cdot}) q(S_{1:T})$$

for the Student t distribution.

The evidence lower bound ($\mathcal{L}$) is

$$\mathcal{L} = E[\log p(\Phi)] + E[\log p(S \mid Y, \Phi)] + E[\log p(Y \mid \Phi)] - E[\log q(\Phi)] - E[\log q(S)]$$

For any factor in the mean field family, I solve for the optimal variational density by taking the conditional evidence lower bound that solves variational parameters conditional on the others. For example, with Φ, I have the conditional evidence lower bound

$$\mathcal{L}^* = E[\log p(\Phi, S, Y)] - E[\log q(\Phi)] = E_{q(\Phi)} \left(E_{q(S)}[\log p(\Phi, S, Y)] - E_{q(S)}[\log q(\Phi)] \right)$$

The optimal solution is then

$$q^*(\Phi) = \exp\left\{ E_{q(S)}[\log p(\Phi \mid S, Y)] \right\} \tag{12}$$

A detailed explanation can be found in Ormerod and Wand (2010). In the next steps, I find the optimal solution for each parameter.

The result from VI is the optimal variational density of the model parameter. If an analytical solution for the conditional expectation of a variational parameter is available, it is used to update the other variational parameters. Otherwise, I must find a numerical approximation by sampling.

Variational algorithms to approximate the Markov switching model have been developed in Beal (2003), Ghahramani and Hinton (2000), Gruhl and Sick (2016). However, these works are mainly based on the Baum-Welch algorithm, which is not familiar in economic literature. The Baum-Welch algorithm belongs to the class of expectation maximisation algorithm. This iterative algorithm increases the conditional posterior until the maximum is reached. At each iteration, we find the forward and backward probabilities as follows:

$$F_t = p(Y_{1:t}, s_t \mid \theta) \qquad\qquad B_t = p(Y_{t+1:T} \mid s_t, \theta)$$

Then, the marginal of each state s_t can be found by the Bayesian rule

$$p(s_t \mid Y_{1:T}, \theta) = \frac{F_t B_t}{\sum_{s_t} F_t B_t}$$

The Chib (1996)'s algorithm also belongs to the class of the expectation maximisation algorithm. At each iteration, we have the forward probabilities using the Bayesian rule from time 1 to time T:

$$F_t = p(Y_{1:t} \mid s_t)$$

Then at time T, we move backward to time 1:

$$p(s_t \mid Y_{1:T}, \theta) = \sum_{s_{t+1}} p(s_t, s_{t+1} \mid Y_{1:T}, \theta) \tag{13}$$

Both methods gives the posterior marginal of the latent state via a recursive process. However, The former is popular in computer science and engineering, while the latter is more common in economics (Frühwirth-Schnatter, 2006). In this paper, I propose a novel algorithm to multivariate dimension with the idea from Chib (1996)'s MCMC algorithm that is common in Bayesian econometric literature.

3.2.1 $q(\Phi)$ with multivariate Normal distribution

Under the multivariate normal distribution, the conditional evidence lower bound is

$$\begin{aligned}\mathcal{L}^* = E&\left[\log p(M, \Sigma) + \log p(P) + \log p(S \mid Y, P) + \sum_{t=1}^{T} \log p(y_t \mid M, \Sigma, P, S)\right] \\ &- E[\log q(M, \Sigma) + \log q(P) + \log q(S)]\end{aligned}$$

Details can be found in Appendix C.

1. P

   Note that $\log p(s_1 = k)$ is a constant with respect to P. I can ignore this term and start from $t = 2$. The optimal solution for $q(p_{k\cdot})$ is

   $$q^*(p_k) \sim \mathbb{D}ir(\overline{\alpha}_k)$$

   where $\overline{\alpha}_{kj} = \alpha_{kj} + \sum_{t=2}^{T} \phi_{t-1,k,t,j}$ with

   $$E_{q(S)}[\mathbf{1}(s_{t-1} = k, s_t = j)] = \phi_{t-1,k,t,j}$$

   as the conditional expectation of the transition probability of state k at time $t - 1$ to state j at time t.

2. M, Σ

   The optimal solution is then

   $$q^*(M_k, \Sigma_k) \propto \exp\left\{E_{q(-M,\Sigma)}\left[\log p(M_k, \Sigma_k) + \sum_{t=1}^{T} \log p(y_t \mid S, M, \Sigma)\right]\right\}$$

Let denote

$$\begin{aligned}
\overline{h}_k &= h + \sum_{t=1}^{T} \phi_{tk} \\
\overline{\mu}_k &= \overline{h}_k^{-1} \left(h\mu + \sum_{t=1}^{T} \phi_{tk} y_t \right) \\
\overline{n}_k &= n + \sum_{t=1}^{T} \phi_{tk} \\
\overline{\Psi}_k &= \Psi + \sum_{t=1}^{T} \phi_{tk} y_t y_t' - \overline{h}_k \overline{\mu}_k \overline{\mu}_k' + h\mu\mu'
\end{aligned}$$

Then I find that the variational density of $q^*(M_k, \Sigma_k)$ is a Normal - Inverse Wishart distribution $\mathbb{N} - \mathbb{IW}(\overline{\mu}_k, \overline{h}_k, \overline{n}_k, \overline{\Psi}_k)$.

3.2.2 $q(\Phi)$ with multivariate Student t distribution

The evidence lower bound is

$$\begin{aligned}
\mathcal{L} = E[&\log p(M, \Sigma) + \log p(\nu) + \sum_{t=1}^{T} \log p(\tau_t \mid \nu) + \log p(P) + \log p(S \mid Y, P) \\
&+ \sum_{t=1}^{T} \log p(y_t \mid M, \Sigma, \tau, P, S)] \\
&- E[\log q(M, \Sigma) + \log q(\nu) + \sum_{t=1}^{T} \log q(\tau_t) + \log q(P) + \log q(S)]
\end{aligned}$$

Conditional expectations can be found in Appendix E.2.

1. P

   This part is similar to the multivariate Normal distribution because the new terms ν and τ are not involved in estimating $q(P)$.

2. M, Σ

   The optimal solution is

$$q^*(M_k, \Sigma_k) \propto \exp\left[E_{q(S),q(\tau)}[\log p(M_k, \Sigma_k) + \sum_{t=1}^{T} \log p(y_t \mid s_t = k, M_k, \Sigma_k, \tau_t)] \right]$$

   Let denote

$$\overline{h}_k = h + \sum_{t=1}^{T} \phi_{tk} E\left[\tau_t\right]$$

$$\overline{m}_k = \overline{h}^{-1}\left(h\mu + \sum_{t=1}^{T} \phi_{tk} E\left[\tau_t\right] y_t\right)$$

$$\overline{n}_k = n + \sum_{t=1}^{T} \phi_{tk}$$

$$\overline{\Psi}_k = \Psi + \sum_{t=1}^{T} \phi_{tk} E\left[\tau_t\right] y_t y_t' - \overline{h}\overline{\mu}_k\overline{\mu}_k' + h\mu\mu'$$

I find that the variational density of $q^*(M_k, \Sigma_k)$ is a Normal Inverse Wishart distribution $\mathbb{N} - \mathbb{IW}(\overline{\mu}_k, \overline{h}_k, \overline{n}_k, \overline{\Psi}_k)$.

3. τ

   The optimal solution is

$$q^*(\tau_t) \propto \exp\left[E[\log p(\tau_t \mid \nu_{s_t}, s_t) + \log p(y_t \mid M_{s_t}, \Sigma_{s_t}, \tau_t)]\right]$$

   I find that this functional form implies

$$q^*(\tau_t) \sim G\left(\frac{E(\nu_{s_t})}{2} + \frac{N}{2}, \frac{E(\nu_{s_t})}{2} + \frac{1}{2}E\left[(y_t - M_{s_t})' \Sigma_{s_t}^{-1} (y_t - M_{s_t})\right]\right)$$

4. ν

   The optimal solution is

$$q^*(\nu_k) \propto \exp\left[E\left(\log p(\nu_k) + \sum_{t=1}^{T} \log p(\tau_t \mid \nu_k, s_t = k)\right)\right]$$

   The difficulty is that at this point, because there is no exponential conjugacy in this solution, I cannot update the variational parameters straightforwardly. Unlike the MCMC algorithm, there is no room for the Metropolis-Hastings rejection method.

   I introduce three approaches to find degrees of freedom.

   - Conditional calibration

     This is a comparable version to MCMC-CC. I estimate sample degrees of freedom conditional on other parameters. I use the variance formula to recover the degrees of freedom. This approach can be denoted as VI-CC because it is equivalent to MCMC-CC.

   - Approximation

     This approach is proposed by Christmas and Everson (2010). I use Stirling's approximation for $\log\Gamma\left(\frac{\nu_k}{2}\right)$ in order to have my proposed prior as a conjugate prior. Therefore, the optimal variational density is

$$\nu_k \sim \mathbb{G}\left(v_\nu + \frac{\sum_{t=1}^{T}\phi_{tk}}{2}, s_\nu + \frac{\sum_{t=1}^{T}\phi_{tk} E\left(\tau_t - \log\tau_t\right)}{2} - \frac{\sum_{t=1}^{T}\phi_{tk}}{2}\right)$$

     Let denote this approach as VI-A that is comparable to MCMC-A.

- Proposing a Gamma distribution of $q(\nu)$

  I propose $q^*(\nu) \sim \mathbb{G}(\overline{v}_\nu, \overline{s}_\nu)$. We find $\overline{v}_\nu$ and $\overline{s}_\nu$ that minimise the distance

  $$E[\log q(\nu_k)] - E\left[\log p(\nu_k) + \sum_{t=1}^{T} \log p(\tau_t \mid \nu_k, s_t = k)\right]$$

  with the term inside the bracket as the complete likelihood of v_k.

  Since I assume the Student t distribution has the second moment, its degrees of freedom must be greater than 2. Only the result satisfying this condition will be kept. It is not necessary to update ν every iteration. I can update other parameters for 10 iterations, then update hyperparameters for ν. It will save computational costs. I denote this approach as VI.

3.2.3 $q(S)$

The conditional evidence lower bound on S is

$$\mathcal{L}^* = E[\log p(S, Y \mid \Phi)] - E[\log q(S)]$$

The optimal solution is as

$$\begin{aligned} q^*(S_{1:T}) &= \exp\left\{E_{q(\Phi)}\left[\log p(S_{1:T}, y_{1:T} \mid \Phi) - \log p(y_{1:T} \mid \Phi)\right]\right\} \\ &\propto \exp\left\{E_{q(\Phi)}\left[\log p(S_{1:T}, y_{1:T} \mid \Phi)\right]\right\} \end{aligned}$$

I make a comparison between the complete log-likelihood of the conditional posterior in Equation 5 with the complete log-likelihood of the variational density, $q^*(S_{1:T})$. First, the log of the complete likelihood in Equation 5 is proportional to

$$\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{1}(s_t = k, s_{t-1} = j) \log p_{jk} + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{1}(s_t = k) \log p(y_t \mid \Phi, s_t)$$

Second, the log-likelihood of the optimal solution from the VI where I take the conditional expectation with respect to Φ is

$$\begin{aligned} &E_{q(\Phi)}\left[\sum_{t=2}^{T} \log p(s_t \mid s_{t-1}, P)\right] + E_{q(\Phi)}\left[\sum_{t=1}^{T} \log p(y_t \mid s_t, \Phi)\right] \\ &= \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{1}(s_{t-1} = k, s_t = j) E_{q(\Phi)}[\log p_{kj}] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{1}(s_t = k) E_{q(\Phi)}[\log p(y_t \mid \Phi, s_t)] \end{aligned}$$

The log of the complete likelihood in Equation 5 is similar to the loglikelihood of the optimal solution. It suggests that the optimal solution shares the same type of density with the conditional posterior. However, their parameters are not the same. For example, the parameters for the transition matrix in the conditional posterior density are p_{kj} and in the optimal solution through VI for $q(p_{kj})$ is $\exp(E_{q(p_{kj})}[\log p_{kj}])$. Let denote

$$\widetilde{p}_{kj} = \exp(E_{q(\Phi)}[\log p_{kj}]) \quad \widetilde{p}(y_t \mid \Phi, s_t) = \exp(E_{q(\Phi)}[\log p(y_t \mid \Phi)])$$

with the first equality involving normalisation. These newly defined parameters are already found in the previous steps. Hence, I can find $q(s_t)$ in the same way I find $p(s_t \mid y_{1:T}, \Phi)$, i.e., through a forward filtering backward smoothing algorithm.

In order to find the marginal variational density of each s_t, $q^*(s_t)$, I note that the structure of this variational distribution is similar to the conditional posterior distribution, $p(S \mid Y, \Phi)$. Therefore, I can write that

$$\begin{aligned} q^*(S_{1:T}) &\propto \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} \widetilde{p}_{jk}^{\mathbf{1}(s_t=k, s_{t-1}=j)} \prod_{t=1}^{T} \prod_{k=1}^{K} \widetilde{p}(y_t \mid \Phi, s_t)^{1(s_t=k)} \\ &\propto \widetilde{p}(y_1 \mid s_1, \Phi) \prod_{t=2}^{T} \widetilde{p}(s_t \mid s_{t-1}, P) \widetilde{p}(y_t \mid s_t, \Phi) \end{aligned}$$

This expression is similar to Equation 5, the complete likelihood of the conditional posterior distribution of S. Because the conditional posterior distribution, $p(s_t \mid Y, \Phi)$, is estimated through the forward-backward algorithm, I can do the same to find the variational density, $q^*(s_t)$, at each t.

I use the same Chib (1996)'s algorithm as in MCMC. For any t, the variational forward filtering is

$$\widetilde{F}(s_t) \propto \frac{\widetilde{p}(y_t \mid s_t, \Phi) \sum_{s_{t-1}} \widetilde{p}(s_t \mid s_{t-1}, P) \widetilde{F}(s_{t-1})}{\sum\limits_{s_t} \widetilde{p}(y_t \mid s_t, \Phi) \sum_{s_{t-1}} \widetilde{p}(s_t \mid s_{t-1}, P) \widetilde{F}(s_{t-1})}$$

with the normalising constant $\widetilde{p}(y_t \mid y_{1:t-1}, M, \Sigma)$.

The forward pass at time $t = 1$ for this optimal solution is

$$\widetilde{F}(s_1) = \frac{\widetilde{p}(s_1) \widetilde{p}(y_1 \mid s_1, \Phi)}{\sum_{s_1} \widetilde{p}(s_1) \widetilde{p}(y_1 \mid s_1, \Phi)}$$

with the normalising constant $\widetilde{p}(y_1 \mid \Phi)$.

I follow the backward algorithm in Equation 15 to smooth the variational density of states as follows

$$q^*(s_t) \propto \widetilde{p}(s_t \mid y_{1:T}, \Phi) = \sum_{s_{t+1}} \widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi) \widetilde{p}(s_{t+1} \mid y_{1:T}, \Phi)$$

with the first term

$$\widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi) = \frac{\widetilde{p}(s_{t+1} \mid s_t, \Phi) \widetilde{p}(s_t \mid y_{1:t}, \Phi)}{\sum_{s_t} \widetilde{p}(s_{t+1} \mid s_t, \Phi) \widetilde{p}(s_t \mid y_{1:t}, \Phi)}$$

Similarly, the temporal transition between two states is

$$q^*(s_t, s_{t+1}) \propto \widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi) \widetilde{p}(s_{t+1} \mid y_{1:T}, \Phi)$$

The likelihood of the data is

$$\exp(E_{q(\Phi)}[\log p(Y \mid \Phi)]) = \widetilde{p}(y_{1:T} \mid \Phi) = \widetilde{p}(y_1 \mid \Phi) \prod_{t=2}^{T} \widetilde{p}(y_t \mid y_{1:t-1}, \Phi) \tag{14}$$

On the right-hand side, each factor in this product is a normalising constant of its corresponding forward pass. Therefore, I store all these constants to calculate the log likelihood of the ELBO.

This approach can be applied to the multivariate Student t distribution. The reason is that the functional form of the Student t distribution can be rewritten as a Normal distribution, a member of the exponential family. Because a distribution of the exponential family can be rewritten with an exponential function, the log in Equation 14 only removes this exponential function. Hence, the expectation and the exponential function in Equation 14 enter the likelihood without changing its kernel density. If there is no exponential conjugacy, this variation-backward smoothing algorithm fails to work.

It is important to note that I must maintain numerical stability. Errors may appear during the execution of the algorithm. The reason lies in the use of likelihood and joint likelihood. These values are extremely small and easily out of range in any computing programme. Hence, I take note of these likelihoods and their relevant terms to avoid inaccurate approximations, as in Appendix F. In addition, my algorithm will keep the log values until the last step before reverting them by an exponential function.

3.2.4 Identification restrictions

Previously developed VI algorithms do not emphasise applications with restrictions. However, restrictions are significant in identifying the bull and bear states in economic literature. For example, a sign restriction on the sum of the mean in each state The bear state should have a negative mean return, while the bull state should show a positive one. My VI algorithm must address this identification problem.

I address this identification problem by including restrictions in the VI algorithm. Those restrictions enter directly into the ELBO optimisation problem via the truncated prior in 9 and 11. Since the distribution of M is truncated by this restriction, I must find the expectation related to M under this truncation. Because there is no closed-form solution for the expectation of a truncated distribution, I find it through sampling.

Among all parameters, only M is related to the truncated distribution. Hence, the restricted VI algorithm is similar to the above ones except for the parts that have M. In particular, I re-evaluate the lower bounds of M and S.

The optimal variational density for M is

$$q(M_1 \mid \Sigma_1, \cdot) = f_N(M_1 \mid \overline{\mu}_1, \overline{h}_1, \Sigma_1) I\left(\iota' M_1 < 0\right) \qquad q(M_2 \mid \Sigma_2 \cdot) = f_N(M_2 \mid \overline{\mu}_2, \overline{h}_2, \Sigma_2) I\left(\iota' M_2 > 0\right)$$

Under this optimal solution, the expectation with respect to $q(M)$ must be re-evaluated. Details are in Appendix E.2.

Regarding S, I do not need to revise the previous solution for the optimal density. However, I need to use the new expectations relating to the truncated distribution. These expected values are similar to those I find in the case of M.

3.2.5 Monitoring ELBO

The convergence of both restricted and unconstrained algorithms is similar. However, since there is a sampling step, the numerical values calculated should result in an erratic lower bound. Therefore, I monitor the convergence by observing several iterations (defined as Ω iterations) as in Tran et al. (2017). Instead of observing each iteration, I observe the convergence through the mean of the lower bounds from these iterations.

The majority of the expected values in the evidence lower bound are available directly (see Appendix D) except for the expectation of the log likelihood of the data with respect to the variational distribution q. Note that this is the normalising constant in the forward filtering step. Hence, I have

$$E_{q(\Phi,S)}[\log q(S)] = E_{q(S)}[E_{q(\Phi)}[\log p(S, Y \mid \Phi)] - E_{q(\Phi)}[\log p(Y \mid \Phi)]]$$

So the evidence lower bound is

$$\begin{aligned}\mathcal{L} &= E_q[\log p(\Phi)] - E_q[\log q(\Phi)] + E_q[\log p(Y \mid \Phi)] \\ &= E_q[\log p(P)] - E_q[\log q(P)] + E_q[\log p(M, \Sigma)] - E_q[\log q(M, \Sigma)] + E_q[\log p(Y \mid \Phi)]\end{aligned}$$

The last term on the right hand side is available in Equation 14.

Another approach to monitoring the convergence of this algorithm is through the loglikelihood of the data, $E_q[\log p(Y \mid \Phi)]$. This is also available through my algorithm.

3.3 Forecasting

Suppose there are h steps ahead. I start with one step ahead. The conditional predictive density is as follows

$$p(y_{t+1} \mid y_{1:t}, \Phi) = \sum_{s_{t+1}} p(y_{t+1} \mid y_{1:t}, s_{t+1}, \Phi) \left(\sum_{s_t} p(s_{t+1} \mid s_t, \Phi) p(s_t \mid y_{1:t}, \Phi) \right)$$

This functional form implies a mixture of multivariate normal distributions with the weight being the probability $p(s_{t+1} \mid y_{1:t}, \Phi)$. This weight is computed by the transition matrix and the current state at time t. Note that, although there is a case of Student t distribution, it can be rewritten as a Normal distribution. Therefore, I have the same way of forecasting.

I find the unconditional predictive density by integrating out the model parameters Φ. Suppose there is a sample of G draws from the conditional posterior distribution $p(\Phi \mid y_{1:t})$. The unconditional predictive density

$$p(y_{t+1} \mid y_{1:t}) = \int p(y_{t+1} \mid y_{1:t}, \Phi) p(\Phi \mid y_{1:t}) d\Phi$$

is calculated unbiasedly from

$$p(y_{t+1} \mid y_{1:t}) \approx \frac{1}{G} \sum_{g=1}^{G} p(y_{t+1} \mid y_{1:t}, \Phi^{(g)})$$

The h-step ahead forecast joint density is the multiplication of multiple 1-step-ahead forecast densities as

$$p(y_{t+1:t+h} \mid y_{1:t}) = \prod_{i=1}^{h} p(y_{t+i} \mid y_{1:t+i-1})$$

Each predictive density on the right-hand side is computed separately. The model that is better at forecasting will have a higher value of predictive likelihood.

It is possible to derive the conditional expected forecasting mean and variance as

$$\begin{aligned}
\widehat{y}_{t+h} &= E(y_{t+h} \mid y_{1:t}, \Phi) \\
&= \sum_{k=1}^{K} \sum_{s_{t+h}} \cdots \sum_{s_t} E(y_{t+h} \mid s_{t+h}, \Phi) p(s_{t+h} \mid s_{t+h-1}, \Phi) \cdots p(s_{t+1} \mid s_t, \Phi) p(s_t \mid y_{1:t}, \Phi) \\
\widehat{V}_{t+h} &= Var(y_{t+h} \mid y_{1:t}, \Phi) \\
&= \sum_{k=1}^{K} \sum_{s_{t+h}} \cdots \sum_{s_t} E(y_{t+h} y'_{t+h} \mid s_{t+h}, \Phi) p(s_{t+h} \mid s_{t+h-1}, \Phi) \cdots p(s_{t+1} \mid s_t, \Phi) p(s_t \mid y_{1:t}, \Phi) \\
&- \left(\sum_{k=1}^{K} \sum_{s_{t+h}} \cdots \sum_{s_t} E(y_{t+h} \mid s_{t+h}, \Phi) p(s_{t+h} \mid s_{t+h-1}, \Phi) \cdots p(s_{t+1} \mid s_t, \Phi) p(s_t \mid y_{1:t}, \Phi) \right) \\
&\left(\sum_{k=1}^{K} \sum_{s_{t+h}} \cdots \sum_{s_t} E(y_{t+h} \mid s_{t+h}, \Phi) p(s_{t+h} \mid s_{t+h-1}, \Phi) \cdots p(s_{t+1} \mid s_t, \Phi) p(s_t \mid y_{1:t}, \Phi) \right)'
\end{aligned}$$

where

$$E(y_{t+h} y'_{t+h} \mid s_{t+h}, \Phi) = var(y_{t+h} \mid s_{t+h}, \Phi) + E(y_{t+h} \mid s_{t+h}, \Phi) E(y_{t+h} \mid s_{t+h}, \Phi)'$$

With the MCMC algorithm, I can apply the forecasting formula immediately. Regarding the VI method, the suitable probability density functions are the variational densities $q^*(P), q^*(S)$ and $q^*(M, \Sigma)$ and the "tilde" probabilities, for example, we replace $p(s_t \mid y_{1:t}, \Phi)$ with $q^*(s_t)$, $p(s_{t+1} \mid s_t, \Phi)$ with $\widetilde{p}(s_{t+1} \mid s_t, \Phi)$. Regarding the multivariate Student t distribution, I extend it to $q^*(\tau)$ and $q^*(\nu)$. Although I do not have the sample of G drawn directly from the conditional posterior, the variational density of P in the last iteration is suitable to find the expected value. Each conditional predictive density is computed separately.

4 Algorithm comparison

I compare the performance of two algorithms in terms of how fitted their results are to in-sample data and how close their point forecasting value is to the out-of-sample data. I use the computing system SPARTAN with an Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz running MATLAB 2021b. Regarding the out-of-sample performance, I also measure the economic performance of investment strategies formed by two algorithms. In addition, I compare the amount of time that needs to be consumed to achieve the final result for two algorithms.

I carry out multiple simulation studies. These simulation results highlight that the VI algorithm may lead to similar inferences compared to the MCMC algorithm. In addition, the forecasting performance of the VI is not necessarily worse. This paper is not the first to observe it. It may surpass the prediction from MCMC instead. The finding is consistent with Frazier et al. (2022, 2021).

In conclusion, I find the VI algorithm provides fast and accurate inference and comparable forecasting performance. Its results are close to or identical to the results from the MCMC algorithm. However, the VI consumes less time.

4.1 Simulation

In this section, I perform several simulations in order to compare the in-sample and out-of-sample performance between algorithms. I have two sets of simulated data: one with two assets and the other with 30 assets. The focus is on the novel approach with identification restrictions.

4.1.1 A bivariate example

I conduct a simulation to measure the performance of new algorithms. First, let denote a data of $Y = y_{it}$ where $i = 1 : N; t = 1 : T$. Suppose there are only two states $k = 1, 2$. The data generating process is as follows

$$y_{it} \sim \begin{cases} \mathbb{N}(M^{(1)}, \Sigma^{(1)}) & \text{if } 1 \leq t < 300 \\ \mathbb{N}(M^{(2)}, \Sigma^{(2)}) & \text{if } 300 \leq t < 600 \\ \mathbb{N}(M^{(1)}, \Sigma^{(1)}) & \text{if } 600 \leq t \leq 1000 \end{cases}$$

We expect our algorithms to be able to identify the latent states, bear and bull, under restrictions. The true parameters under restrictions are set as

$$M^1 \sim \mathbb{N}([-0.5 \quad -0.5]', \Sigma^1) I(\iota' M^1 < 0) \qquad\qquad M^2 \sim \mathbb{N}([1 \quad 2]', \Sigma^2) I(\iota' M^2 > 0)$$

A similar set up is generated for the Student t distribution. The corresponding distributions are

$$\mathbb{T}(M^{(1)}, \Sigma^{(1)}, \nu^{(1)}) \qquad\qquad \mathbb{T}(M^{(2)}, \Sigma^{(2)}, \nu^{(2)})$$

I start with a simple example with $N = 2$, i.e., 2 asset series. In order to test the performance of the new VI algorithm, I carry out this simulation under a comparable MCMC algorithm that performs with the same set of priors and data. For example: the MCMC-A algorithm is comparable to the VI-A algorithm. For the VI algorithm, the tolerance level is 10^{-16}, $\Omega = 20$, $n = 0.5$. Regarding the MCMC algorithm, the total iterations for the MCMC algorithm are 2,000 with 1,000 burn-ins.

Table 1: Estimated parameters with 2 assets under Normal distribution

	Unrestricted				Restricted					
Parameters	MCMC		VI		MCMC		VI		True	
Mean of the bear state ($\widehat{M}^1$)	-0.01	1.00	-0.01	1.00	-0.49	0.49	-0.50	0.50	-0.50	0.50
Mean of the bull state ($\widehat{M}^2$)	1.04	0.96	1.04	0.96	1.04	0.96	1.04	0.96	1	1
Variance of the bear state ($\widehat{V}^1$)	0.95	0.04	0.95	0.04	0.94	0.04	0.95	0.04	1	0
	0.04	1.05	0.04	1.04	0.04	1.05	0.04	1.05	0	1
Variance of the bull state ($\widehat{V}^2$)	5.33	3.37	5.32	3.36	5.30	3.34	5.30	3.35	5	3
	3.37	5.22	3.37	5.22	3.35	5.18	3.35	5.20	3	5

Notes: Table 1 contains the posterior mean of parameters and the optimal variational parameters of the multivariate Markov switching model estimated by MCMC algorithm and VI algorithm, respectively. This table includes both results with and without restrictions. The true parameters are also provided.

The restricted model can identify the bull and bear states, while the unrestricted version cannot. Table 1 shows the estimated results from the normal data. Traditional MCMC and VI algorithms find no bear state. In particular, all states show a positive mean of returns. Only when we control for the restrictions can the algorithm detect the bear state. The same insight can be found with the student t distribution as in Table 12. I plot these estimated densities against the true densities of both states with normal data in Figure 6. Estimated densities overlap, and they are also quite near the true ones. However, the unrestricted results are far from the true ones. To conclude, I find that the new algorithm can adapt restrictions that fulfil the purpose of state identification.

Estimates of conditional probabilities are the same between two algorithms. It is presented visually as in Figure 2, 3 and 7. In addition, the correlation between two estimated posterior probabilities is almost 1 as in Table 2.

Tables 1 and 12 present parameter estimates. The estimated mean and variance/scale of two methods are similar and close to the true value. Notably, none of the estimates of the degrees of freedom are exact. However, these estimates can reflect a distribution with thicker tails. Overall, this implies that the results from one run of the VI algorithm are similar to the ones from MCMC.

Table 2: Correlation between conditional posterior probabilities and the true probabilities

Number of assets	2					30			
Distribution	Normal		Student t			Normal	Student t		
Correlation	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000

Notes: Table 2 contains the correlation between the mean of the conditional posterior of probabilities and the optimal variational probabilities in the multivariate Markov switching model estimated by the MCMC algorithm and the VI algorithm respectively. With 2 assets under Normal distribution, we have the correlation from left to right as without and with restrictions. With 30 assets under Normal distribution, we have a restricted result. Under the Student t distribution, from left to right, we have the correlation between results from MCMC-CC vs. VI-CC, MCMC vs. VI, and MCMC-A vs. VI-A, respectively.

Table 3: Computation time (in seconds)

	2 Assets			30 Assets		
	VI	MCMC	Proportion	VI	MCMC	Proportion
'Normal'	7.749	102.46	7.56%	12.190	1195	1.02%
'Student t (CC)'	21.919	243.80	8.99%	24.905	1259	1.98%
'Student t (A)'	25.077	260.97	9.61%	35.866	1384	2.59%
'Student t'	39.264	255.35	15.38%	55.919	3209	1.74%

Note: Table 3 presents the computational cost of estimating the multivariate Markov switching model in order to identify the bull and bear states. Results are measured in seconds. The last column is the proportion of the computing time using VI over the one with MCMC.

In my simulation, the VI algorithm converges faster than the MCMC algorithm. Table 3 presents the time required to run each algorithm. The MCMC method takes about 10 to 100 times longer than the VI method. For further investigation, we use the time trace plot of loglikelihood and Geweke (1992)'s plot to detect the convergence of the MCMC approach. Results are in Figure 2 and Figure 7. These graphs indicate that the MCMC starts to show convergence after the VI convergence. It is worth mentioning that we need multiple posterior draws from the MCMC algorithm after it shows convergence. Hence, we can conclude that the VI consumes less computational time than the MCMC.

4.1.2 A multivariate example

I extended the simulation to 30 time series. The data-generation process is designed so that the simulation is close to the real data, with 1000 time periods and 20 regime switches. As a consequence, the duration of the regime is about 5 years in order to mimic the duration of a business cycle. My prior for the transition matrix favours persistence. As a result of expectations for the duration of the business cycle, $p_{k\cdot}$ is then $(p_{11}, p_{12}) \sim Dir(99, 1)$ and $(p_{22}, p_{21}) \sim Dir(99, 1)$. I keep a similar setting for the prior and initial values as in the previous subsection.

I find the same insights as in the simulation with the two asset series. The algorithm with identification restrictions can find the bear and bull states. Figure 10 shows a high similarity visually of each state. Table 2 emphasizes that the identification of the bull and bear state is the same between two technique. The parameters estimated by VI are close to the MCMC and to the true ones. Table 4 presents the average Euclidean distance of the estimated results to the true values. We find a minor discrepancy between the estimated moments and the true ones. Additional, the estimated results are also close to each other.

Table 4: Average Euclidean distance to the true value

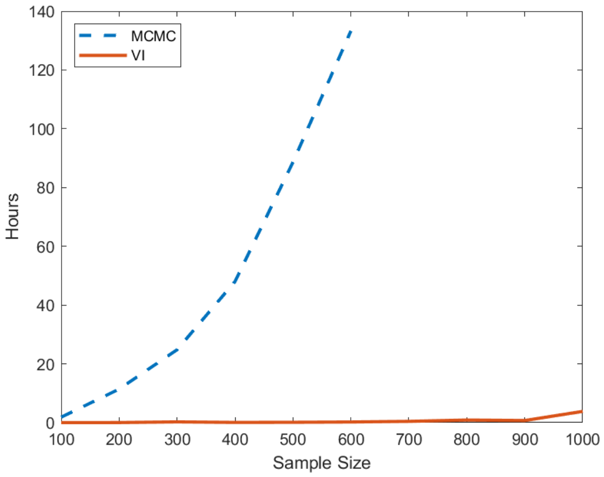
	Bear state		Bull State	
	Mean	Variance	Mean	Variance
MCMC	0.216	0.041	0.010	0.003
VI	0.216	0.041	0.010	0.003
Difference between MCMC and VI	0	0	0	0

Note: Table 4 presents the average Euclidean distance from the posterior and optimal variational mean and variance from MCMC and VI estimation to the true values. The last row is the direct comparison between results of two methods.

I extend the simulation studies to multiple sample sizes with $T = 100$ and N ranging from 100 to 1000.

Figure 1 presents the computing time between two approaches over multiple simulated data sets. The

Figure 1: Computing time under different sample sizes



Notes: Figure 1 compares the computing time of the MCMC and VI algorithms for multiple sample sizes.

computing time of the VI method is quite stable, costing up to 3 hours for a sample with 1000 series. However, the computing cost is much higher in the case of the MCMC method, taking 5 days for the sample with 600 series. Given the tendency of the plot, we expect the huge amount of time needed for the data with 100 series. In our simulation, we cannot infer the posterior within 10 days. This represents a significant improvement from our novel approach.

4.2 Computational cost

I compare the computational costs of the MCMC and VI algorithms. The criterion is the big $\mathcal{O}$ notation that captures the time complexity of the algorithm. Notably, I have a restricted optimisation problem that causes difficulties in finding time complexity.

The algorithm for each parameter is often clear about its complexity. It comes harder when I consider the forward filtering backward smoothing algorithm and the truncated distribution moments. Regarding the former, in literature, the time complexity of the forward-backward algorithm is well-known as $\mathcal{O}(K^2T)$ in one iteration.

However, it is not easy to find the time complexity of algorithms under restrictions. The issue arises from drawing a vector that satisfies a linear restriction and then using this sample to estimate parameters. In particular, it is to solve for the parameters of a truncated multivariate normal distribution. Approximation methods are considered in the literature as Botev (2017), Genz and Bretz (2009), Geweke (1991) and Griffiths et al. (2002). In this paper, I choose Botev (2017)'s method because of its unbiased and generally reliable result. The complexity of Botev (2017)'s method is $\mathcal{O}(N^3)$. This sampling technique ensures that we can apply multiple restrictions to the mean. Notably, the downside of this method is the curse of dimension, as well as the other methods, including MCMC. This approach is reliable with $N \leq 100$.

Table 5: Time complexity of MCMC and VI algorithms

	Unrestricted MCMC	Unrestricted VI	Restricted MCMC	Restricted VI
P	$\mathcal{O}(KT)$	$\mathcal{O}(T)$	$\mathcal{O}(KT)$	$\mathcal{O}(T)$
M, Σ	$\mathcal{O}(KT)$	$\mathcal{O}(K)$	$\mathcal{O}(KT)$	$\mathcal{O}(K)$
S	$\mathcal{O}(K^2T)$	$\mathcal{O}(K^2T)$	$\mathcal{O}(K^2T)$	$\mathcal{O}(K^2T)$
Constraint	-	-	$\mathcal{O}(N^3)$	$\mathcal{O}(N^3)$
Expected values	-	$\mathcal{O}(K)$	-	$\mathcal{O}(K)$
$\widetilde{y}$	-	$\mathcal{O}(KT)$	-	$\mathcal{O}(KT)$
τ	$\mathcal{O}(KT)$	$\mathcal{O}(KT)$	$\mathcal{O}(KT)$	$\mathcal{O}(KT)$
ν	$\mathcal{O}(K)$	$\mathcal{O}(K)$	$\mathcal{O}(K)$	$\mathcal{O}(K)$
Total: $\mathbb{N}$	$\mathcal{O}(K^2T)$	$\mathcal{O}(K^2T)$	$\mathcal{O}(K^2T + N^3)$	$\mathcal{O}(K^2T + N^3)$
Total: $\mathbb{T}$	$\mathcal{O}(K^2T)$	$\mathcal{O}(K^2T)$	$\mathcal{O}(K^2T + N^3)$	$\mathcal{O}(K^2T + N^3)$

Table 5 presents the time complexity of all algorithms in one iteration. In both problems, the time complexity of two algorithms is the same under one iteration. Hence, the VI approach should gain a significant advantage from much faster convergence. This gain is around the ratio between the number of iterations needed.

4.3 In-sample and out-of-sample performance

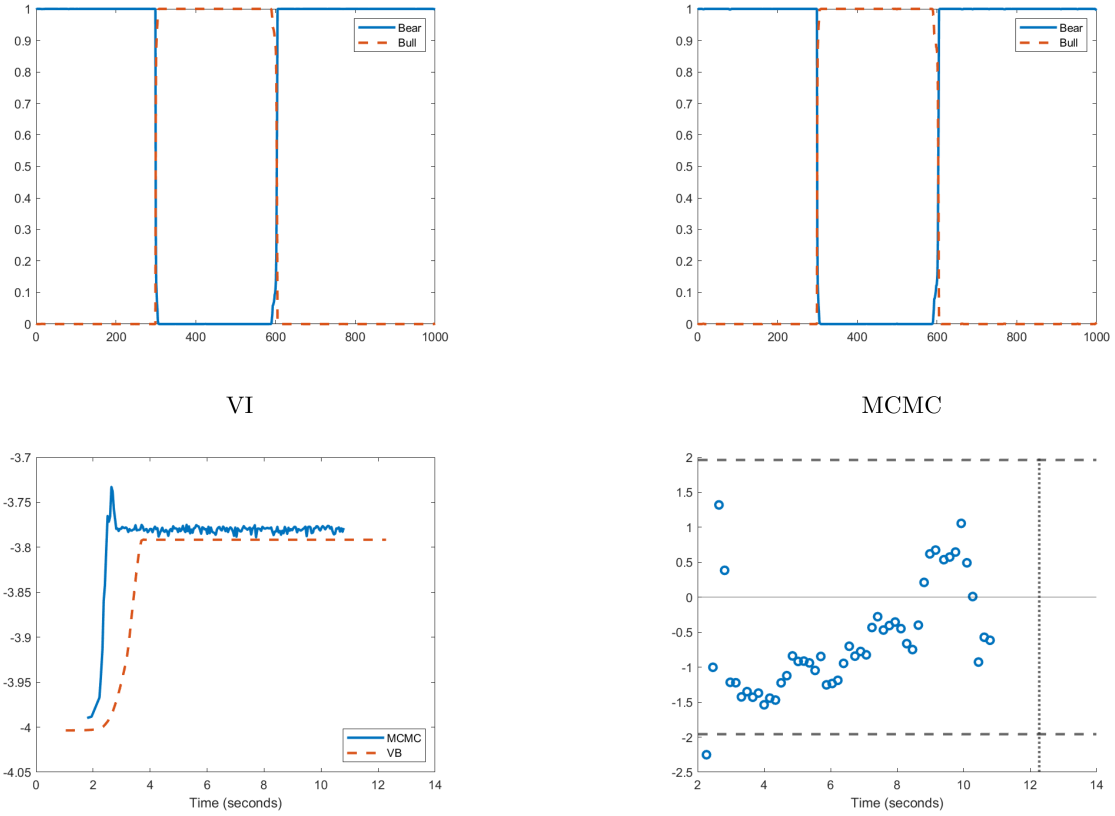
I compare both the in-sample performance and the out-of-sample performance between algorithms. Regarding the in-sample performance, I compare the results of estimators of the mean and covariance between two algorithms. The difference from the true values is measured by the Kulback-Leibler divergence. The higher value indicates the estimates are further from the true parameters.

The above approach does not compare the estimated probability. One solution is to compute the log scores of each estimate. Another criterion is the mean squared error (Chan and Yu, 2022). I generate Z dataset following the same data generating process as in equation 1 with $T = 100$ periods. For each data set $y_t^{(j)} = \{y_1^{(j)}, \ldots, y_N^{(j)}\}$ for $j = 1, \cdots, Z$, I draw G samples of posterior values $p(\Phi, S \mid y_{1:T}^{(g)})$. The log score is as

$$\frac{1}{Z}\frac{1}{G}\frac{1}{T}\sum_{i=1}^{Z}\sum_{g=1}^{G}\sum_{t=1}^{T}\log p(y_t^{(i)} \mid \Phi^{(g)})$$

The estimation shows that a higher log score is more favorable. And for each mean matrix $\widehat{M} =$

Figure 2: Comparison between algorithm without restrictions for 2 assets



VI

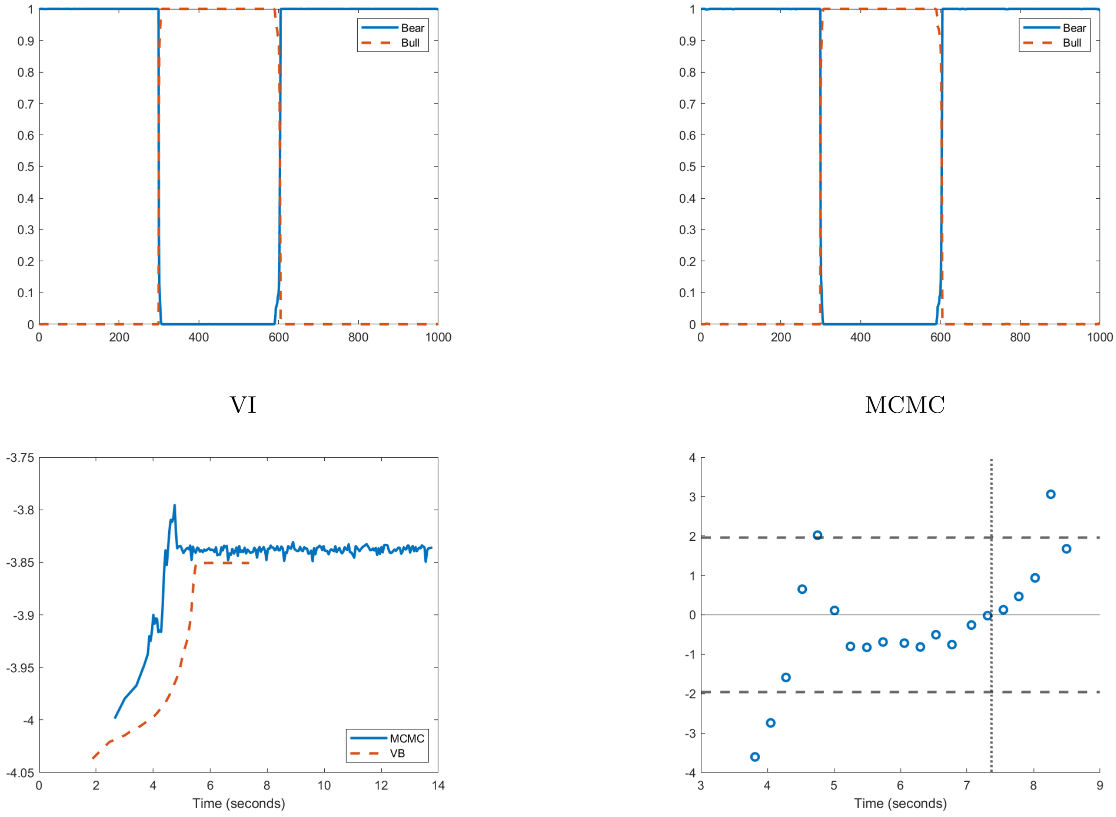
MCMC

VI

MCMC

Notes: Figure 2 compares the results of the MCMC and VI algorithms without restrictions. The first row is the conditional posterior probabilities of states and the optimal variational probabilities of states. The second row is, from left to right, the time trace of the average log likelihood, and the Geweke (1992)'s plots. Both data are normally distributed.

Figure 3: Comparison between algorithm with restrictions for 2 assets



VI MCMC

VI MCMC

Notes: Figure 3 compares the results of the MCMC and VI algorithms with and without restrictions. The first row is the result under the unrestricted approach. The second row is the result of the restricted approach. For each row, from left to right, we have the conditional posterior probabilities of states, the optimal variational probabilities of states, the time trace of the average log likelihood, and the Geweke (1992)'s plots. Both data are normally distributed.

$\{\widehat{M}_1, \cdots, \widehat{M}_T\}'$ where each $\widehat{M}_t = \sum_{s_t} p(s_t \mid y_{1:T}) M_{s_t}$, I calculate the mean squared errors as

$$MSE_j(\widehat{M}) = \frac{\sum\limits_{t=1}^{T} (\widehat{M}_t - M_t)^2}{T}$$

where M_t is the true value of the mean.

In order to measure the in-sample performance of the VI algorithm, I capture the average difference between two sets of results from two algorithms using multiple datasets. In this simulation, I choose $R = 100$ sets that are generated similarly from the same distribution and similar regime switching. In each run, I capture the Kullback-Leibler divergence Kullback and Leibler (1951) of the estimates against true distributions (applicable to the Normal distribution only), log score, and mean squared errors. The average results for three criteria are in Table 6. All the discrepancies are so small that I can find the two estimates very close to each other. This implies that the approximating technique provides a result as good as an exact estimation.

Table 6: Average in-sample performance under Normal distribution

	30 assets			2 assets		
	KL divergence	Log score	MSE	KL divergence	Log score	MSE
MCMC	1.443	-50.138	0.476	0.042	-3.406	0.095
VI	1.441	-50.138	0.479	0.050	-3.402	0.122
Difference	-0.002	0	0.003	0.009	0.003	0.027

Notes: Table 6 contains the average of Kulback-Leibler divergence, log score, and mean squared errors of the estimation from the MCMC and VI algorithms. The last row, "Difference", records the gap between the criterion calculated from the VI algorithm and the one computed from the MCMC algorithm. Data contains 2 assets and 30 assets that are from Normal distribution.

To compare the out-of-sample performance, I use the log predictive scores, which is the average of the log predictive densities, and the mean squared of forecasting errors (Gefang et al., 2022) as

$$MSFE(\widehat{y}_{t+1:t+h}) = \frac{\sum\limits_{j=1}^{h} (\widehat{y}_{t+j} - y_{t+j})^2}{h}$$

where the data is divided into two parts: the training sample is from time 1 to time t, the forecasting sample is the rest. I use the training sample to estimate the model and exploit the rest to compute the predictive likelihood values.

Average results of these measures are in Table 7 and Table 14. The differences between these measures are relatively small. This conclusion is consistently small in Normal distributions and Student t distribution. This implies a comparable forecasting performance.

4.4 Investment strategy

I compare the economic value of the two algorithms through an investment strategy. In particular, I investigate the out-of-sample forecasting performance for each algorithm. The statistics required are predictive mean and predictive covariance, which are computed along with the predictive probability of a state. The prediction is calculated from the posterior simulations. For example, given each draw of posterior value, $\Phi^{(g)}$ and $S^{(g)}$ conditional on the observed sata $y_{1:t}$, s_{t+1} is simulated as in Equation 2 and predicted returns $\widehat{y}_{t+1}$ are simulated by Equation 1. The simulated data $\widehat{y}_{t+1}^{(g)}$ is a sample of the predictive distribution $p(y_{t+1} \mid y_{1:t})$. An estimator of the predictive mean and variance can be computed directly.

For each period in the forecasting horizon, I estimate the model recursively and make the investment decision. Model comparison in terms of log predictive likelihood and investment strategy is conducted on the same out-of-sample window. I consider two strategies: a market timing portfolio and a mean-variance portfolio.

4.4.1 Market timing portfolio

I consider three strategies in this subsection. The first market timing strategy (Strategy 1) is "buy and hold". An investor buys an equally weighted (EW) index at time 1 and keeps his investment until time T. There is no timing in this plan.

The other two strategies are defined based on the predicted probability of market states. In particular, an investor chooses an allocation between the EW index and risk-free assets. The second strategy (Strategy 2) is that the investment is on the EW index if the predicted probability of the bull state, $p(s_t = \mathit{bull} \mid \mathcal{F}_{t-1})$, is more than a threshold, ε. Otherwise, he invests in risk-free assets. $\mathcal{F}_{t-1}$ is the set of all information up to time $t-1$. In my context, this set contains $y_{1:t-1}$.

The third method of market timing (Strategy 3) is to use the predicted probability to assign the proportion of the investment. For example, if the predicted probability of the bull state is 75%, the investor allocates 75% of his wealth to risk-free assets and the rest to the EW index.

I compare the results of each strategy against multiple criteria, such as descriptive statistics and the performance fee. The latter is the amount that the investor pays in order to keep two strategies equivalent in utility. It is calculated based on the quadratic utility function as

$$U(r_t^{pf}) = (1 + r_t^{pf}) - \frac{\gamma}{1+\gamma}(1 + r_t^{pf})^2$$

where r_t^{pf} are the returns to portfolio and γ is the investor's risk aversion coefficient. The performance fee is denoted as Δ. This fee is the cost that equals the utility received from two returns. I find it as

$$\sum_{t=1}^{T} U(r_t^{pf_{MCMC}}) = \sum_{t=1}^{T} U(r_t^{pf_{VI}} - \Delta)$$

where $MCMC$ and VI denote the results estimated through the MCMC and VI algorithms, respectively. If utility from portfolios suggested by the MCMC algorithm is higher, Δ is negative, and vice versa. I can also compare the performance fees between using the buy-and-hold strategy and strategies using different estimation methods.

4.4.2 Mean-variance portfolio

In this part, I focus on two strategies. The weight of each index is allowed to vary in order to solve the corresponding maximisation problem of each strategy. First, I solve for the global minimum variance portfolio in the optimisation problem as

$$\min_{w_t} w_t' \widehat{\Sigma}_t w_t \quad \text{s.t} \quad w_t' \iota = 1$$

where w_t is the weight vector at time t, $\widehat{\Sigma}_t$ is the predicted variance-covariance matrix at time t based on the information up to time $t-1$, and ι is the vector of 1 with a conformable length. The solution is as

$$w_t^{min} = \frac{\widehat{\Sigma}_t^{-1} \iota}{\iota' \widehat{\Sigma}_t^{-1} \iota}$$

The algorithm that yields a lower variance is the better one.

Second, I maximise the Sharpe ratio as in the maximisation problem

$$\max_{w_t} \frac{\widehat{M}_t' w_t}{w_t' \widehat{\Sigma}_t w_t} \quad \text{s.t} \quad w_t' \iota = 1$$

The optimal solution for weight is as

$$w_t^{max} = \frac{\widehat{\Sigma}_t^{-1} \widehat{M}_t}{\iota' \widehat{\Sigma}_t^{-1} \widehat{M}_t}$$

I use this optimal weight to find the predictive returns and the ex-post Sharpe ratio. The algorithm yielding a higher ex-post Sharpe ratio is the better one.

For both measures, I compute their averages through different algorithms.

I expand my comparison to the out-of-sample performance. I choose $h = 100$ steps ahead. Prediction is carried out as in subsection 3.3. For each step ahead, I measure the mean squared forecasting errors, log predictive score, global minimum variance, Sharpe ratio, and performance fee. Among investment strategies, I choose the risk free rate of 0 for the whole horizon, and the cutoff to switch, n, in Strategy 2 is 0.5. I record the performance fee of switching from strategy 2 or 3 to strategy 1 under the forecasting

Table 7: The average out-of-sample performance under Normal distribution

2 assets								
	MSFE	LPS	Min.Variance	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
'MCMC'	11.140	-3.191	3.963	**0.528**	**1.219**	2.310	-0.007	-0.007
'VI'	**10.251**	**-3.000**	**2.952**	0.504	1.118	**2.218**	-0.007	**0**
'Difference'	-0.888	0.191	-1.011	0.024	0.101	0.092	0	0
30 assets								
	MSFE	LPS	Min.Variance	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
'MCMC'	142.14	-36.748	0.013	36.967	**1.614**	0.044	-0.014	**0**
'VI'	**142.13**	**-36.731**	**0.012**	**40.436**	1.612	**0.040**	-0.014	**0**
'Difference'	-0.007	0.017	-0.001	-3.469	0.001	0.004	0	0

Notes: Table 7 consists of the measures of the out-of-sample performance for the MCMC and VI algorithms. All values are in average. MSFE is the mean squared errors of forecasting errors. LPS means the log predictive score. Min.Variance implies the global minimum variance. Sharpe ratio captures the ex-post Sharpe ratio. Mean is the mean of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. SD is the standard deviation of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. Columns Strategy 2 and 3 present the performance fee of switching from these strategies towards "buy-and-hold" strategy. The row Difference shows the gap between the results from the VI and MCMC algorithms.

results of both algorithms. In addition, I can observe the gap in this fee for the two strategies using results from two techniques.

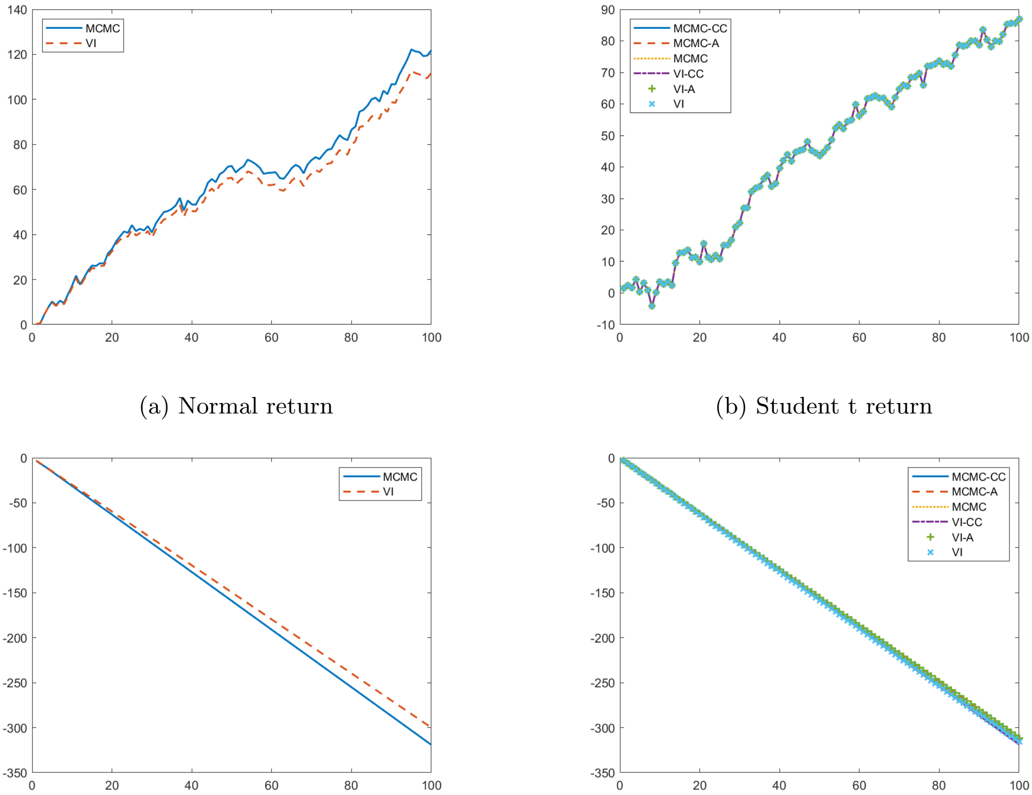
Average results for the whole window are in Table 7 and Table 14. In addition, Figure 4 presents a comparison in cumulative mean squared errors and log score. I can find that the performance of two algorithms is equivalent because the gap between indicators is minor. The special case is that the investments under Strategy 3 are preferred over the "buy-and-hold" strategy.

5 Applications

I applied my novel algorithms to two sets of data. In this section, only restricted algorithms are considered. I use the same computing system as in my simulation. The first data set is the monthly industry portfolio returns from the Kenneth French library (French, 2022). The second data set is the monthly returns from equity from CRSP data (CRSP, 2022). All returns are annualised. The monthly risk-free rate is annualised data from the Kenneth French Library.

Both sets involve multiple time series of returns. The difference between two data sets is twofold. First, the CRSP data may include more time series than the first one. The reason is that there are many firms on the S&P 500 list, and the number of firms should be greater than the number of industries. Second,

Figure 4: The out-of-sample performance for 2 assets



(a) Normal return

(b) Student t return

(c) Normal log likelihood

(d) Student t log likelihood

Notes: Figure 4 presents the out-of-sample comparison between two inference approaches. The first row is the cumulative one-period-ahead predictive return that is calculated from the optimal ex-ante Sharpe ration. The second row is the cumulative predictive loglikehood. The first column is the result of the Normal data. The next two columns are the results from the Student t distribution.

the industry portfolio data suffers less from the survivorship bias because only existing, well-performing stocks are included. If there is this bias, the data from the S&P 500 does not involve firms that perform poorly. Therefore, it may be less biassed to use the states identified from this data to imply the business cycle.

Table 8: Monthly returns descriptive statistics

	Start	End	N	T	Mean	SD	Excess Kurtosis
EW34	1926-Jan	2022-Mar	34	1155	0.113	1.237	1391.9
EW59	1931-Mar	2022-Mar	59	1093	0.115	1.209	3080.4
EW103	1951-Jan	2022-Mar	103	855	0.107	0.967	2234.8
IP40	1926-Jul	2022-Oct	40	1156	1.033	8.609	2795.9

Note: Table 8 presents several descriptive statistics for all four data sets. In particular, I have the starting month, the ending month, the number of series, duration, mean, standard deviation, skewness and kurtosis. The multivariate skewness and kurtosis are calculated by the formula in Mardia (1970).

Regarding the CRSP monthly returns, I use returns from equity, excluding dividends. The data set should be as long as possible in order to convey more observations. However, firms are frequently dropped from the data. Therefore, the trade-off is unavoidable. The longer the span, the less firms stay. I then chose three subsets from the original data. The first set (EW34) has 34 stocks, spanning from January 1926 to March 2022. The second set (EW59) has 59 stocks, spanning from January 1951 to March 2022. The third set (EW103) has 103 stocks spanning from January 19 to March 2022. Their total number of months is 1155, 1093, and 855, respectively. In this setting, I allow for the longest possible data while maintaining a high number of firms.

The industry portfolio return data (IP40) is extracted from the Kenneth F. French library. The returns exclude dividends. It spans from July 1926 to October 2022. Hence, I have 1156 time periods and 40 time series in this data.

Table 8 presents the descriptive statistics of these data sets. The multivariate kurtosis are calculated by methods in Mardia (1970). The positive excess kurtosis may imply leptokurtism. Although we cannot conclude about the population property using these sample values, it is worth assuming both the Normal and Student t distribution assumptions.

Table 9: Computation time (in seconds)

	Normal			Student t								
	MCMC	VI	Proportion	MCMC-CC	MCMC-A	MCMC	VI-CC	VI-A	VI	Corresponding Proportion		
'EW34'	1,477.3	35.8	2.42%	9,776.2	8,893.9	8,514.8	35.9	26.2	68.1	0.37%	0.29%	0.80%
'EW59'	3,585.3	38.1	1.06%	19,669	21,672	21,157	30.1	30.4	58.9	0.15%	0.14%	0.27%
'EW103'	7,471.6	107.4	1.43%	42,113	39,582	41,858	50.1	51.0	90.2	0.12%	0.13%	0.22%
'IP40'	9,610.5	16.3	0.17%	10,567	10,424	10,630	26.3	23.9	58.4	0.25%	0.23%	0.55%

Note: Table 9 presents the computational cost of estimating the multivariate Markov switching model. Results are measured in seconds.

Table 10: Descriptive statistics of two states under restrictions

	EW34				EW59				EW103				IP40			
	Normal															
	M^1	std^1	M^2	std^2	M^1	std^1	M^2	std^2	M^1	std^1	M^2	std^2	M^1	std^1	M^2	std^2
MCMC	-0.025	1.838	0.098	0.847	-0.026	1.815	0.093	0.847	-0.040	1.560	0.093	0.832	-0.041	8.566	0.079	3.471
VI	-0.026	1.845	0.101	0.838	-0.028	1.874	0.092	0.848	-0.022	1.443	0.075	0.812	-0.162	13.248	0.610	5.871
	Student t															
	M^1	std^1	M^2	std^2	M^1	std^1	M^2	std^2	M^1	std^1	M^2	std^2	M^1	std^1	M^2	std^2
MCMC-CC	-0.023	1.474	0.104	1.007	-0.043	1.674	0.090	0.965	-0.289	1.256	0.111	0.871	-0.070	8.570	0.312	4.229
MCMC-A	-0.024	1.256	0.102	0.976	-0.041	1.357	0.090	0.946	-0.292	1.276	0.110	0.887	-0.065	8.188	0.322	4.338
MCMC	-0.023	1.221	0.105	0.970	-0.039	1.336	0.090	0.943	-0.204	1.359	0.102	0.881	-0.069	7.927	0.381	4.335
VI-CC	-0.014	1.273	0.111	0.909	-0.027	1.397	0.123	0.958	-0.004	0.973	0.113	0.808	-0.183	9.956	0.341	7.057
VI-A	-0.019	1.203	0.113	0.966	-0.040	1.327	0.096	0.936	-0.073	1.045	0.127	0.870	-0.234	9.238	0.484	6.835
VI	-0.019	2.344	0.113	0.995	-0.027	1.498	0.122	1.086	-0.004	1.320	0.115	0.836	-0.183	10.943	0.322	8.011

Notes: Table 10 contains average values of estimated means and variances of two states by MCMC and VI algorithms. M^i denotes the average of means of returns in State i. std^i denotes the average of standard deviations of returns in State i. The results are on three equally-weighted stock returns series and one industry portfolio returns.

The risk aversion coefficient is $\gamma = 2$ as in Campbell and Cochrane (1999). The threshold for Strategy 2 is $\varepsilon = 0.5$. I choose the prior of the transition matrix as the business cycle is every 5 years on average.

Table 9 shows the computation time of each algorithm for all data sets. My VI algorithm demonstrates a clear advantage in speed of estimation. Similar to my simulation, the MCMC algorithm requires much more time to finish. For example, MCMC may take data from the 103 series around 11 hours to complete, while VI needs only less than 3 minutes.

I justify the VI method through estimated parameters and forecasting performance. Table 10 presents the average of estimates for all datasets. Results from two algorithms are comparable. Discrepancies from these results get larger under the larger dimensions. In general, I find the mean of asset returns is lower on average and is more dispersed in the bear state. This is common in all datasets of this section.

Figure 14, 15, 16, and 17 (in Appendix H) show estimates of the correlation matrix in each state . In all states, I mostly find positive correlations. Visually, the results from VI and MCMC are very similar.

Figure 5 presents the dynamics of forecasting performance between two algorithms. I have plots of cumulative predictive returns, which are the returns using the optimal weight of the ex ante Sharp ratio, and cumulative predictive loglikelihood. Results from VI and MCMC are close to each other.

The two sets of results from multivariate Normal distribution and multivariate Student t distribution are very similar. Table 10 presents that the bear state exhibits more volatility than the bull state. Descriptive statistics from different methods are more distant under higher dimensions.

The forecasting criteria show a mixed preference for both methods. Table 11 demonstrates that the forecasting performance of the portfolio with 34 assets is better with the assumption of multivariate Student t distribution under mean square forecasting errors, log predictive score, Sharpe ratio, minimum variance and performance fee. Meanwhile, Table 16 and 17 imply that the results under multivariate Normal distribution become more favorable when the dimension is larger. Table 18 presents a mixed preference between two multivariate distributions. I often find that the higher Sharpe ratio aligns with the higher mean of returns with high risk while the lower minimum variance accompanies with the lower standard deviation in the Sharpe ratio. Investment strategies using variational inference perform similarly to those using the MCMC method. I find that strategies using the MS model perform better than the "buy-and-hold" strategy. In addition, as the dimension increases, more criteria favor the multivariate normal assumption.

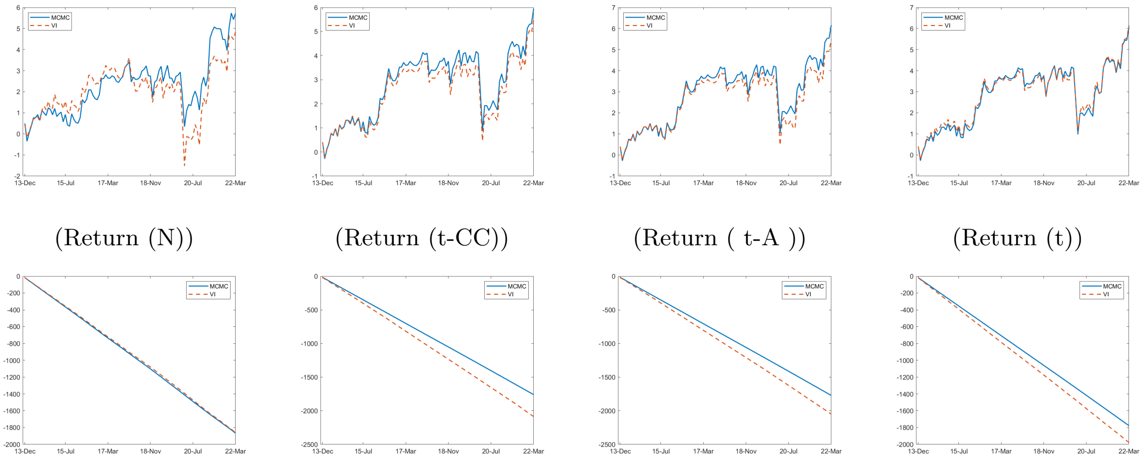
Table 11 shows the average criteria for forecasting and investment performance (Results of other datasets are in Appendix G.). Results in terms of which method is better are mixed. However, these criteria present values that are close to each other. Therefore, empirical results indicate that my new VI algorithm is comparable to the MCMC method in terms of forecasting performance.

Table 11: Out-of-sample performance for a portfolio with 34 assets

	MSFE	LPS	Min.Var	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
				Normal				
MCMC	49.468	-18.666	0.263	0.121	0.057	0.471	-0.001	0
VI	49.272	-18.607	0.348	0.089	0.048	0.546	-0.001	0
				Student t				
MCMC-CC	49.297	**-17.610**	0.186	0.130	0.060	0.456	**0**	0
MCMC	49.302	-17.776	0.159	0.135	0.061	0.449	**0**	0
MCMC-A	49.289	-17.734	0.162	**0.137**	**0.062**	0.449	**0**	0
VI-CC	49.283	-20.875	**0.154**	0.122	0.055	**0.448**	**0**	0
VI	49.361	-19.778	0.232	**0.137**	**0.062**	0.449	-0.001	0
VI-A	**49.233**	-20.511	0.157	0.119	0.055	0.460	**0**	0

Notes: Table 11 stores the measures of out-of-sample performance for both algorithms for the data of an equally-weighted portfolio with 34 assets. MSFE is the mean squared errors of forecasting errors. LPS means log predictive score. Min.Var implies the global minimum variance. Sharpe ratio captures the ex-post Sharpe ratio. Mean is the mean of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. SD is the standard deviation of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. Columns Strategy 2 and 3 present the performance fee of switching from these strategies towards "buy-and-hold" strategy.

Figure 5: Cumulative predictive returns and loglikelihood for EW34



(Return (N))

(Return (t-CC))

(Return (t-A))

(Return (t))

(LLH (N))

(LLH (t-CC))

(LLH (t-A))

(LLH (t))

Notes: Figure 5 consists of out-of-sample results from both VI and MCMC algorithms for the portfolio of 34 equally-weighted assets. The first column on the left is the result of the Normal assumption. The next columns are the results of Student t assumption. The upper row contains the cumulative predictive returns that are from the optimal weight calculated by the ex-ante Sharpe ratio. The second row includes the cumulative predictive loglikehood.

In all distribution assumptions, we find a positive correlation between asset returns. Figure 14, 15, 16 and 17 depict these strong associations in both bull and bear states. The correlation is stronger under the data from 40 industry portfolio.

It is difficult to conclude whether the data under higher dimension belongs to the multivariate Normal or Student t distribution in my application. I find it convinced to assume both types of distributions in further applications.

The forecasting and investment applications demonstrate the value of the novel variational inference in exploiting information and weighting each asset in the portfolio.

6 Conclusion

This paper proposes a novel VI algorithm to provide inference for the multivariate Markov switching model. I contribute to the literature in several facets. The new method gives fast and accurate results in comparison to an equivalent MCMC approach. The new technique incorporates important restrictions that require identifying the hidden states easily and tractably. The forward filtering backward smoothing algorithm is similar to the well-known Chib (1996)'s algorithm in economic literature.

This new VI algorithm estimates the multivariate Markov switching model quickly and accurately. I measure its accuracy through in-sample and out-of-sample performance. Regarding the in-sample performance, a comparison is made between the average values of Kullback-Leibler divergence and the true distribution of the mean, log score, and mean squared errors from 100 different samples under the same data generation process. My novel approach yields results similar to those of a comparable MCMC method.

Regarding out-of-sample performance, I measure the average of mean squared forecasting error, log predictive score, and economic value in investment strategy. There are three strategies: buy and hold an equally weighted portfolio; use the probability of the bull state to signal the time to invest in a risk-free asset; and use the predicted probability to allocate the invested weight. My algorithm finds a similar result to an equivalent MCMC method.

The significant gain from this new algorithm is its computational time. As the time complexity in each iteration is similar between the VI and MCMC since the VI quickly converges, it gains a significant improvement in computational time from exploiting this approach. In my simulation, the time consumed by VI is about a couple hundred times shorter than the comparable MCMC algorithm. In my application, the new method reduces the estimation time from 11 hours to 3 minutes.

My novel algorithm includes a restriction to identify the bull and bear states from stock returns. This restriction is crucial in economic literature, but it is not addressed elsewhere. The reason is that this restriction impacts the assumption of exponential conjugacy in VI. Without this conjugate assumption, it is difficult to find an analytical solution. Since my algorithm is capable of incorporating it, results are identified, and I can interpret their economic meaning.

My method follows the popular Chib (1996)'s algorithm that is well-known in economic literature. The literature on variational inference only focuses on the Baum-Welch forward-backward algorithm, which is uncommon in economics. My method is more familiar to economists who apply the variational inference technique in their research.

I apply this method to the S&P 500 and Kenneth datasets to identify the bull and bear states of the market and compare their forecasting performance for investment. Results suggest that the new method shows close forecasting performance while using significantly less computational time. The MCMC algorithm is a popular solution to both versions of the Markov switching model and provides exact inference and good predictive performance.

The algorithm has its weaknesses. As the solution for my restricted VI algorithm is tractable, values in the solution can be obtained through sampling only. In my proposed algorithm, I use Botev's sampling technique, which is modern and reliable. However, this technique claims its weakness once the dimension of the data is greater than 100.

References

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference.* University of London, University College London (United Kingdom).

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148.

Campbell, J. Y. and Cochrane, J. H. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of political Economy*, 107(2):205–251.

Chan, J. C. and Yu, X. (2022). Fast and accurate variational inference for large bayesian vars with stochastic volatility. *Journal of Economic Dynamics and Control*, 143:104505.

Chan, K. F., Treepongkaruna, S., Brooks, R., and Gray, S. (2011). Asset market linkages: Evidence from financial, commodity and real estate assets. *Journal of Banking & Finance*, 35(6):1415–1426.

Chatzis, S. P., Kosmopoulos, D. I., and Varvarigou, T. A. (2008). Robust sequential data modeling using an outlier tolerant hidden markov model. *IEEE transactions on pattern analysis and machine intelligence*, 31(9):1657–1669.

Chen, S.-S. (2009). Predicting the bear stock market: Macroeconomic variables as leading indicators. *Journal of Banking & Finance*, 33(2):211–223.

Chevallier, J. (2012). Global imbalances, cross-market linkages, and the financial crisis: A multivariate markov-switching analysis. *Economic Modelling*, 29(3):943–973.

Chib, S. (1996). Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, 75(1):79–97.

Chib, S., Nardari, F., and Shephard, N. (2002). Markov chain monte carlo methods for stochastic volatility models. *Journal of Econometrics*, 108(2):281–316.

Christmas, J. and Everson, R. (2010). Robust autoregression: Student-t innovations using variational bayes. *IEEE Transactions on Signal Processing*, 59(1):48–57.

CRSP (2022). Center for research in security prices. [online]. Available at: WRDS `http://wrds-web.wharton.upenn.edu/wrds/` (Accessed: 14 November 2022).

Foti, N., Xu, J., Laird, D., and Fox, E. (2014). Stochastic variational inference for hidden markov models. *Advances in neural information processing systems*, 27.

Frazier, D. T., Loaiza-Maya, R., and Martin, G. M. (2022). Variational bayes in state space models: Inferential and predictive accuracy. *Journal of Computational and Graphical Statistics*, pages 1–12.

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2021). Loss-based variational bayes prediction. *arXiv preprint arXiv:2104.14054*.

French, K. R. (2022). 49 industry portfolios [ex. dividends]. data retrieved from Kenneth R. French data library, `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, volume 425. Springer.

Gefang, D., Koop, G., and Poon, A. (2022). Forecasting using variational bayesian inference in large vector autoregressions with hierarchical shrinkage. *International Journal of Forecasting.*

Genz, A. and Bretz, F. (2009). *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, volume 571, page 578. Fairfax, Virginia: Interface Foundation of North America, Inc.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*, 4:641–649.

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of applied econometrics*, 8(S1):S19–S40.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural computation*, 12(4):831–864.

Griffiths, W. et al. (2002). A gibb's sampler for the parameters of a truncated multivariate normal distribution. *Research paper - University of Melbourne Department of Economics.*

Gruhl, C. and Sick, B. (2016). Variational bayesian inference for hidden markov models with multivariate gaussian output distributions. *arXiv preprint arXiv:1605.08618.*

Guidolin, M. and Timmermann, A. (2007). Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, 31(11):3503–3544.

Guidolin, M. and Timmermann, A. (2008). International asset allocation under regime switching, skew, and kurtosis preferences. *The Review of Financial Studies*, 21(2):889–935.

Haase, F. and Neuenkirch, M. (2023). Predictability of bull and bear markets: A new look at forecasting stock market regimes (and returns) in the us. *International Journal of Forecasting*, 39(2):587–605.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384.

Ji, S., Krishnapuram, B., and Carin, L. (2006). Variational bayes for continuous hidden markov models and its application to active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):522–532.

Kole, E. and Van Dijk, D. (2017). How to identify and forecast bull and bear markets? *Journal of Applied Econometrics*, 32(1):120–139.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Liu, C. and Rubin, D. B. (1995). Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica sinica*, pages 19–39.

Liu, J. and Maheu, J. M. (2018). Improving markov switching models using realized variance. *Journal of Applied Econometrics*, 33(3):297–318.

Maheu, J. M. and McCurdy, T. H. (2000). Identifying bull and bear markets in stock returns. *Journal of Business & Economic Statistics*, 18(1):100–112.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.

McGrory, C. A. and Titterington, D. (2009). Variational bayesian analysis for hidden markov models. *Australian & New Zealand Journal of Statistics*, 51(2):227–244.

Nakajima, J. and Omori, Y. (2012). Stochastic volatility model with leverage and asymmetrically heavy-tailed error using gh skew student's t-distribution. *Computational Statistics & Data Analysis*, 56(11):3690–3704.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

Quiroz, M., Nott, D. J., and Kohn, R. (2022). Gaussian variational approximations for high-dimensional state space models. *Bayesian Analysis*, 1(1):1–28.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden markov model. *Journal of applied econometrics*, 13(3):217–244.

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). Variational bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882.

Turner, C. M., Startz, R., and Nelson, C. R. (1989). A markov model of heteroskedasticity, risk, and learning in the stock market. *Journal of Financial Economics*, 25(1):3–22.

Appendices

A Student t distribution

Let each vector $y_t = \{y_{1t}, y_{2t}, \cdots, y_{it}, \cdots, y_{N_t}\}$ identically independently distributed in $Y = \{y_t\}_{t=1}^{T}$ follow a multivariate Student t-distribution, $t(\mu, \Sigma, \nu)$, where μ is the mean, Σ is the scale (a positive definite matrix) and ν is the degrees of freedom. Its probability distribution function is

$$f(y_t \mid \mu, \Sigma, \nu) = \frac{\Gamma\left[(\nu + N)/2\right]}{\Gamma(\nu/2)\nu^{N/2}\pi^{N/2}\left|\mathbf{\Sigma}\right|^{1/2}}\left[1 + \frac{1}{\nu}(y_t - \mu)^{'}\Sigma^{-1}(y_t - \mu)\right]^{-(\nu+N)/2}$$

where the expression outside the brackets is the normalizing constant, μ is the mean, $\Sigma > 0$ is the scale, $\nu > 0$ is the degree of freedom. Note that:

$$\begin{aligned} \text{mean}(y_t) &= \mu \text{ if } \nu > 1 \\ \text{var}(y_t) &= \frac{\nu}{\nu - 2}\Sigma \text{ if } \nu > 2 \end{aligned}$$

We re-write the probability distribution function as

$$\begin{aligned} y_t \mid \mu, \Sigma, \tau_t &\sim \mathbb{N}(\mu, \Sigma\tau_t^{-1}) \\ \tau_t \mid \nu &\sim \mathbb{G}\left(a, b\right) \end{aligned}$$

with $a = b = \frac{\nu}{2}$ as in Liu and Rubin (1995). It follows that the probability distribution function of this Gamma distribution is

$$f_G(\tau_t \mid \nu) = \Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\frac{\nu}{2}\right)^{\nu/2}\tau_t^{\nu/2-1}\exp\left\{-\frac{\nu\tau_t}{2}\right\}$$

We choose prior as follows

$$\begin{aligned} M \mid \Sigma &\sim \mathbb{N}(\mu, h^{-1}\Sigma) \\ \Sigma &\sim \mathbb{IW}\left(n, \Psi\right) \\ \nu &\sim \mathbb{G}(v_\nu, s_\nu) \end{aligned}$$

The likelihood is

$$\prod_{t=1}^{T} f(y_t \mid \mu, \Sigma, \nu, \tau) = \prod_{t=1}^{T}(2\pi)^{-N/2}\tau_t^{N/2}|\Sigma|^{-1/2}\exp\left\{\frac{\tau_t}{2}(y_t - \mu)^{'}\Sigma^{-1}(y_t - \mu)\right\}$$

The MCMC algorithm for the Student t distribution needs to scale the auxiliary variable τ in order for it to have the mean of 1.

B MCMC algorithm

B.1 Parameter space

I have parameters for the Normal distribution case as

1. $\{M_k\}_{k=1}^{K}$: the mean coefficient of the returns.
2. $\{\Sigma_k\}_{k=1}^{K}$: the variance of the returns. And both also vary by states.
3. P: transition matrix. The element p_{ij} captures the probability of state i switching to state j.
4. $S = \{s_1, s_2, \ldots, s_T\}$: the state.

Regarding the Student t distribution, I have some additional parameter

1. $\{\nu_k\}_{k=1}^{K}$: the degrees of freedom

B.2 Variance of a Student t

I use the variance formula of Student t distribution to estimate the degrees of freedom that is

$$V = \frac{\nu}{\nu - 2}\Sigma$$

In particular, I find the sample variance covariance matrix, $\widehat{V}$, the scale matrix, Σ, and then recover the degrees of freedom, ν. Note that two sides of the equation are two matrices. I find the value of ν that minimises the Frobenius matrix distance between two matrices.

Suppose that there are two positive definite matrices X and Y that have the same dimension $M \times N$. Their Frobenius matrix distance is

$$F = \sqrt{\sum_{i}^{M}\sum_{j}^{N} |a_{ij}|^2} = \sqrt{tr(AA')}$$

where a_{ij} is the element on row i^{th} and column j^{th} of matrix $A = X - Y$ with dimension $M \times N$.

B.3 Stirling's approximation

The Stirling's approximation is that

$$\log\Gamma\left(\frac{\nu_k}{2}\right) \approx \left(\frac{\nu_k}{2} - \frac{1}{2}\right)\log\left(\frac{\nu_k}{2}\right) - \left(\frac{\nu_k}{2}\right) + \frac{1}{2}\log 2\pi$$

$$\Gamma\left(\frac{\nu_k}{2}\right) = \sqrt{\frac{2\pi}{\nu_k/2}}\left(\frac{\nu_k}{2e}\right)^{\nu_k/2}\left(1+O\left(\frac{2}{\nu_k}\right)\right)$$

where the big O implies the error terms in this approximation. In this scenario, the ratio between $\Gamma\left(\frac{\nu_k}{2}\right)$ and $\sqrt{\frac{2\pi}{\nu_k/2}}\left(\frac{\nu_k}{2e}\right)^{\nu_k/2}$ is at most proportional to $1+O\left(\frac{2}{\nu_k}\right)$ when $\nu_k/2$ is sufficiently large. It indicates that the approximation is not good when ν_k is close to the lower bound of 2 since the true term is twice of the approximated representation.

The prior of ν_k is

$$p(\nu_k) \propto \nu_k^{v_\nu - 1} \exp\left\{-\nu_k s_\nu\right\}$$

The likelihood is

$$p(\tau_{\cdot k}, S \mid M, \Sigma, y) \propto \prod_{t=1}^{T}\left(\Gamma^{-1}\left(\frac{\nu_k}{2}\right)\left(\frac{\nu_k}{2}\right)^{\frac{\nu_k}{2}} \tau_t^{\frac{\nu_k}{2}-1} \exp\left\{-\frac{\nu_k \tau_t}{2}\right\}\right)^{\mathbf{1}(s_t=k)}$$

Applying the Stirling's approximation and expand the likelihood, we find

$$\begin{aligned} p(\tau_{\cdot k}, S \mid M, \Sigma, y) &\propto \prod_{t=1}^{T}\left(\left(1+O\left(\frac{2}{\nu_k}\right)\right)^{-1}\left(\frac{\nu_k}{2}\right)^{\frac{1}{2}}\right)^{\mathbf{1}(s_t=k)} \\ &\quad\times\left(\exp\left\{\frac{\nu_k}{2}\right\}\exp\left\{\left(\frac{\nu_k}{2}-1\right)\log\tau_t\right\}\exp\left\{-\frac{\nu_k\tau_t}{2}\right\}\right)^{\mathbf{1}(s_t=k)} \\ &\propto \prod_{t=1}^{T}\left(\left(1+O\left(\frac{2}{\nu_k}\right)\right)^{-1}\left(\frac{\nu_k}{2}\right)^{\frac{1}{2}}\exp\left\{-\frac{\nu_k}{2}\left(\tau_t - \log\tau_t - 1\right)\right\}\right)^{\mathbf{1}(s_t=k)} \end{aligned}$$

My conditional posterior of ν_k is

$$\begin{aligned} p(\nu_k \mid \tau_{\cdot k}, S) &\propto \nu_k^{v_\nu - 1}\Gamma^{-\sum_{t=1}^{T}\mathbf{1}(s_t=k)}\left(\frac{\nu_k}{2}\right)\left(\frac{\nu_k}{2}\right)^{\sum_{t=1}^{T}\mathbf{1}(s_t=k)\nu_k/2} \\ &\quad \exp\left\{-\nu_k\left(\frac{\sum_{t=1}^{T}\mathbf{1}(s_t=k)(\tau_t-\log\tau_t)}{2}+s_\nu\right)\right\} \\ &\propto \nu_k^{v_\nu-1}\left(\frac{\nu_k}{2}\right)^{\sum_{t=1}^{T}\mathbf{1}(s_t=k)\nu_k/2}\exp\left\{-\nu_k\left(\frac{\sum_{t=1}^{T}\mathbf{1}(s_t=k)(\tau_t-\log\tau_t-1)}{2}+s_\nu\right)\right\} \\ &\quad\times\left(1+O\left(\frac{2}{\nu_k}\right)\right)^{-\sum_{t=1}^{T}\mathbf{1}(s_t=k)} \end{aligned}$$

Suppose that we have a good approximation,

$$p(\nu_k \mid \tau, S) \propto \nu_k^{\left(v_\nu - 1 + \frac{\sum_{t=1}^{T}\mathbf{1}(s_t=k)}{2}\right)} \exp\left\{-\nu_k\left(\frac{\sum_{t=1}^{T}\mathbf{1}(s_t=k)(\tau_t-\log\tau_t)}{2}+s_\nu-\frac{\sum_{t=1}^{T}\mathbf{1}(s_t=k)}{2}\right)\right\}$$

B.4 Forward filtering backward smoothing

B.4.1 Forward filtering

The forward filtering of states at time t is $F(s_t) = p(s_t \mid y_{1:t}, \Phi)$. So if $t = 1$, I have

$$p(s_1 \mid y_1, \Phi) = \frac{p(s_1)p(y_1 \mid s_1, M, \Sigma)}{p(y_1 \mid \Phi)} = \frac{p(s_1)p(y_1 \mid s_1, M, \Sigma)}{\sum_{s_1} p(s_1)p(y_1 \mid s_1, M, \Sigma)}$$

where I utilise that $p(s_1) = p(s_1 \mid y_1, M, \Sigma)$. If $t = 2$, I have

$$\begin{aligned} p(s_2 \mid y_{1:2}, \Phi) &= \frac{p(s_2, y_2 \mid y_1, \Phi)}{p(y_2 \mid y_1, \Phi)} \\ &= \frac{p(y_2 \mid s_2, y_1, M, \Sigma)p(s_2 \mid y_1, \Phi)}{\sum_{s_2} p(s_2, y_2 \mid y_1, \Phi)} \\ &= \frac{p(y_2 \mid s_2, y_1, M, \Sigma)p(s_2 \mid y_1, \Phi)}{\sum_{s_2} p(y_2 \mid s_2, y_1, M, \Sigma)p(s_2 \mid y_1, \Phi)} \end{aligned}$$

where

$$p(s_2 \mid y_1, \Phi) = \sum_{s_1} p(s_2 \mid s_1, y_1, P)p(s_1 \mid y_1, \Phi) = \sum_{s_1} p(s_2 \mid s_1, y_1, P)F(s_1)$$

Hence, for any time t, the forward pass is

$$p(s_t \mid y_{1:t}, \Phi) = \frac{p(y_t \mid s_t, M, \Sigma)p(s_t \mid y_{1:t-1}, \Phi)}{\sum_{s_t} p(y_t \mid s_t, M, \Sigma)p(s_t \mid y_{1:t-1}, \Phi)}$$

with $p(s_t \mid y_{1:t-1}, \Phi) = \sum\limits_{s_{t-1}} F(s_{t-1})p(s_t \mid s_{t-1}, P)$. Once I are at the terminal T, the forward is

$$p(s_T \mid y_{1:T}, \Phi) = \frac{p(y_t \mid s_T, M, \Sigma) \sum\limits_{s_{T-1}} p(s_{T-1} \mid y_{1:T-1}, \Phi)p(s_T \mid s_{T-1}, P)}{\sum\limits_{s_T} p(y_t \mid s_T, M, \Sigma) \sum\limits_{s_{T-1}} p(s_{T-1} \mid y_{1:T-1}, \Phi)p(s_T \mid s_{T-1}, P)}$$

This is also the conditional posterior distribution of s_T.

B.4.2 Backward smoothing

The backward smoother goes iteratively from time T to time 1. Note that the marginal density for s_T is available for the forward filtering step. Hence, I start to apply the backward smoother proposed by Chib (1996) to find

$$p(s_{T-1} \mid y_{1:T}, \Phi) = \sum_{s_T} p(s_{T-1} \mid s_T, y_{1:T}, \Phi)p(s_T \mid y_{1:T}, \Phi) \tag{15}$$

This method is feasible under the assumption of knowing the state at time T. The last term on the right hand side is the forward filtering at time T. I focus on the first term

$$p(s_{T-1} \mid s_T, y_{1:T}, \Phi) = \frac{p(y_T \mid s_T, s_{T-1}, y_{1:T-1}, \Phi)p(s_{T-1} \mid s_T, y_{1:T-1}, \Phi)}{p(y_T \mid s_T, y_{1:T-1}, \Phi)}$$

Given s_T, s_{T-1} is irrelevant to y_t in the first term. This term is now similar to the denominator. The second term is as

$$p(s_{T-1} \mid s_T, y_{1:T-1}, \Phi) = \frac{p(s_T \mid s_{T-1}, \Phi)p(s_{T-1} \mid y_{1:T-1}, \Phi)}{\sum\limits_{s_{T-1}} p(s_T \mid s_{T-1}, \Phi)p(s_{T-1} \mid y_{1:T-1}, \Phi)}$$

I then deduce that

$$p(s_{T-1} \mid s_T, y_{1:T}, \Phi) = \frac{p(s_T \mid s_{T-1}, \Phi)p(s_{T-1} \mid y_{1:T-1}, \Phi)}{\sum\limits_{s_{T-1}} p(s_T \mid s_{T-1}, \Phi)p(s_{T-1} \mid y_{1:T-1}, \Phi)}$$

The marginal probability of s_{T-1} from Equation 15

$$p(s_{T-1} \mid y_{1:T}, \Phi) = \sum_{s_T} p(s_T \mid y_{1:T}, \Phi) \frac{p(s_T \mid s_{T-1}, \Phi)p(s_T \mid y_{1:T-1}, \Phi)}{\sum_{s_{T-1}} p(s_T \mid s_{T-1}, \Phi)p(s_{T-1} \mid y_{1:T-1}, \Phi)}$$

Analogously, for any t, I have the conditional posterior density of s_t is as

$$p(s_t \mid y_{1:T}, \Phi) = \sum_{s_{t+1}} p(s_t \mid s_{t+1}, y_{1:T}, \Phi)p(s_{t+1} \mid y_{1:T}, \Phi)$$

The first term is

$$p(s_t \mid s_{t+1}, y_{1:T}, \Phi) = \frac{p(y_{t+1:T} \mid s_t, s_{t+1}, y_{1:t}, \Phi)p(s_t \mid s_{t+1}, y_{1:t}, \Phi)}{p(y_{t+1:T} \mid s_{t+1}, y_{1:t}, \Phi)}$$

Under the same argument that $y_{t+1:T}$ depends on s_{t+1}, I safely cancel $p(y_{t+1:T} \mid s_t, s_{t+1}, y_{1:t}, \Phi)$ and the denominator. The conditional posterior density for s_t is

$$p(s_t \mid y_{1:T}, \Phi) = \sum_{s_{t+1}} p(s_{t+1} \mid y_{1:T}, \Phi) \frac{p(s_{t+1} \mid s_t, \Phi)p(s_t \mid y_{1:t}, \Phi)}{\sum_{s_t} p(s_{t+1} \mid s_t, \Phi)p(s_t \mid y_{1:t}, \Phi)}$$

The joint marginal density of s_t and s_{t+1} is also available in each time

$$p(s_t, s_{t+1} \mid y_{1:T}, \Phi) = p(s_t \mid s_{t+1}, y_{1:T}, \Phi)p(s_{t+1} \mid y_{1:T}, \Phi)$$

C Variational inference without restrictions

C.1 Multivariate Normal distribution

1. P

   The conditional evidence lower bound on P is

   $$\mathcal{L}^* = E[\log p(P) + \log p(S \mid P)] - E[\log q(P)]$$

   For a specific row k in P, I have the conditional evidence lower bound as

   $$\mathcal{L} = E[\log p(p_{k\cdot}) + \log p(s_1 = k) + \log p(s_t \mid s_{t-1} = k, p_{k\cdot})] - E[\log q(p_{k\cdot})]$$

Note that $\log p(s_1 = k)$ is a constant with respect to P. I can ignore this term and start from $t = 2$. The optimal solution for $q(p_{k\cdot})$ is then

$$q^*(p_{k\cdot}) \propto \exp\left\{E_{q(-p_{k\cdot})}\left[\log p(p_{k\cdot}) + \log(s_t \mid s_{t-1} = k, p_{k\cdot})\right]\right\}$$

Focus on the exponent part

$$\begin{aligned}&\log \prod_{j=1}^{K} p_{kj}^{\alpha_{kj}-1} + \log \prod_{j=1}^{K}\prod_{t=K}^{T} p_{kj}^{\mathbf{1}(s_{t-1}=k, s_t=j)} \\ &= \sum_{j=1}^{K}(\alpha_{kj}-1)\log p_{kj} + \sum_{j=1}^{K}\sum_{t=2}^{T}\mathbf{1}(s_{t-1}=k, s_t=j)\log p_{kj}\end{aligned}$$

Take the conditional expectation with the note that only S is relevant

$$\left[\sum_{j=1}^{K}(\alpha_{kj}-1) + \sum_{j=1}^{K}\sum_{t=2}^{T} E_{q(S)}[\mathbf{1}(s_{t-1}=k, s_t=j)]\right]\log p_{kj}$$

It suggests that

$$q^*(p_k) \sim \mathbb{D}ir(\overline{\alpha}_k)$$

where $\overline{\alpha}_{kj} = \alpha_{kj} + \sum_{t=2}^{T}\phi_{t-1,k,t,j}$ with $E_{q(S)}[\mathbf{1}(s_{t-1}=k, s_t=j)] = \phi_{t-1,k,t,j}$

2. M,Σ

The conditional evidence lower bound on M, Σ is

$$\mathcal{L} = E[\log p(M,\Sigma) + \log(Y \mid S, M, \Sigma)] - E[\log q(M,\Sigma)]$$

Hence, for each regime k, the evidence lower bound is

$$\mathcal{L} = E\left[\log p(M_k, \Sigma_k) + \sum_{t=1}^{T}\log p(y_t \mid S, M, \Sigma)\right] - E[\log q(M_k, \Sigma_k)]$$

The optimal solution is then

$$q^*(M_k, \Sigma_k) \propto \exp\left\{E_{q(-M,\Sigma)}\left[\log p(M_k,\Sigma_k) + \sum_{t=1}^{T}\log p(y_t \mid S, M, \Sigma)\right]\right\}$$

Focus on the exponent part

$$\begin{aligned}&\frac{N}{2}\log\left(\frac{h}{2\pi}\right) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(M_k-\mu)'h\Sigma_k^{-1}(M_k-\mu) - \frac{n+N+1}{2}\log|\Sigma_k| - \frac{1}{2}tr(\Sigma_k^{-1}\Psi) \\ &+ \sum_{t=1}^{T}\mathbf{1}(s_t=k)\left(-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right)\end{aligned}$$

Take the conditional expectation

$$\frac{N}{2}\log\left(\frac{h}{2\pi}\right) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(M_k-\mu)'h\Sigma_k^{-1}(M_k-\mu) - \frac{n+N+1}{2}\log|\Sigma_k| - \frac{1}{2}tr(\Sigma_k^{-1}\Psi)$$

$$+\sum_{t=1}^{T}\phi_{tk}\left(-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right)$$

Let denote

$$\begin{aligned}
\overline{h}_k &= h + \sum_{t=1}^{T}\phi_{tk} \\
\overline{\mu}_k &= \overline{h}_k^{-1}\left(h\mu + \sum_{t=1}^{T}\phi_{tk}y_t\right) \\
\overline{n}_k &= n + \sum_{t=1}^{T}\phi_{tk} \\
\overline{\Psi}_k &= \Psi + \sum_{t=1}^{T}\phi_{tk}y_t y_t' - \overline{h}_k\overline{\mu}_k\overline{\mu}_k' + h\mu\mu'
\end{aligned}$$

I deduce that the term inside the exponential function is proportional to

$$-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(M_k - \overline{\mu}_k)'\overline{h}_k\Sigma_k^{-1}(M_k - \overline{\mu}_k) - \frac{\overline{n} + N + 1}{2}\log|\Sigma_k| - \frac{1}{2}tr(\Sigma_k^{-1}\overline{\Psi}_k)$$

It suggests that the variational density of $q^*(M_k, \Sigma_k)$ is a Normal Inverse Wishart distribution $\mathbb{N} - \mathbb{IW}(\overline{\mu}_k, \overline{h}_k, \overline{n}_k, \overline{\Psi}_k)$.

C.2 Multivariate Student t distribution

The evidence lower bound is

$$\begin{aligned}
\mathcal{L} = E[\log p(M,\Sigma) + \log p(\nu) + \sum_{t=1}^{T}\log p(\tau_t \mid \nu) + \log p(P) + \log p(S \mid Y, P) \\
+ \sum_{t=1}^{T}\log p(y_t \mid M, \Sigma, \tau, P, S)] \\
- E[\log q(M,\Sigma) + \log q(\nu) + \sum_{t=1}^{T}\log q(\tau_t) + \log q(P) + \log q(S)]
\end{aligned}$$

1. P

This part is similar to the multivariate normal distribution.

2. M, Σ

The conditional evidence lower bound for M_k and Σ_k is

$$\mathcal{L}^* = E\left[\log p(M_k, \Sigma_k) + \sum_{t=1}^{T}\log p(y_t \mid s_t = k, M_k, \Sigma_k, \tau_t)\right] - E\left[\log q(M_k, \Sigma_k)\right]$$

The optimal solution is

$$q^*(M_k, \Sigma_k) \propto \exp\left[E_{q(S),q(\tau)}[\log p(M_k, \Sigma_k) + \sum_{t=1}^{T} \log p(y_t \mid s_t = k, M_k, \Sigma_k, \tau_t)]\right]$$

Focus on the expected value, I have

$$\begin{aligned}
&-\frac{1}{2}\log|\Sigma_k| - \frac{h}{2}(M_k-\mu)'\Sigma_k^{-1}(M_k-\mu) - \frac{n+N+1}{2}\log|\Sigma_k| - \frac{1}{2}Tr(\Psi\Sigma_k^{-1}) \\
&-\sum_{t=1}^{T}\mathbf{1}(s_t=k)\left(-\frac{N}{2}\log(\tau_{tk}) + \frac{1}{2}\log|\Sigma_k| + (y_t-M_k)'\tau_t\Sigma_k^{-1}(y_t-M_k)\right)
\end{aligned}$$

Taking the conditional expectation

$$\begin{aligned}
&-\frac{n+N+2}{2}\log|\Sigma_k| - \frac{h}{2}(M_k-\mu)'\Sigma_k^{-1}(M_k-\mu) - \frac{1}{2}Tr(\Psi\Sigma_k^{-1}) \\
&-\sum_{t=1}^{T}\phi_{tk}E_q\left[\frac{1}{2}\log|\Sigma_k| + \frac{\tau_t}{2}(y_t-M_k)'\Sigma_k^{-1}(y_t-M_k)\right]
\end{aligned}$$

Let denote

$$\begin{aligned}
\overline{h}_k &= h + \sum_{t=1}^{T}\phi_{tk}E_q\left[\tau_t\right] \\
\overline{m}_k &= \overline{h}^{-1}\left(h\mu + \sum_{t=1}^{T}\phi_{tk}E_q\left[\tau_t\right]y_t\right) \\
\overline{n}_k &= n + \sum_{t=1}^{T}\phi_{tk} \\
\overline{\Psi}_k &= \Psi + \sum_{t=1}^{T}\phi_{tk}E_q\left[\tau_t\right]y_ty_t' - \overline{h}\overline{\mu}_k\overline{\mu}_k' + h\mu\mu'
\end{aligned}$$

I deduce that the term inside the exponential function is proportional to

$$-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(M_k-\overline{\mu}_k)'\overline{h}_k\Sigma_k^{-1}(M_k-\overline{\mu}_k) - \frac{\overline{n}_k+N+1}{2}\log|\Sigma_k| - \frac{1}{2}tr(\Sigma_k^{-1}\overline{\Psi}_k)$$

It suggests that the variational density of $q^*(M_k, \Sigma_k)$ is a Normal Inverse Wishart distribution $\mathbb{N}-\mathbb{IW}(\overline{\mu}_k, \overline{h}_k, \overline{n}_k, \overline{\Psi}_k)$.

3. τ

The conditional evidence lower bound for τ_t is

$$\mathcal{L}^* = E[\log p(\tau_t \mid \nu_k, s_t = k) + \log p(y_t \mid M_k, \Sigma_k, \tau_t)] - E[\log q(\tau_t)]$$

The optimal solution is

$$q^*(\tau_t) \propto \exp\left[E[\log p(\tau_t \mid \nu_k, s_t = k) + \log p(y_t \mid M_k, \Sigma_k, \tau_t)]\right]$$

Focus on the expected value, I have

$$\begin{aligned}&\mathbf{1}(s_t = k)\Bigg(\frac{\nu_k}{2}\log\left(\frac{\nu_k}{2}\right) - \log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2} - 1\right)\log\tau_t - \frac{\nu_k}{2}\tau_t \\ &+ \frac{N}{2}\log\tau_t - \frac{\tau_t}{2}(y_t - M_k)'\Sigma^{-1}(y_t - M_k)\Bigg)\end{aligned}$$

Taking the conditional expectation, I find it safe to remove the constant relating to the state

$$\begin{aligned}&\left(\frac{E(\nu_k)}{2} - 1\right)\log\tau_t - \frac{E(\nu_k)}{2}\tau_t + \frac{N}{2}\log\tau_t - E\left[\frac{\tau_t}{2}(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right] \\ &= \left(a + \frac{N}{2} - 1\right)\log\tau_t - \left(b + E\left[\frac{1}{2}(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right]\right)\tau_t\end{aligned}$$

This functional form implies $q^*(\tau_t) \sim G\left(\frac{E(\nu_k)}{2} + \frac{N}{2}, \frac{E(\nu_k)}{2} + \frac{1}{2}E\left[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right]\right)$. Conditional expectations relating to τ_t must be scaled. We scale $E[\tau_t]$ in order that the average of this expectation under state k is one. We also scale the $E[\log\tau_t]$. However, this task is straight forward since it is equivalent to adding a constant to the unscaled value.

4. ν

The conditional evidence lower bound for ν_k is

$$\mathcal{L}^* = E_q[\log p(\nu_k) + \sum_{t=1}^{T}\log p(\tau_t \mid \nu_k, s_t = k)] - E_q[\log q(\nu_k)]$$

The optimal solution is

$$q^*(\nu_k) \propto \exp\left[E_q[\log p(\nu_k) + \sum_{t=1}^{T}\log p(\tau_t \mid \nu_k, s_t = k)]\right]$$

Focus on the expected value, I have

$$(v_\nu - 1)\log\nu_k - \nu_k s_\nu + \sum_{t=1}^{T}\mathbf{1}(s_t = k)\left(-\log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2}\left(\frac{\nu_k}{2}\right) - \nu_k\left(\frac{\tau_t - \log\tau_t}{2}\right)\right)$$

Taking the conditional expectation

$$(v_\nu - 1)\log\nu_k - \nu_k s_\nu + \sum_{t=1}^{T}\phi_{tk}\left(-\log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2}\left(\frac{\nu_k}{2}\right) - \nu_k\left(\frac{E_{q(\tau)}\left(\tau_t - \log\tau_t\right)}{2}\right)\right)$$

The work encounters difficulty at this point. Because there is no exponential conjugacy in this solution, I cannot update the variational parameters straightforward. Unlike the MCMC algorithm, there is no room for the use of Metropolist-Hastings rejection method.

In this part, I introduce three approaches to find the degrees of freedom.

- Conditional calibration

I estimate sample degrees of freedom conditional on other parameters. I use the variance formula to recover the degrees of freedom.

$$V_k = \frac{\nu_k}{\nu_k - 2}\Sigma_k \tag{16}$$

In particular, I find the sample variance covariance matrix, $\widehat{V}$, the scale matrix, Σ, and then recover the degrees of freedom, ν. Note that two sides of the equation are two matrices. I find the value of ν that minimises the Frobenius matrix distance between two matrices.

Suppose that there are two positive definite matrices X and Y that have the same dimension $M \times N$. Their Frobenius matrix distance is

$$F = \sqrt{\sum_i^M \sum_j^N |a_{ij}|^2} = \sqrt{tr(AA')}$$

where a_{ij} is the element on row i^{th} and column j^{th} of matrix $A = X - Y$ with dimension $M \times N$.

Provided that sample variance coveriance matrix, V_k, and the expected value of the scale Σ_k, it is feasible to recover ν_k. Note that ν_k must be greater than 2. I update the value of degrees of freedom after several iterations. This approach can be denoted as VI-CC because it is equivalent to MCMC-CC.

- Approximation

  This approaches is proposed by Chatzis et al. (2008). I use the Stirling's approximation for $\log\Gamma\left(\frac{\nu_k}{2}\right)$ in order to have my proposed prior as a conjugate prior. Therefore

$$\begin{aligned}&(v_\nu - 1)\log(\nu_k) - \nu_k s_\nu + \sum_{t=1}^{T} \phi_{tk}\left(-\left[\left(\frac{\nu_k}{2} - \frac{1}{2}\right)\log\left(\frac{\nu_k}{2}\right) - \left(\frac{\nu_k}{2}\right) + \frac{1}{2}\log 2\pi\right]\right)\\&+ \sum_{t=1}^{T} \phi_{tk}\left(\frac{\nu_k}{2}\log\left(\frac{\nu_k}{2}\right) - \nu_k\left(\frac{E_q\left(\tau_t - \log\tau_t\right)}{2}\right)\right)\\&\doteq (v_\nu - 1)\log\left(\nu_k\right) - \nu_k s_\nu + \sum_{t=1}^{T} \phi_{tk}\left(\frac{1}{2}\log\left(\frac{\nu_k}{2}\right) - \nu_k\left(\frac{E_q\left(\tau_t - \log\tau_t\right)}{2} - \frac{1}{2}\right)\right)\end{aligned}$$

  It suggests that the variational density as

$$\nu_k \sim \mathbb{G}\left(v_\nu + \frac{\sum_{t=1}^T \phi_{tk}}{2}, s_\nu + \frac{\sum_{t=1}^T \phi_{tk} E_q\left(\tau_t - \log\tau_t\right)}{2} - \frac{\sum_{t=1}^T \phi_{tk}}{2}\right)$$

  Let denote this approach as VI-A that is comparable to MCMC-A.

- Proposing a Gamma distribution of $q(\nu)$

  I propose $q^*(\nu) \sim \mathbb{G}(\overline{v}_\nu, \overline{s}_\nu)$. We find $\overline{v}_\nu$ and $\overline{s}_\nu$ that minimize the distance

$$E[\log q^*(\nu_k)] - E[\log p(\nu_k) + \sum_{t=1}^{T} \log p(\tau_t \mid \nu_k, s_t = k)]$$

The first term is

$$-\log\Gamma\left(\overline{v}_\nu\right)+\overline{v}_\nu\log(\overline{s}_\nu)+(\overline{v}_\nu-1)E[\log\nu_k]-\overline{s}_\nu E[\nu_k]$$

The second term is

$$\begin{aligned}&-\log\Gamma\left(v_\nu\right)+v_\nu\log(s_\nu)+(v_\nu-1)E[\log\nu_k]-s_\nu E[\nu_k]\\+\sum_{t=1}^{T}\phi_{tk}&\left[-\log\Gamma\left(\frac{\nu_k}{2}\right)+\frac{\nu_k}{2}\log\left(\frac{\nu_k}{2}\right)+\left(\frac{\nu_k}{2}-1\right)E[\log\tau_t]-\frac{\nu_k}{2}E[\tau_t]\right]\end{aligned}$$

Since I assume the Student t distribution has the second moment, its degrees of freedom must be greater than 2. Only the result satisfying this condition will be kept. It is not necessary to update ν every iteration. I can update other parameters for 20 iterations, then update hyperparameters for ν. It will save my computational cost. I denote this approach as VI.

C.3 S

The conditional evidence lower bound on S is

$$\mathcal{L}^*=E[\log p(S,Y\mid\Phi)]-E[\log q(S)]$$

The optimal solution is as

$$\begin{aligned}q^*(S_{1:T})&=\exp\left\{E_{q(\Phi)}\left[\log p(S_{1:T},y_{1:T}\mid\Phi)-\log p(y_{1:T}\mid\Phi)\right]\right\}\\&\propto\exp\left\{E_{q(\Phi)}\left[\log p(S_{1:T},y_{1:T}\mid\Phi)\right]\right\}\end{aligned}$$

I make a comparison between the complete log-likelihood of the conditional posterior in Equation 5 with the complete log-likelihood of the variational density, $q^*(S_{1:T})$. First, the log of the complete likelihood in Equation 5 is proportional to

$$\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\mathbf{1}(s_t=k,s_{t-1}=j)\log p_{jk}+\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbf{1}(s_t=k)\log p(y_t\mid\Phi,s_t)$$

Second, the log-likelihood of the optimal solution from the VI where I take the conditional expectation with respect to Φ is

$$\begin{aligned}&E_{q(\Phi)}\left[\sum_{t=2}^{T}\log p(s_t\mid s_{t-1},P)\right]+E_{q(\Phi)}\left[\sum_{t=1}^{T}\log p(y_t\mid s_t,\Phi)\right]\\=&\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\mathbf{1}(s_{t-1}=k,s_t=j)E_{q(\Phi)}[\log p_{kj}]+\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbf{1}(s_t=k)E_{q(\Phi)}[\log p(y_t\mid\Phi,s_t)]\end{aligned}$$

The log of the complete likelihood in Equation 5 is similar to the loglikelihood of the optimal solution. It suggests that the optimal solution shares the same type of density with the conditional posterior. However, their parameters are not the same. For example: the parameters for the transition matrix in the conditional posterior density is p_{kj} and in the optimal solution through VI for $q(p_{kj})$ is $\exp(E_{q(p_{kj})}[\log p_{kj}])$. Let denote

$$\widetilde{p}_{kj} = \exp(E_{q(\Phi)}[\log p_{kj}]) \quad \widetilde{p}(y_t \mid M_k, \Sigma_k) = \exp(E_{q(\Phi)}[\log p(y_t \mid \Phi)])$$

with the first equation involving normalisation. These newly defined parameters are already found in the previous steps. Hence, I can find $q(s_t)$ in the same way I find $p(s_t \mid y_{1:T}, \Phi)$, i.e. through forward filtering backward smoothing algorithm.

In order to find the marginal variational density of each s_t, $q^*(s_t)$, I note that the structure of this variational distribution is similar to the conditional posterior distribution, $p(S \mid Y, \Phi)$. Therefore, I can write that

$$\begin{aligned} q^*(S_{1:T}) &\propto \prod_{t=2}^{T}\prod_{j=1}^{K}\prod_{k=1}^{K} \widetilde{p}_{jk}^{\mathbf{1}(s_t=k, s_{t-1}=j)} \prod_{t=1}^{T}\prod_{k=1}^{K} \widetilde{p}(y_t \mid \Phi, s_t)^{1(s_t=k)} \\ &\propto \widetilde{p}(y_1 \mid s_1, \Phi) \prod_{t=2}^{T} \widetilde{p}(s_t \mid s_{t-1}, P)\widetilde{p}(y_t \mid s_t, \Phi) \end{aligned}$$

This expression is similar to Equation 5, the complete likelihood of the conditional posterior distribution of S. Because the conditional posterior distribution, $p(s_t \mid Y, \Phi)$, is estimated through the forward-backward algorithm, I can do the same to find the variational density, $q^*(s_t)$, at each t.

I use the same Chib (1996)'s algorithm as in MCMC. For any t, the variational forward filtering is

$$\widetilde{F}(s_t) \propto \frac{\widetilde{p}(y_t \mid s_t, \Phi) \sum_{s_{t-1}} \widetilde{p}(s_t \mid s_{t-1}, P)\widetilde{F}(s_{t-1})}{\sum\limits_{s_t} \widetilde{p}(y_t \mid s_t, \Phi) \sum_{s_{t-1}} \widetilde{p}(s_t \mid s_{t-1}, P)\widetilde{F}(s_{t-1})}$$

with the normalizing constant $\widetilde{p}(y_t \mid y_{1:t-1}, M, \Sigma)$.

The forward pass at time $t = 1$ for this optimal solution is

$$\widetilde{F}(s_1) = \frac{\widetilde{p}(s_1)\widetilde{p}(y_1 \mid s_1, \Phi)}{\sum_{s_1} \widetilde{p}(s_1)\widetilde{p}(y_1 \mid s_1, \Phi)}$$

with the normalizing constant $\widetilde{p}(y_1 \mid \Phi)$.

I follow the backward algorithm in Equation 15 to smooth the variational density of states as follows

$$q^*(s_t) \propto \widetilde{p}(s_t \mid y_{1:T}, \Phi) = \sum_{s_{t+1}} \widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi)\widetilde{p}(s_{t+1} \mid y_{1:T}, \Phi)$$

with the first term

$$\widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi) = \frac{\widetilde{p}(s_{t+1} \mid s_t, \Phi)\widetilde{p}(s_t \mid y_{1:t}, \Phi)}{\sum_{s_t} \widetilde{p}(s_{t+1} \mid s_t, \Phi)\widetilde{p}(s_t \mid y_{1:t}, \Phi)}$$

Similarly, the temporal transition between two states is

$$q^*(s_t, s_{t+1}) \propto \widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi)\widetilde{p}(s_{t+1} \mid y_{1:T}, \Phi)$$

To find the likelihood of the data, I note

$$\exp(E_{q(\Phi)}[\log p(Y \mid \Phi)]) = \widetilde{p}(y_{1:T} \mid \Phi) = \widetilde{p}(y_1 \mid \Phi) \prod_{t=2}^{T} \widetilde{p}(y_t \mid y_{1:t-1}, \Phi) \tag{17}$$

On the right hand side, each factor in this product is a normalizing constant of its corresponding forward pass. Therefore I store all these constants to calculate the log likelihood of in the $\mathcal{L}$.

It is important to note that I must maintain a numerical stability. The reason lies in the use of likelihood and joint likelihood. These values are extremely small and easy to be out of the range in any computing programs. Hence, I take log of these likelihoods and their relevant terms to avoid inaccurate approximation as in Appendix F. In addition, my algorithm will keep the log value until the last step before reverting them by exponential function.

C.4 Variational forward filtering backward smoothing

The forward pass at time $t = 1$ for this optimal solution is

$$\widetilde{F}(s_1) = \frac{\widetilde{p}(s_1)\widetilde{p}(y_1 \mid s_1, M, \Sigma)}{\sum_{s_1} \widetilde{p}(s_1 \mid)\widetilde{p}(y_1 \mid s_1, M, \Sigma)}$$

with the normalizing constant $\widetilde{p}(y_1 \mid M, \Sigma)$. It is necessary to normalize this forward pass before applying it into the next period. Then, for any t, the forward filtering is

$$\widetilde{F}(s_t) \propto \frac{\widetilde{p}(y_t \mid s_t, \Phi) \sum_{s_{t-1}} \widetilde{p}(s_t \mid s_{t-1}, P)\widetilde{F}(s_{t-1})}{\sum\limits_{s_t} \widetilde{p}(y_t \mid s_t, \Phi) \sum_{s_{t-1}} \widetilde{p}(s_t \mid s_{t-1}, P)\widetilde{F}(s_{t-1})}$$

where

$$E_{q(\Phi)}[\log \widetilde{p}(y_t \mid s_t = k, \Phi)] = -\frac{1}{2}E[\log |\Sigma_k|] - \frac{1}{2}E[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)] + const$$

and the normalising constant $\widetilde{p}(y_t \mid y_{1:t-1}, M, \Sigma)$.

Then, at time T, the optimal variational density is

$$q^*(s_T) = \widetilde{F}(s_T) = \frac{\widetilde{p}(y_t \mid s_T, M, \Sigma) \sum\limits_{s_{T-1}} \widetilde{F}(s_{T-1})\widetilde{p}(s_T \mid s_{T-1}, P)}{\sum\limits_{s_T} \widetilde{p}(y_t \mid s_T, M, \Sigma) \sum\limits_{s_{T-1}} \widetilde{F}(s_{T-1})\widetilde{p}(s_T \mid s_{T-1}, P)}$$

I also follow Chib (1996)'s backward algorithm in Equation 15 to smooth the state variational density. The other variational densities at other times can be found backwardly as follows

$$q^*(s_t) \propto \widetilde{p}(s_t \mid y_{1:T}, \Phi) = \sum_{s_{t+1}} \widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi)\widetilde{p}(s_{t+1} \mid y_{1:T}, \Phi)$$

with the first term

$$\widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi) = \frac{\widetilde{p}(s_{t+1} \mid s_t, \Phi)\widetilde{p}(s_t \mid y_{1:t}, \Phi)}{\sum_{s_t} \widetilde{p}(s_{t+1} \mid s_t, \Phi)\widetilde{p}(s_t \mid y_{1:t}, \Phi)}$$

Similarly, the temporal transition between two states is

$$q^*(s_t, s_{t+1}) \propto \widetilde{p}(s_t \mid s_{t+1}, y_{1:T}, \Phi)\widetilde{p}(s_{t+1} \mid y_{1:T}, \Phi)$$

D Expectation in variational inference

D.1 Expected values for multivariate Normal distribution

To estimate the evidence lower bound, several expectations requires to be computed.

- $E_{q(\Phi)}[\log p_{kj}] = E_{q(P)}[\log p_{kj}] = \psi(\alpha_{kj}) - \psi\left(\sum_{j=1}^K \alpha_{kj}\right)$
- $E_{q(\Phi)}[\log p(y_t \mid s_t = k, \Phi)] = -\frac{1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(M,\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)] + const$

  where

$$\begin{aligned} E_{q(\Sigma)}[\log|\Sigma_k|] &= -E_{q(\Sigma)}[\log|\Sigma_k^{-1}|] \\ &= -N\log 2 + \log|\overline{\Psi}_k| - \sum_{i=1}^N \psi\left(\frac{\overline{n}_k + 1 - i}{2}\right) \\ E_{q(\Sigma)}[\Sigma_k^{-1}] &= \overline{n}_k\overline{\Psi}_k^{-1} \\ E_{q(M,\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)] &= \frac{N}{\overline{h}_k} + \overline{n}_k(y_t - \overline{\mu}_k)'\overline{\Psi}_k^{-1}(y_t - \overline{\mu}_k) \end{aligned}$$

  The last line is in Appendix E.2

- $E[\log p(p_{kj})] = \log\Gamma\left(\sum\limits_{j=1}^K \alpha_{kj}\right) - \sum\limits_{j=1}^K \log\Gamma(\alpha_{kj}) + \sum\limits_{l=1}^K(\alpha_{kl} - 1)E_{q(P)}[\log p_{kl}]$
- $E_q[\log p(M_k, \Sigma_k)]$

  That is

$$\begin{aligned} &\frac{N}{2}\log\left(\frac{h}{2\pi}\right) - \frac{1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(M,\Sigma)}[(M_k - \mu)'h\Sigma_k^{-1}(M_k - \mu)] \\ &+ \frac{n}{2}\log|\Psi| - \frac{nN}{2}\log 2 - \frac{N(N-1)}{4}\log\pi \\ &- \sum_{i=1}^n \log\Gamma\left(\frac{n+1-i}{2}\right) - \frac{n+N+1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(\Sigma)}[tr(\Psi\Sigma_k^{-1})] \\ &\doteq -\frac{1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(M,\Sigma)}[(M_k - \mu)'h\Sigma^{-1}(M_k - \mu)] \\ &- \frac{n+N+1}{2}E_{q(\Sigma)}[\log|\Sigma|] - \frac{1}{2}E_{q(\Sigma)}[tr(\Psi\Sigma^{-1})] \end{aligned}$$

  where

$$\begin{aligned} E_{q(M,\Sigma)}[(M_k - \mu)'h\Sigma_k^{-1}(M_k - \mu)] &= \frac{Nh}{\overline{h}_k} + (\overline{\mu}_k - \mu)'h\overline{n}_k\overline{\Psi}_k^{-1}(\overline{\mu}_k - \mu) \\ E_{q(\Sigma)}[tr(\Psi\Sigma^{-1})] &= \overline{n}_k tr(\Psi\overline{\Psi}_k^{-1}) \end{aligned}$$

- $E_q[\log q(p_k)]$

  That is the negative entropy of p_k

$$\begin{aligned}
&\sum_{k=1}^{K} \log\Gamma(\overline{\alpha}_k) - \Gamma\log\left(\sum_{k=1}^{K}\overline{\alpha}_k\right) + \left(\sum_{k=1}^{K}\overline{\alpha}_k - 2\right)\psi\left(\sum_{k=1}^{K}\overline{\alpha}_k\right) - \sum_{k=1}^{K}(\overline{\alpha}_k - 1)\psi\left(\overline{\alpha}_k\right)\\
&= \sum_{k=1}^{K} \log\Gamma(\overline{\alpha}_k) - \psi\left(\sum_{k=1}^{K}\overline{\alpha}_k\right) - \sum_{k=1}^{K}(\overline{\alpha}_k - 1)\psi\left(\overline{\alpha}_k\right)
\end{aligned}$$

- $E_q[\log q(M_k, \Sigma_k)]$

  That is

$$\begin{aligned}
&\frac{N}{2}\log\left(\frac{\overline{h}_k}{2\pi}\right) - \frac{1}{2}E[\log|\Sigma_k|] - \frac{1}{2}E[(M_k - \overline{\mu}_k)'\overline{h}_k\Sigma_k^{-1}(M_k - \overline{\mu}_k)]\\
&+ \frac{N}{2}\log\left|\overline{\Psi}_k\right| - \frac{\overline{n}_k N}{2}\log 2 - \frac{N(N-1)}{4}\log\pi\\
&- \sum_{i=1}^{N}\log\Gamma\left(\frac{\overline{n}_k + 1 - i}{2}\right) - \frac{\overline{n}_k N}{2} - \frac{\overline{n}_k + N + 1}{2}E[\log|\Sigma|]\\
&= \frac{N}{2}\log\left(\frac{\overline{h}_k}{2\pi}\right) - \frac{1}{2}E[\log|\Sigma_k|] - \frac{N}{2}\\
&+ \frac{N}{2}\log\left|\overline{\Psi}_k\right| - \frac{\overline{n}_k N}{2}\log 2 - \frac{N(N-1)}{4}\log\pi\\
&- \sum_{i=1}^{N}\log\Gamma\left(\frac{\overline{n}_k + 1 - i}{2}\right) - \frac{\overline{n}_k N}{2} - \frac{\overline{n}_k + N + 1}{2}E[\log|\Sigma|]
\end{aligned}$$

D.2 Expected values for multivariate Student t distribution

To estimate the evidence lower bound, several expectations requires to be computed.

- $E_{q(\Phi)}[\log p_{kj}] = E_{q(P)}[\log p_{kj}] = \psi(\alpha_{kj}) - \psi\left(\sum_{j=1}^{K}\alpha_{kj}\right)$

- $E_{q(\Phi)}[\log p(y_t \mid s_t = k, \Phi)] = \frac{N}{2}E_{q(\tau)}[\log\tau_t] - \frac{1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(M,\Sigma,\tau)}[(y_t - M_k)'\tau_t\Sigma_k^{-1}(y_t - M_k)] + const$

  where

$$\begin{aligned}
E_q(\tau_t) &= \frac{E_q(\nu_k) + N}{E_q(\nu_k) + E_q\left[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right]}\\
E_q\left[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right] &= \frac{N}{\overline{h}_k} + \overline{n}_k(y_t - \overline{\mu}_k)'\overline{\Psi}_k^{-1}(y_t - \overline{\mu}_k)\\
E_q[\nu_k] &= \frac{\overline{v}_{\nu,k}}{\overline{s}_{\nu,k}}\\
E_{q(\tau_t)}\left(\log\tau_t\right) &= -\log\left\{b + E_q\left[\frac{1}{2}(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right]\right\}\\
&+ \psi\left(a + \frac{N}{2}\right)
\end{aligned}$$

$$\begin{aligned}
E_{q(\Sigma)}[\log|\Sigma_k|] &= -E_{q(\Sigma)}[\log|\Sigma_k^{-1}|] \\
&= -N\log 2 + \log|\overline{\Psi}_k| - \sum_{i=1}^{N}\psi\left(\frac{\overline{n}_k+1-i}{2}\right) \\
E_{q(\Sigma)}[\Sigma_k^{-1}] &= \overline{n}_k\overline{\Psi}_k^{-1} \\
E_{q(M,\Sigma,\tau)}[(y_t-M_k)'\tau_t\Sigma_k^{-1}(y_t-M_k)] &= \frac{NE_q(\tau_t)}{\overline{h}_k} + \overline{n}_k(y_t-\overline{\mu}_k)'E_q(\tau_t)\overline{\Psi}_k^{-1}(y_t-\overline{\mu}_k)
\end{aligned}$$

The last line is in Appendix E.2.

- $E_q[\log p(p_{kj})] = \log\Gamma\left(\sum_{j=1}^{K}\alpha_{kj}\right) - \sum_{j=1}^{K}\log\Gamma(\alpha_{kj}) + \sum_{l=1}^{K}(\alpha_{kl}-1)E_{q(P)}[\log p_{kl}]$
- $E_q[\log p(\tau_t)] = (a-1)E_q[\log\tau_t] - bE_q[\tau_t]$
- $E_q[\log p(\nu_k)] = (v_\nu - 1)E_q[\log\nu_k] - s_\nu E_q[\nu_k]$ where

$$\begin{aligned}
E_q[\log\nu_k] &= -\log\overline{s}_{\nu,k} + \psi(\overline{v}_{\nu,k}) \\
E_q[\nu_k] &= \frac{\overline{v}_{\nu,k}}{\overline{s}_{\nu,k}}
\end{aligned}$$

- $E_q[\log p(M_k,\Sigma_k)]$

  That is

$$\begin{aligned}
&\frac{N}{2}\log\left(\frac{h}{2\pi}\right) - \frac{1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(M,\Sigma)}[(M_k-\mu)'h\Sigma_k^{-1}(M_k-\mu)] \\
&+\frac{n}{2}\log|\Psi| - \frac{nN}{2}\log 2 - \frac{N(N-1)}{4}\log\pi \\
&-\sum_{i=1}^{n}\log\Gamma\left(\frac{n+1-i}{2}\right) - \frac{n+N+1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(\Sigma)}[tr(\Psi\Sigma_k^{-1})] \\
&\doteq -\frac{1}{2}E_{q(\Sigma)}[\log|\Sigma_k|] - \frac{1}{2}E_{q(M,\Sigma)}[(M_k-\mu)'h\Sigma_k^{-1}(M_k-\mu)] \\
&-\frac{n+N+1}{2}E_{q(\Sigma)}[\log|\Sigma|] - \frac{1}{2}E_{q(\Sigma)}[tr(\Psi\Sigma^{-1})]
\end{aligned}$$

  where

$$\begin{aligned}
E_{q(M,\Sigma)}[(M_k-\mu)'h\Sigma_k^{-1}(M_k-\mu)] &= \frac{Nh}{\overline{h}_k} + (\overline{\mu}_k-\mu)'h\overline{n}_k\overline{\Psi}_k^{-1}(\overline{\mu}_k-\mu) \\
E_{q(\Sigma)}[tr(\Psi\Sigma^{-1})] &= \overline{n}_k tr(\Psi\overline{\Psi}_k^{-1})
\end{aligned}$$

- $E_q[\log q(p_k)]$

  That is the negative entropy of p_k

$$\begin{aligned}
&\sum_{k=1}^{K}\log\Gamma(\overline{\alpha}_k) - \Gamma\log\left(\sum_{k=1}^{K}\overline{\alpha}_k\right) + \left(\sum_{k=1}^{K}\overline{\alpha}_k - 2\right)\psi\left(\sum_{k=1}^{K}\overline{\alpha}_k\right) - \sum_{k=1}^{K}(\overline{\alpha}_k-1)\psi\left(\overline{\alpha}_k\right) \\
&= \sum_{k=1}^{K}\log\Gamma(\overline{\alpha}_k) - \psi\left(\sum_{k=1}^{K}\overline{\alpha}_k\right) - \sum_{k=1}^{K}(\overline{\alpha}_k-1)\psi\left(\overline{\alpha}_k\right)
\end{aligned}$$

- $E_q[\log q(M_k, \Sigma_k)]$

  That is

$$\begin{aligned}
&\frac{N}{2}\log\left(\frac{\overline{h}_k}{2\pi}\right) - \frac{1}{2}E[\log|\Sigma_k|] - \frac{1}{2}E[(M_k-\overline{\mu}_k)'\overline{h}_k\Sigma_k^{-1}(M_k-\overline{\mu}_k)] \\
&+ \frac{N}{2}\log\left|\overline{\Psi}_k\right| - \frac{\overline{n}_k N}{2}\log 2 - \frac{N(N-1)}{4}\log\pi - \sum_{i=1}^{N}\log\Gamma\left(\frac{\overline{n}_k+1-i}{2}\right) \\
&- \frac{\overline{n}_k N}{2} - \frac{\overline{n}_k+N+1}{2}E[\log|\Sigma|] \\
&= \frac{N}{2}\log\left(\frac{\overline{h}_k}{2\pi}\right) - \frac{1}{2}E[\log|\Sigma_k|] - \frac{N}{2} \\
&+ \frac{N}{2}\log\left|\overline{\Psi}_k\right| - \frac{\overline{n}_k N}{2}\log 2 - \frac{N(N-1)}{4}\log\pi - \sum_{i=1}^{N}\log\Gamma\left(\frac{\overline{n}_k+1-i}{2}\right) \\
&- \frac{\overline{n}_k N}{2} - \frac{\overline{n}_k+N+1}{2}E[\log|\Sigma|]
\end{aligned}$$

- $E_q[\log q(\tau_t)] = \left(a + \frac{N}{2}\right)E_q[\log\tau_t] - \left(b + E_q\left[\frac{1}{2}(y_t-M_k)'\Sigma_k^{-1}(y_t-M_k)\right]\right)E_q[\tau_t]$

- $E_q[\log q(\nu_k)] = \left(v_\nu + \frac{\sum_{t=1}^{T}\phi_{tk}}{2} - 1\right)E_q[\log\nu_k] - \left(s_\nu + \frac{\sum_{t=1}^{T}\phi_{tk}E_q\left(\tau_t - \log\tau_t\right)}{2} - \frac{\sum_{t=1}^{T}\phi_{tk}}{2}\right)E_q[\nu]$

E Variational inference with restrictions

E.1 Optimal variational density

Regarding M, restrictions enter its prior. Distribution of M is a truncated distribution. Particularly, the probability density function $p(M)$ needs to consider a new normalizing constant. Hence, it impacts on the approximation of $q(M, \Sigma)$ so that I re-derive its optimal solution. I focus on state 1, and state 2 should follow the same direction. Its exponent part is

$$\begin{aligned}
&\frac{N}{2}\log\left(\frac{h}{2\pi}\right) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(M_1-\mu)'h\Sigma_1^{-1}(M_1-\mu) - \frac{n+N+1}{2}\log|\Sigma_1| - \frac{1}{2}tr(\Sigma_1^{-1}\Psi) \\
&+ \sum_{t=1}^{T}\mathbf{1}(s_t=1)\left(-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(y_t-M_1)'\Sigma_k^{-1}(y_t-M_1)\right) + \log I(\iota' M_1 < 0)
\end{aligned}$$

where the last term indicates a log of a constant relating to M_1. I find this last term as a represent of the new normalizing constant.

Then the conditional expectation is as

$$\begin{aligned}
&\frac{N}{2}\log\left(\frac{h}{2\pi}\right) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(M_1-\mu)'h\Sigma_1^{-1}(M_1-\mu) - \frac{n+N+1}{2}\log|\Sigma_1| - \frac{1}{2}tr(\Sigma_1^{-1}\Psi) \\
&+ \sum_{t=1}^{T}\phi_{t1}\left(-\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(y_t-M_1)'\Sigma_1^{-1}(y_t-M_1)\right) + \log I(\iota' M_1 < 0)
\end{aligned}$$

Therefore, restrictions on the M will result in the variational density as

$$q(M_1 \mid \Sigma_1, \cdot) = f_N(\overline{\mu}_1, \overline{h}_1, \Sigma_1) I\,(\iota' M_1 < 0) \qquad q(M_2 \mid \Sigma_2 \cdot) = f_N(\overline{\mu}_2, \overline{h}_2, \Sigma_2) I\,(\iota' M_2 > 0)$$

Under this optimal solution, the terms, $E_{q(M,\Sigma)}[\log p(M, \Sigma)]$, $E_{q(M,\Sigma)}[\log q(M, \Sigma)]$ and $E_{q(\Phi)}[\log p(y_t \mid s_t = k, \Phi)]$, need to be re-evaluated. For example, the issue lies on the expression, $E_{q(M,\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)]$. Under the restrictions, I derive it as

$$E_{q(M,\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)] = tr(\overline{h}_k^{-1} E[\Sigma_k^{-1}]\Sigma_k^*) + \overline{n}_k(y_t - \overline{\mu}_k^*)'\overline{\Psi}_k^{-1}(y_t - \overline{\mu}_k^*)$$

The restriction impacts on the mean of $q(M \mid \Sigma)$. Therefore, I must find the mean $\overline{\mu}_k^*$ and the variance Σ_k^* instead. It can be done via sampling as in Botev (2017). Other derivations can be found in Appendix E.2. Note that I must find the mean and variance of M_k before any further updating step.

Regarding S, I do not need to revise the previous solution of the optimal density. However, I need to use the new expectations relating to the truncated distribution. These expected values are similar to ones I find in the case of M.

E.2 Expectations under restrictions

I derive the expectation with multivariate Normal and Student t distribution with and without restrictions. I start with the expectation without restriction for the Normal distribution case as follows

$$E_{q(M,\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)] = E_{q(\Sigma)}\left\{E_{q(M|\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)]\right\}$$

Since $M_k \sim q^*(M_k \mid \Sigma_k) = N(\overline{\mu}_k, \overline{h}_k^{-1}\Sigma_k)$, I re-write the term inside the bracket as

$$\begin{aligned}
&E_{q(M|\Sigma)}[(y_t - \overline{\mu}_k + \overline{\mu}_k - M_k)'\Sigma_k^{-1}(y_t - \overline{\mu}_k + \overline{\mu}_k - M_k)] \\
&= E_{q(M|\Sigma)}[(y_t - \overline{\mu}_k)'\Sigma_k^{-1}(y_t - \overline{\mu}_k)] + 2E_{q(M|\Sigma)}[(y_t - \overline{\mu}_k)'\Sigma_k^{-1}(\overline{\mu}_k - M_k)] \\
&+ E_{q(M|\Sigma)}[(\overline{\mu}_k - M_k)'\Sigma_k^{-1}(\overline{\mu}_k - M_k)]
\end{aligned}$$

The first term on the right hand side does not have M. Hence, it is constant with respect to the expectation. The second term is as follows

$$2E_{q(M|\Sigma)}[(y_t - \overline{\mu}_k)'\Sigma_k^{-1}(\overline{\mu}_k - M_k)] = 2(y_t - \overline{\mu}_k)'\Sigma_k^{-1}(\overline{\mu}_k - E_{q(M|\Sigma)}[M_k]) = 0$$

The last term is

$$\begin{aligned}
E_{q(M|\Sigma)}[(\overline{\mu}_k - M_k)'\Sigma_k^{-1}(\overline{\mu}_k - M_k)] &= E_{q(M|\Sigma)}(tr[(\overline{\mu}_k - M_k)'\Sigma_k^{-1}(\overline{\mu}_k - M_k)]) \\
&= tr(E_{q(M|\Sigma)}[\Sigma_k^{-1}(M_k - \overline{\mu}_k)'(M_k - \overline{\mu}_k)]) \\
&= tr(E_{q(M|\Sigma)}[\Sigma_k^{-1}\overline{h}_k^{-1}\Sigma_k])
\end{aligned}$$

$$= N\overline{h}_k^{-1}$$

To conclude, I receive

$$\begin{aligned} E_{q(M,\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)] &= (y_t - \overline{\mu}_k)'E_{q(\Sigma)}\left\{\Sigma_k^{-1}\right\}(y_t - \overline{\mu}_k) + N\overline{h}_k^{-1} \\ &= (y_t - \overline{\mu}_k)'\overline{n}_k\overline{\Psi}_k^{-1}(y_t - \overline{\mu}_k) + N\overline{h}_k^{-1} \end{aligned}$$

Next, I derive the same expectation but under restriction. Under the restriction on states, I have $M_k \sim q^*(M_k \mid \Sigma_k) = N(\overline{\mu}_k, \overline{h}_k^{-1}\Sigma_k)I(M_k > 0)$, for example. Let $\overline{\mu}_k^* = E(M_k)$ and $\Sigma_k^* = Var(M_k)$. Then I apply the same above approach to get

$$\begin{aligned} E_{q(\Sigma)}\left\{E_{q(M|\Sigma)}[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)]\right\} &= E_{q(\Sigma)}\left\{(y_t - \overline{\mu}_k^*)'\Sigma_k^{-1}(y_t - \overline{\mu}_k^*) + tr(\Sigma_k^{-1}\Sigma_k^*)\right\} \\ &= (y_t - \overline{\mu}_k^*)'E_{q(\Sigma)}[\Sigma_k^{-1}](y_t - \overline{\mu}_k^*) + tr(E_{q(\Sigma)}[\Sigma_k^{-1}]\Sigma_k^*) \end{aligned}$$

Similarly, I have the restricted expectations under multivariate Normal distribution

$$\begin{aligned} E_{q(M,\Sigma)}[(M_k - \mu)'h\Sigma_k^{-1}(M_k - \mu)] &= (\overline{\mu}_k - \mu)'h\overline{n}_k\overline{\Psi}_k^{-1}(\overline{\mu}_k - \mu) + tr(hE_{q(\Sigma)}[\Sigma_k^{-1}]\Sigma_k^*) \\ E_{q(M,\Sigma)}[(M_k - \overline{\mu}_k)'\overline{h}_k\Sigma_k^{-1}(M_k - \overline{\mu}_k)] &= tr(\overline{h}^{-1}E_{q(\Sigma)}[\Sigma_k^{-1}]\Sigma_k^*) \end{aligned}$$

I now consider the Multivariate Student t distribution. Without restriction, I need to find the expectation

$$E_{q(M,\Sigma,\tau)}[(y_t - M_k)'\tau_{tk}\Sigma_k^{-1}(y_t - M_k)] = E_{q(\Sigma)}\left\{E_{q(M|\Sigma)}[(y_t - M_k)'E_{q(\tau)}[\tau_{tk}]\Sigma_k^{-1}(y_t - M_k)]\right\}$$

Since $M_k \sim q^*(M_k \mid \Sigma_k) = N(\overline{\mu}_k, \overline{h}_k^{-1}\Sigma_k)$, I re-write the term inside the bracket as

$$\begin{aligned} &E_{q(M|\Sigma)}[(y_t - \overline{\mu}_k)'E_{q(\tau)}[\tau_{tk}]\Sigma_k^{-1}(y_t - \overline{\mu}_k)] + 2E_{q(M|\Sigma)}[(y_t - \overline{\mu}_k)'E_{q(\tau)}[\tau_{tk}]\Sigma_k^{-1}(\overline{\mu}_k - M_k)] \\ &+ E_{q(M|\Sigma)}[(\overline{\mu}_k - M_k)'E_{q(\tau)}[\tau_{tk}]\Sigma_k^{-1}(\overline{\mu}_k - M_k)] \end{aligned}$$

I use a similar justification as above to have the last term is

$$\begin{aligned} E_{q(M|\Sigma)}[(\overline{\mu}_k - M_k)'E_{q(\tau)}[\tau_{tk}]\Sigma_k^{-1}(\overline{\mu}_k - M_k)] &= tr(E_{q(M|\Sigma)}[E_{q(\tau)}[\tau_{tk}]\Sigma_k^{-1}(M_k - \overline{\mu}_k)'(M_k - \overline{\mu}_k)]) \\ &= NE_{q(\tau)}[\tau_{tk}]\overline{h}_k^{-1} \end{aligned}$$

to conclude that

$$E_{q(M,\Sigma,\tau)}[(y_t - M_k)'\tau_{tk}\Sigma_k^{-1}(y_t - M_k)] = (y_t - \overline{\mu}_k)'E_{q(\tau)}[\tau_{tk}]E_{q(\Sigma)}[\Sigma_k^{-1}](y_t - \overline{\mu}_k) + NE_{q(\tau)}[\tau_{tk}]\overline{h}_k^{-1}$$

Then

$$E_q\left[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right] = (y_t - \overline{\mu}_k)'E_{q(\Sigma)}[\Sigma_k^{-1}](y_t - \overline{\mu}_k) + N\overline{h}_k^{-1}$$

Next, I derive the restricted version for multivariate Student t distribution. Note that $y_t \sim N(M_k, \tau_t^{-1}\Sigma_k)$. Under the same assumption about $\overline{\mu}_k^*$ and Σ_k^*, I have

$$\begin{aligned} E_{q(M,\Sigma,\tau)}[(y_t - M_k)'\tau_{tk}\Sigma_k^{-1}(y_t - M_k)] &= (y_t - \overline{\mu}_k^*)' E_{q(\tau)}[\tau_{tk}] E_{q(\Sigma)}[\Sigma_k^{-1}](y_t - \overline{\mu}_k^*) \\ &\quad + tr(E_{q(\tau)}[\tau_{tk}] E_{q(\Sigma)}[\Sigma_k^{-1}]\Sigma_k^*) \end{aligned}$$

Similarly, I have the restricted expectations under multivariate Student t distribution

$$\begin{aligned} E_{q(M,\Sigma)}\left[(y_t - M_k)'\Sigma_k^{-1}(y_t - M_k)\right] &= (y_t - \overline{\mu}_k^*)' E_{q(\Sigma)}[\Sigma_k^{-1}](y_t - \overline{\mu}_k^*) + tr(E_{q(\Sigma)}[\Sigma_k^{-1}]\Sigma_k^*) \\ E_{q(M,\Sigma)}[(M_k - \mu)' h \Sigma_k^{-1}(M_k - \mu)] &= (\overline{\mu}_k^* - \mu)' h \overline{n}_k \overline{\Psi}_k^{-1}(\overline{\mu}_k^* - \mu) + tr(h E_{q(\Sigma)}[\Sigma_k^{-1}]\Sigma_k^*) \\ E_{q(M,\Sigma)}[(M_k - \overline{\mu}_k)' \overline{h}_k \Sigma_k^{-1}(M_k - \overline{\mu}_k)] &= tr(\overline{h}^{-1} E_{q(\Sigma)}[\Sigma_k^{-1}]\Sigma_k^*) \end{aligned}$$

F Numerical stability

The first formula to keep numerical stability is as

$$\log\left[\sum_{t=1}^{T} \exp(y_t)\right] = y^{max} + \log\left[\sum_{t=1}^{T} \exp(y_t - y^{max})\right]$$

where y^{max} is the maximum element in y.

And I note that the variance-covariance matrix belongs to the class of Hermitian matrix. It is also invertible and its eigen values are all positive. Therefore, I have the property as

$$\log(det(A)) = trace(\log(A))$$

In Matlab, I use *logm* to find the principal logarithm that has all eigenvalues lying between $-\pi$ and π.

G Tables

Table 12: Estimated parameters in Student t distribution with 2 assets

Parameters	Bear state				Bull state			
	$\widehat{M}^1$	$\widehat{\Sigma}^1$		$\widehat{\nu}^1$	$\widehat{M}^2$	$\widehat{\Sigma}^2$		$\widehat{\nu}^2$
True	-0.51	5	3	50	1	1	0	6
	0.49	3	5		1	0	1	
Unrestricted								
MCMC-CC	-0.21	4.86	2.94	28.95	0.97	1.1	-0.1	7.33
	0.88	2.94	4.71		1.08	-0.1	1.09	
MCMC	-0.22	4.86	2.95	23.29	0.97	1.26	-0.1	13.31
	0.88	2.95	4.72		1.08	-0.1	1.24	
MCMC-A	-0.22	4.71	2.86	16.68	0.97	1.18	-0.11	9.78
	0.87	2.86	4.57		1.08	-0.11	1.16	
VI-CC	-0.22	4.80	2.91	26.63	0.97	1.12	-0.1	7.62
	0.88	2.91	4.65		1.08	-0.1	1.12	
VI	-0.21	5.01	3.03	39.97	0.96	1.46	-0.09	39.48
	0.89	3.03	4.87		1.09	-0.09	1.41	
VI-A	-0.22	4.80	2.91	20	0.96	1.35	-0.1	20
	0.87	2.86	4.57		1.08	-0.11	1.16	
Restricted								
MCMC-CC	-0.57	4.84	2.93	27.51	0.97	1.12	-0.1	8.08
	0.54	2.93	4.69		1.08	-0.1	1.11	
MCMC	-0.57	4.75	2.87	17.53	0.97	1.21	-0.1	11.1
	0.53	2.87	4.58		1.08	-0.1	1.2	
MCMC-A	-0.57	4.68	2.83	15.39	0.97	1.15	-0.1	8.96
	0.54	2.83	4.51		1.08	-0.1	1.14	
VI-CC	-0.57	5.06	3.06	72.93	0.98	1.15	-0.1	8.28
	0.54	3.06	4.92		1.08	-0.1	1.14	
VI	-0.57	4.96	3	33.61	0.96	1.48	-0.09	47.33
	0.54	3	4.82		1.08	-0.09	1.43	
VI-A	-0.57	4.68	2.83	15.21	0.97	1.16	-0.1	8.7
	0.54	2.83	4.51		1.08	-0.1	1.14	

Notes: Table 12 contains the posterior mean of parameters and the optimal variational parameters of the multivariate Markov switching model estimated by MCMC algorithm and VI algorithm, respectively. This table includes both results with and without restrictions. The true parameters are also provided.

Table 13: In-sample performance under Student t distribution

Number of assets	2		30	
	Log score	MSE	Log score	MSE
'MCMC-A'	-3.140	0.688	-47.161	65.166
'MCMC-CC'	-3.304	0.375	-47.527	**63.279**
'MCMC'	-3.219	0.466	-47.159	65.169
'VI-A'	**-3.078**	0.444	-46.972	66.371
'VI-CC'	-3.374	**0.342**	-47.323	66.409
'VI'	-3.152	0.344	**-46.950**	66.338

Notes: Table 13 contains the log score and the mean squared errors of the multivariate Markov switching model estimated by MCMC and VI algorithm. All values are in average. The data is from Student t distribution. There are 2 and 30 assets.

Table 14: Out of sample performance with restricted Student t simulation with 2 assets

	MSFE	LPS	Min.Variance	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
'MCMC-A'	12.142	-3.115	4.269	0.361	0.869	2.410	-0.003	-0.003
'MCMC-CC'	12.139	-3.148	**4.160**	0.361	0.869	2.410	-0.003	-0.003
'MCMC'	**12.138**	-3.141	4.235	0.361	0.869	2.410	-0.003	-0.003
'VI-A'	12.142	**-3.114**	4.294	0.361	0.869	2.410	-0.003	-0.003
'VI-CC'	12.140	-3.187	4.329	0.361	0.869	2.410	-0.003	-0.003
'VI'	12.140	-3.159	6.100	0.361	0.869	2.410	-0.003	-0.003

Notes: Table 14 contains the out-of-sample measures of performance of the multivariate Markov switching model estimated by MCMC and VI algorithm. MSFE is the mean squared errors of forecasting errors. LPS means log predictive score. Min.Variance implies the global minimum variance. Sharpe ratio captures the ex-post Sharpe ratio. Mean is the mean of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. SD is the standard deviation of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. Columns Strategy 2 and 3 present the performance fee of switching from these strategies towards "buy-and-hold" strategy. The row Difference shows the gap between the results from VI and MCMC algorithm. Data is consists of 2 assets.

Table 15: Out of sample performance with restricted Student t simulation with 30 assets

	MSFE	LPS	Min.Variance	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
'MCMC-A'	45.212	**-27.326**	0.094	6.129	**1.554**	0.254	-0.015	0
'MCMC-CC'	45.215	-27.493	0.090	5.981	1.543	0.258	-0.015	0
'MCMC'	45.218	-27.380	0.092	6.099	1.552	0.255	-0.015	0
'VI-A'	45.206	-27.329	0.089	6.208	1.551	**0.250**	-0.015	0
'VI-CC'	**45.205**	-27.345	**0.088**	6.204	1.551	**0.250**	-0.015	0
'VI'	49.236	-27.366	0.145	**6.181**	1.553	0.251	-0.015	-0.001

Notes: Table 15 contains the out-of-sample measures of performance of the multivariate Markov switching model estimated by MCMC and VI algorithm. MSFE is the mean squared errors of forecasting errors. LPS means log predictive score. Min.Variance implies the global minimum variance. Sharpe ratio captures the ex-post Sharpe ratio. Mean is the mean of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. SD is the standard deviation of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. Columns Strategy 2 and 3 present the performance fee of switching from these strategies towards "buy-and-hold" strategy. The row Difference shows the gap between the results from VI and MCMC algorithm. Data is consists of 30 assets.

Table 16: Out-of-sample performance for a portfolio with 59 assets

	MSFE	LPS	Min.Var	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
Normal assumption								
MCMC	73.455	-31.120	0.194	0.116	0.052	0.447	0.010	0.001
VI	72.783	-29.432	0.265	**0.215**	**0.095**	0.442	0.001	**0.006**
Student t assumption								
MCMC-CC	**72.659**	-27.664	0.166	0.169	0.070	0.414	**0.007**	0
MCMC	72.667	-27.759	**0.151**	0.170	0.070	**0.410**	**0.007**	0
MCMC-A	72.661	-27.731	0.153	0.172	0.071	0.411	**0.007**	0
VI-CC	72.788	-28.553	0.168	0.137	0.059	0.430	**0.007**	0.001
VI	73.025	-28.153	0.261	0.127	0.055	0.431	0.006	0.001
VI-A	72.986	**-21.289**	0.172	0.131	0.058	0.440	0.006	0.003

Notes: Table 16 stores the measures of out of sample performance for both algorithms to the data of an equally-weighted portfolio with 59 assets. contains the out-of-sample measures of performance of the multivariate Markov switching model estimated by MCMC and VI algorithm. MSFE is the mean squared errors of forecasting errors. LPS means log predictive score. Min.Var implies the global minimum variance. Sharpe ratio captures the ex-post Sharpe ratio. Mean is the mean of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. SD is the standard deviation of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. Columns Strategy 2 and 3 present the performance fee of switching from these strategies towards "buy-and-hold" strategy.

Table 17: Out-of-sample performance for a portfolio with 103 assets

Normal assumption								
	MSFE	LPS	Min.Var	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
MCMC	**116.76**	**-33.392**	0.133	**0.229**	0.097	0.424	**0.005**	0
VI	117.30	-41.671	0.134	0.213	0.085	**0.397**	0.004	0.001
Student t assumption								
MCMC-CC	116.83	-48.798	0.105	0.136	0.078	0.577	0.003	0.002
MCMC	116.83	-48.809	1.797	0.123	0.076	0.615	0.003	0.002
MCMC-A	117.44	-46.031	0.040	0.128	**0.309**	2.418	0.001	**0.004**
VI-CC	116.83	-50.948	0.093	0.128	0.077	0.607	0.003	0.002
VI-A	117.33	-38.905	**0.010**	0.023	0.036	1.582	0.001	**0.004**
VI	117.19	-37.269	0.098	0.061	0.077	1.250	0.004	0.002

Notes: Table 17 stores the measures of out of sample performance for both algorithms to the data of an equally-weighted portfolio with 103 assets. All values are in average. MSFE is the mean squared errors of forecasting errors. Variance includes the minimum variance. Sharpe ratio captures the maximum Sharpe ratio. Columns Strategy 2 and 3 present the performance fee of switching from these strategies towards “buy-and-hold” strategy. The row Difference shows the gap between the results from VI and MCMC algorithm.

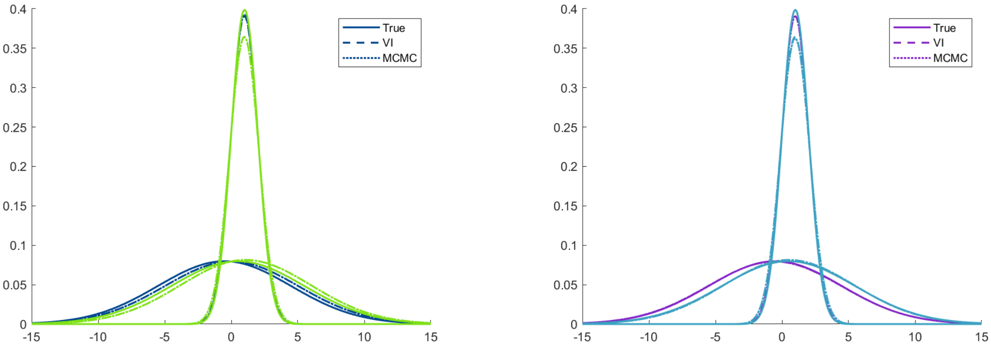
Table 18: Out-of-sample performance for a portfolio with 40 industry portfolio

Normal								
	MSFE	LPS	Min.Var	Sharpe ratio	Mean	SD	Strategy 2	Strategy 3
MCMC	2616.0	**-59.815**	29.591	0.156	0.799	5.135	-0.007	0.001
VI	**2582.1**	-79.352	33.389	**0.213**	**0.955**	4.483	-0.005	0.001
Student t								
MCMC-CC	2622.9	-67.533	26.779	0.176	0.715	4.053	-0.007	**0.002**
MCMC	2613.2	-66.967	23.107	0.170	0.686	4.028	-0.007	0.001
MCMC-A	2615.0	-67.000	24.425	0.164	0.663	4.030	-0.007	0.001
VI-CC	2607.4	-77.116	**14.720**	0.171	0.646	**3.774**	0.001	-0.002
VI	2603.2	-76.971	36.506	0.192	0.745	3.875	0.001	-0.002
VI-A	2604.9	-64.902	24.430	0.171	0.673	3.928	**0.002**	-0.004

Notes: Table 18 stores the measures of out of sample performance for both algorithms to the data of an equally-weighted portfolio with 103 assets. MSFE is the mean squared errors of forecasting errors. LPS means log predictive score. Min.Variance implies the global minimum variance. Sharpe ratio captures the ex-post Sharpe ratio. Mean is the mean of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. SD is the standard deviation of returns yielded from using the optimal weight of the ex-ante Sharpe ratio for allocating the portfolio. Columns Strategy 2 and 3 present the performance fee of switching from these strategies towards “buy-and-hold” strategy.

H Figures

Figure 6: Estimated density of the latent states with and without restrictions for 2 assets

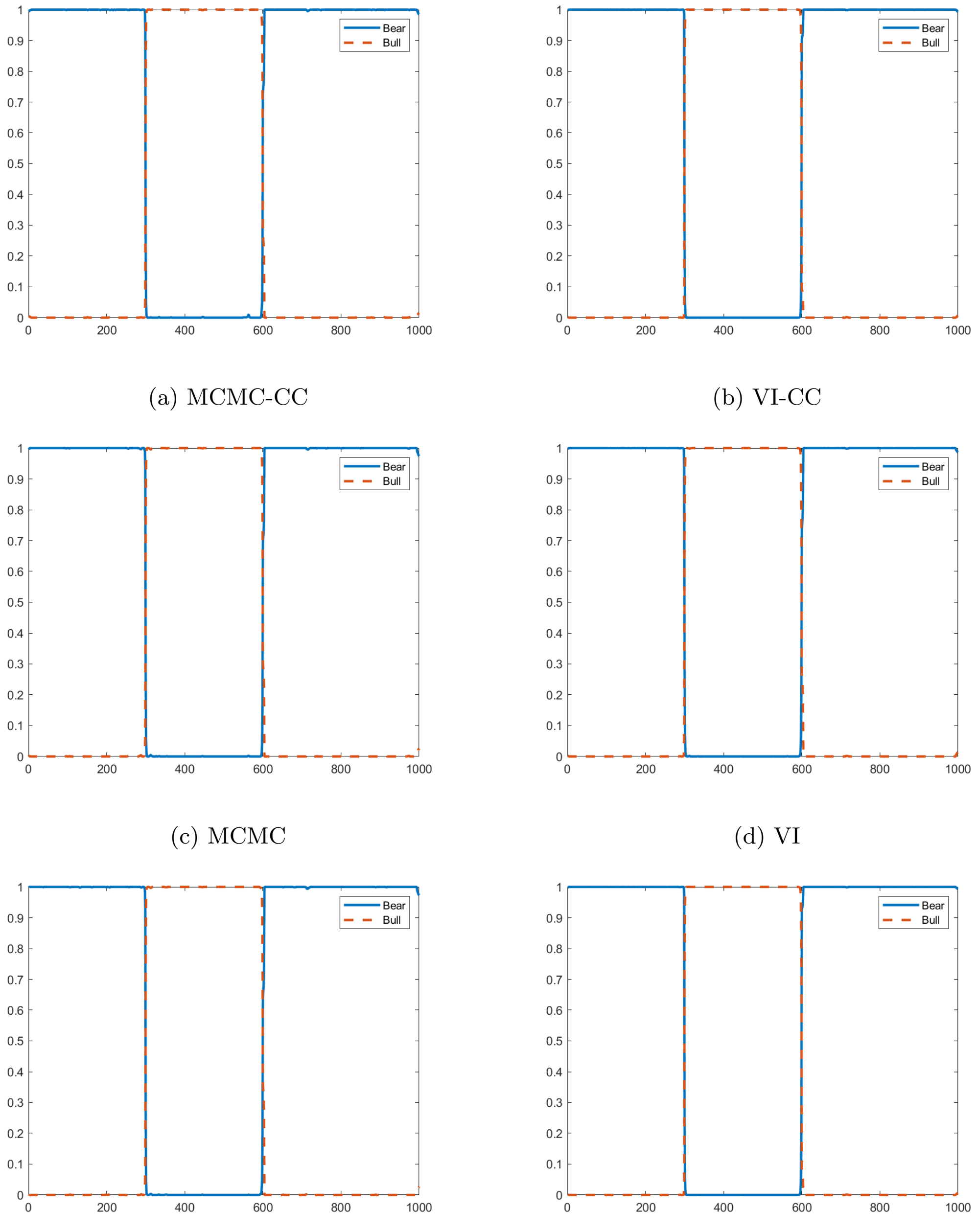


(a) Without restriction

(b) With restriction

Notes: Figure 6 shows densities of two states. The left graph is the result of the unrestricted MCMC and VI algorithm. The right graph is the result of the restricted MCMC and VI algorithm. Both data are normally distributed.

Figure 7: Estimated probabilities of the latent states with restrictions for 2 assets under Student t



(a) MCMC-CC

(b) VI-CC

(c) MCMC

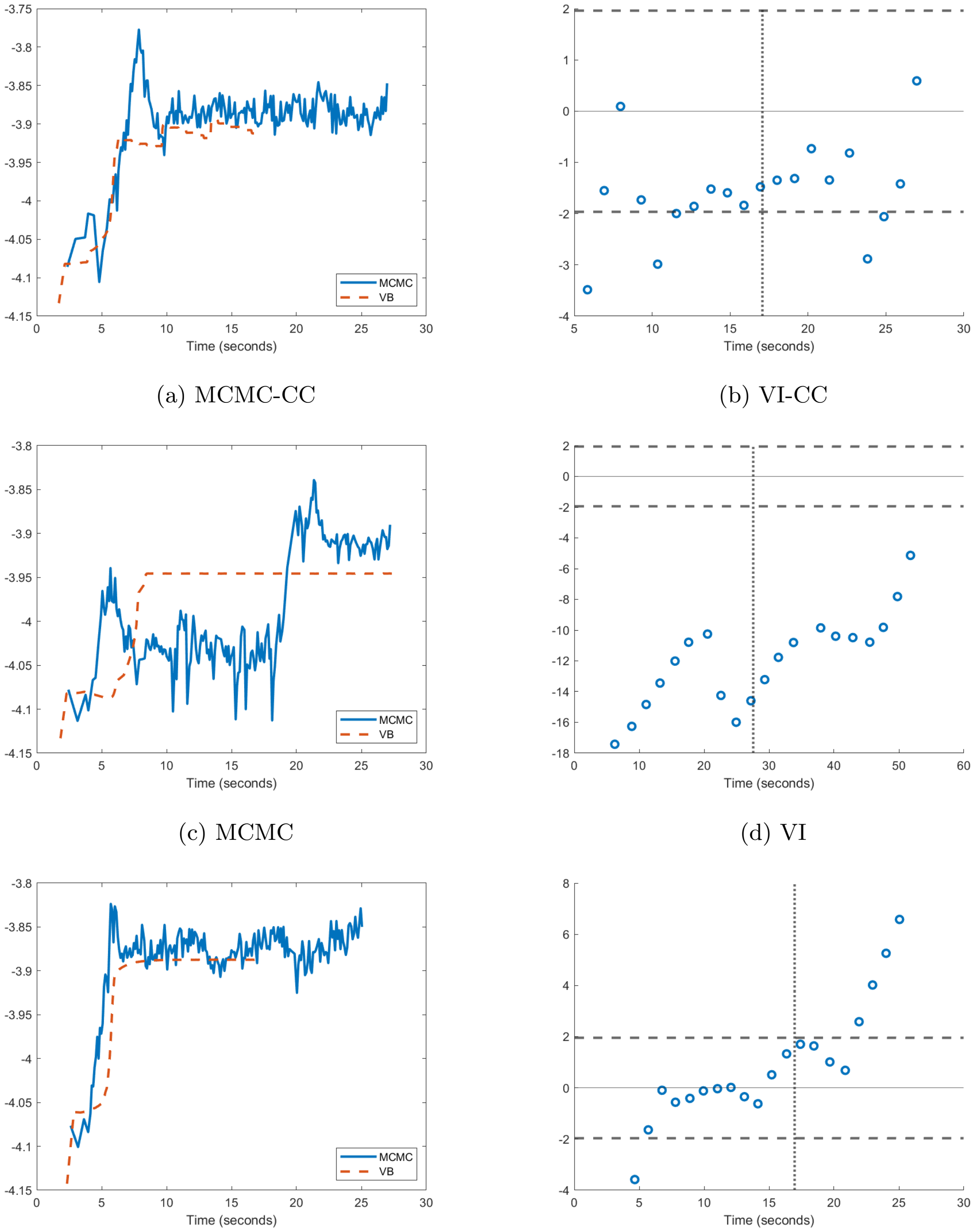
(d) VI

(e) MCMC-A

(f) VI-A

Notes: Figure 7 contains conditional posterior probabilities of states against and the optimal variational probabilities from the MCMC and VI algorithm. The first row results from left to right are MCMC-CC, MCMC and MCMC-A, respectively. The second row results from left to right are VI-CC. VI, and VI-A, respectively.

Figure 8: Time trace plot of Loglikelihood and Geweke's plot with 2 assets



(a) MCMC-CC

(b) VI-CC

(c) MCMC

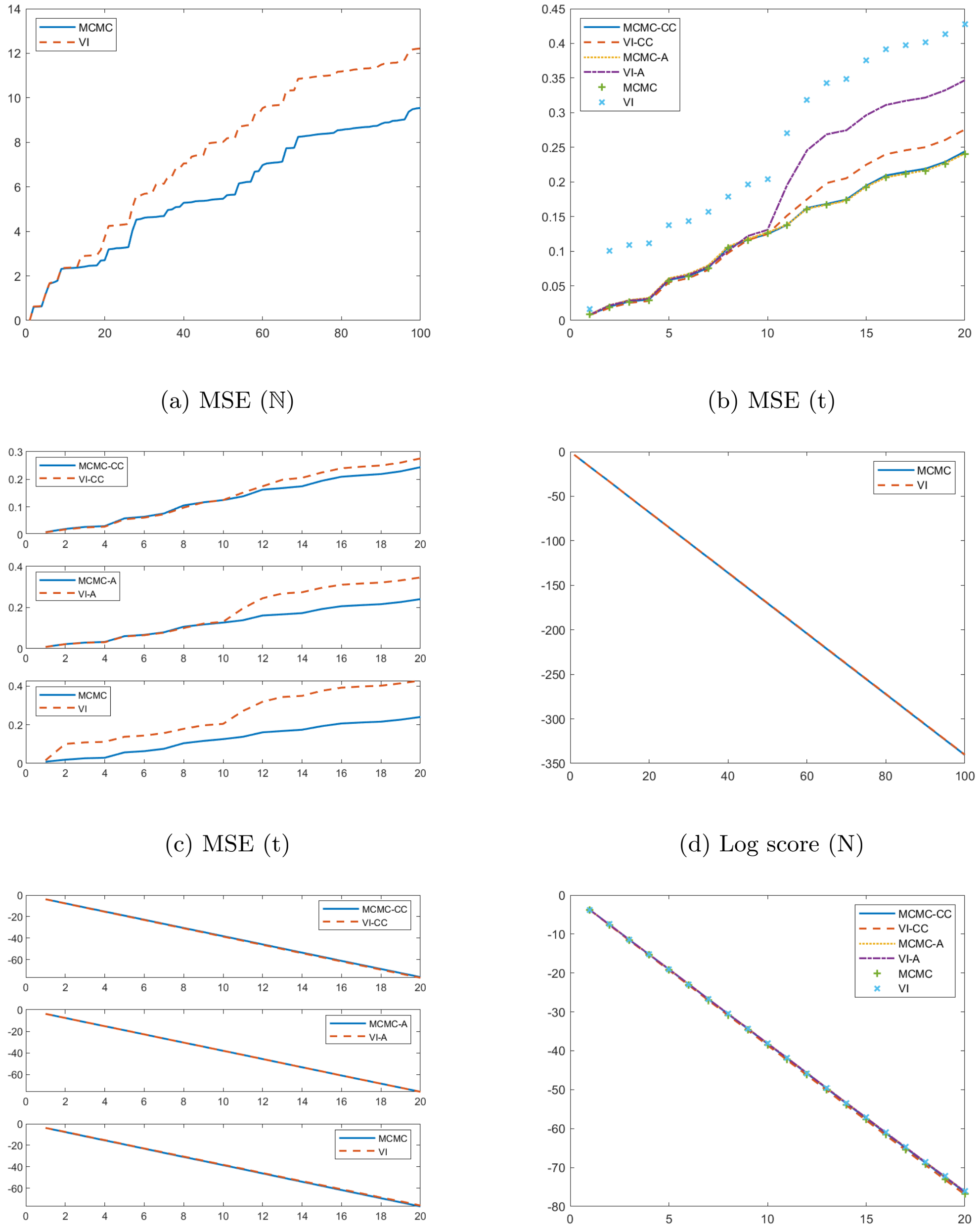
(d) VI

(e) MCMC-A

(f) VI-A

Notes: Figure 8 presents the time trace plots of the average log likelihood and Geweke (1992)'s plots. The first row is the time trace plot: Left: MCMC-CC vs VI-CC; Middle: MCMC vs VI; Right: MCMC-A vs VI-A. The second row is the corresponding Geweke (1992)'s plots. The vertical dotted line indicates the time that the VI converges. Data is from a Student t distribution with 2 assets.

Figure 9: The in-sample performance for 2 assets



(a) MSE (ℕ)

(b) MSE (t)

(c) MSE (t)

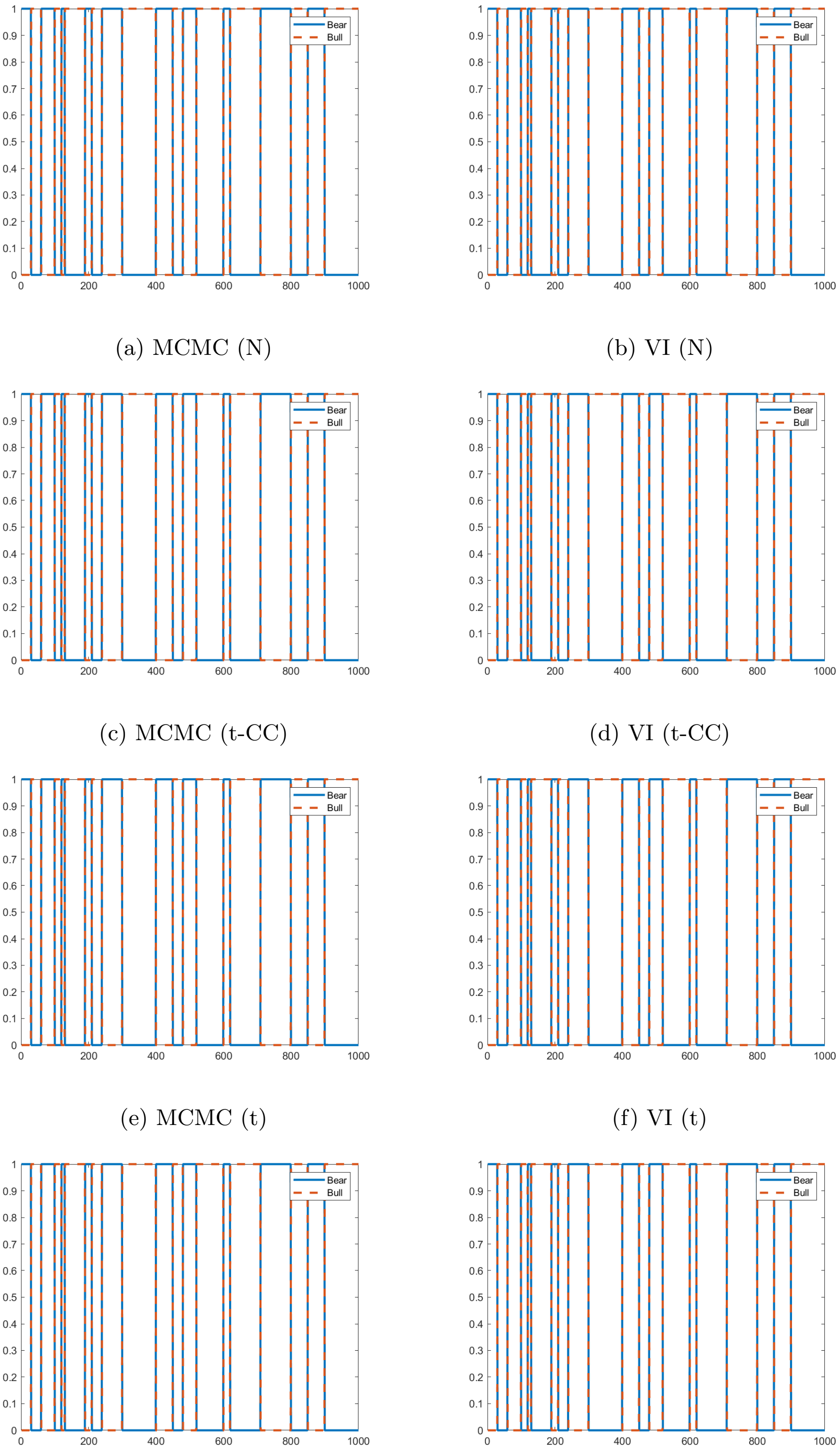
(d) Log score (N)

(e) Log score (N)

(f) Log score (N)

Notes: Figure 9 presents the in-sample comparison between two estimation approaches for 2 assets. The first row is the mean squared of errors. The second row is the log score. The first column shows the results of the Normal data. The next two columns show the results of the Student t data.

Figure 10: Estimated probabilities of the latent states for 30 assets



(a) MCMC (N)

(b) VI (N)

(c) MCMC (t-CC)

(d) VI (t-CC)

(e) MCMC (t)

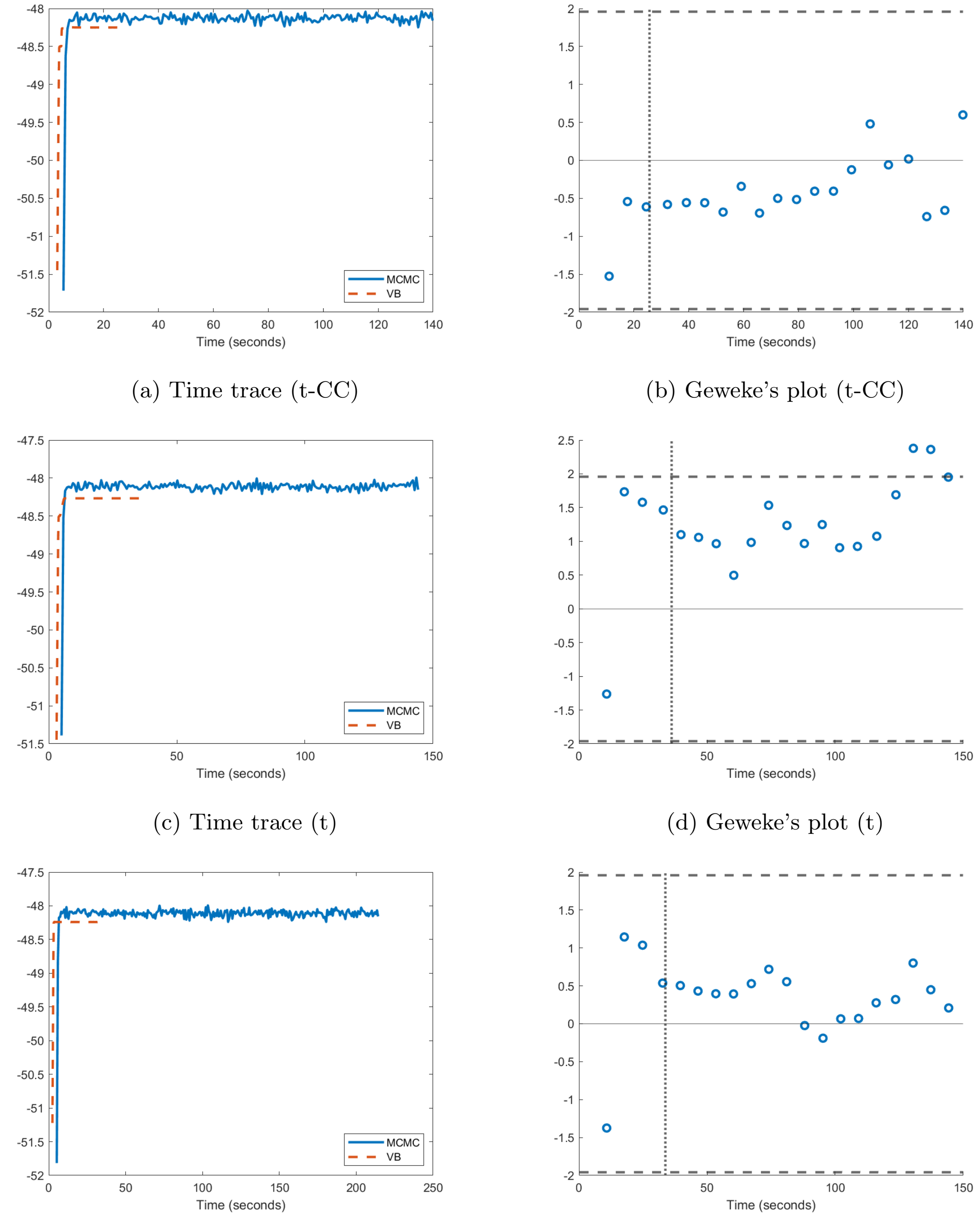
(f) VI (t)

(g) MCMC (t-A)

(h) VI (t-A)

Notes: Figure 10 contains the conditional posterior probabilities of states and the optimal variational probabilities from the MCMC and VI methods. The first column from the left to right are results of Normal distribution. The next columns are results of the Student t distribution where from right to left, each column has results from MCMC-CC & VI-CC, MCMC & VI, and MCMC-A & VI-A, respectively.

Figure 11: Time trace plot of Log likelihood and Geweke's plot with 30 assets



(a) Time trace (t-CC)

(b) Geweke's plot (t-CC)

(c) Time trace (t)

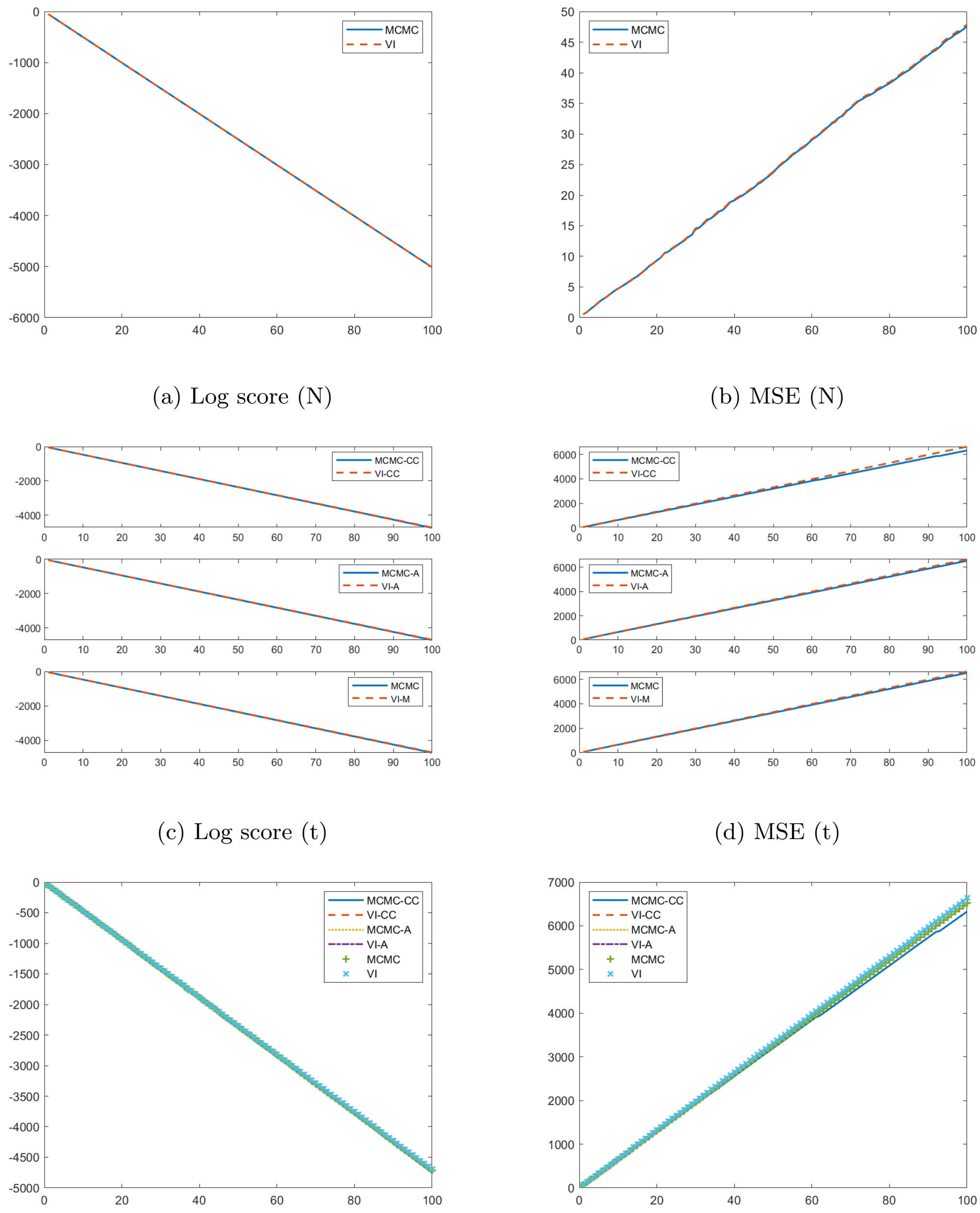
(d) Geweke's plot (t)

(e) Time trace (t-A)

(f) Geweke's plot (t-A)

Notes: Figure 11 presents the time trace plot of the average log likelihood and Geweke (1992)'s plots. The first row is the time trace plot: Left: MCMC-CC vs VI-CC; Middle: MCMC vs VI; Right: MCMC-A vs VI-A. The second row is the corresponding Geweke (1992)'s plots. The vertical dotted line indicates the time that the VI converges. Data is from a Student t distribution with 30 assets.

Figure 12: The in-sample comparison for 30 assets



(a) Log score (N)

(b) MSE (N)

(c) Log score (t)

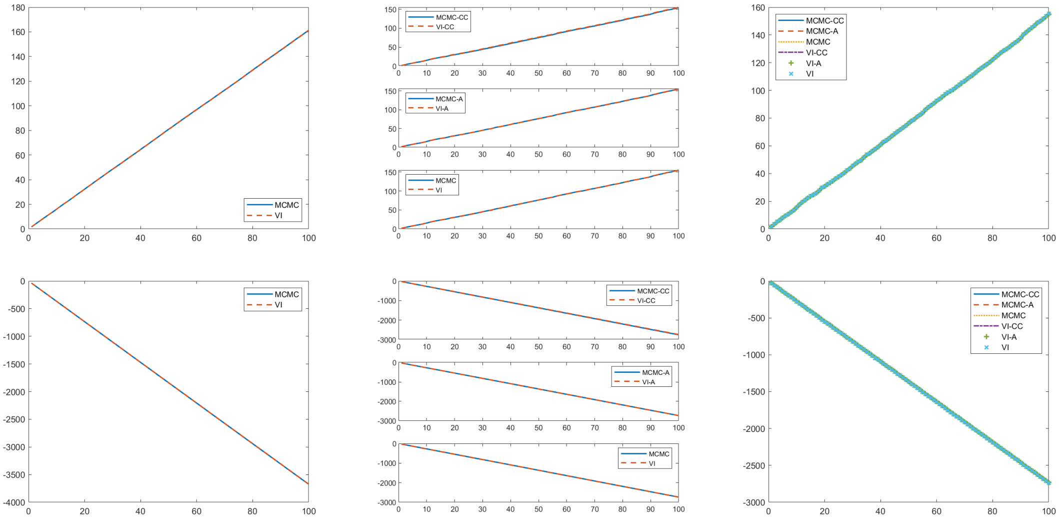
(d) MSE (t)

(e) Log score (t)

(f) MSE (t)

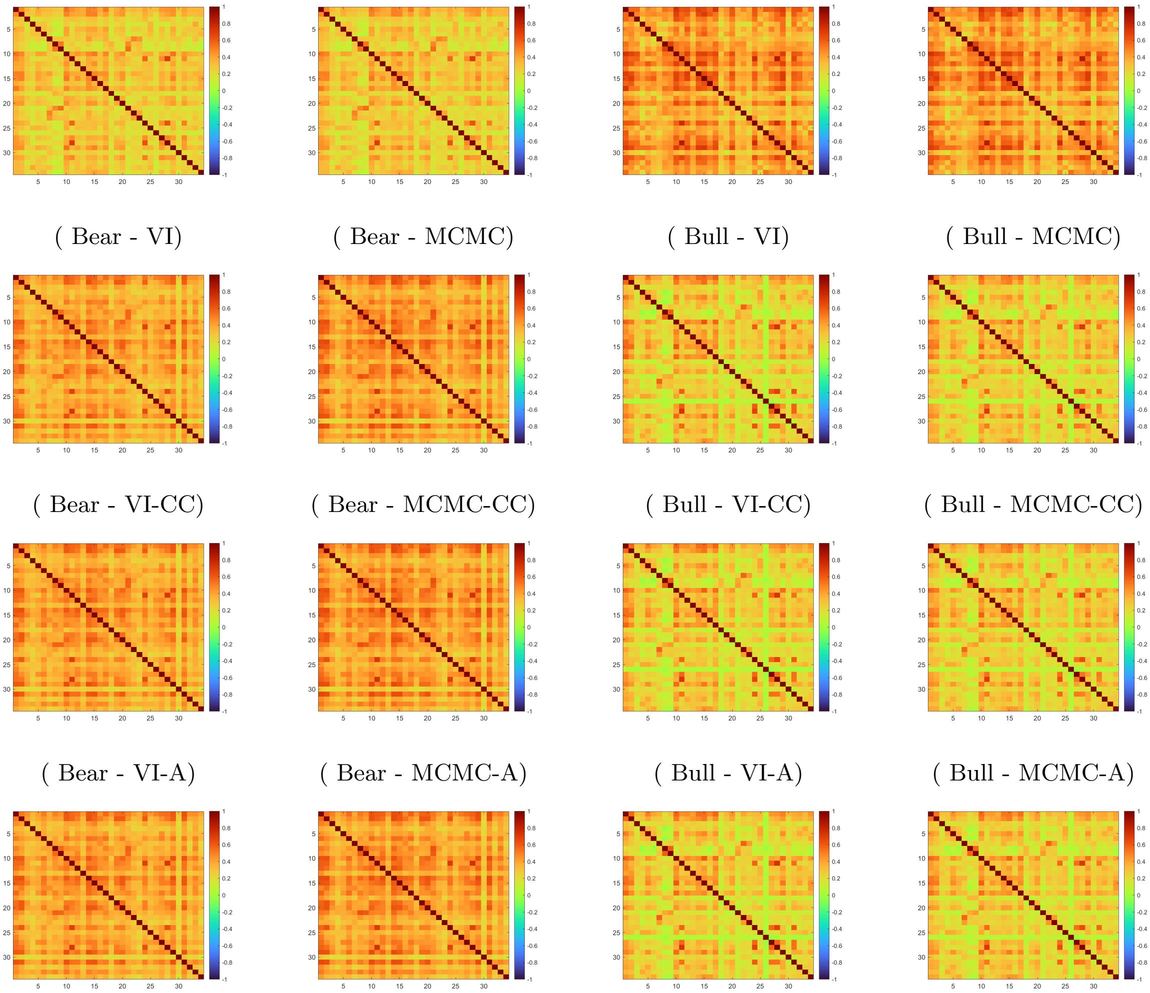
Notes: Figure 12 presents the in-sample comparison between two estimation approaches for 30 assets. The first row is the mean squared of errors. The second row is the log score. The first column shows the results of a Normal data. The next two columns show the results of the Student t data.

Figure 13: The out-of-sample performance for 30 assets



Notes: Figure 13 presents the out-of-sample comparison between two inference approaches. The first row is the cumulative one-period-ahead predictive return that is calculated from the optimal ex-ante Sharpe ratio. The second row is the cumulative predictive loglikehood. The first column is the result of the Normal data. The next two columns are the results of the Student t distribution.

Figure 14: Estimated correlation matrix of bear and bull state of EW34



(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

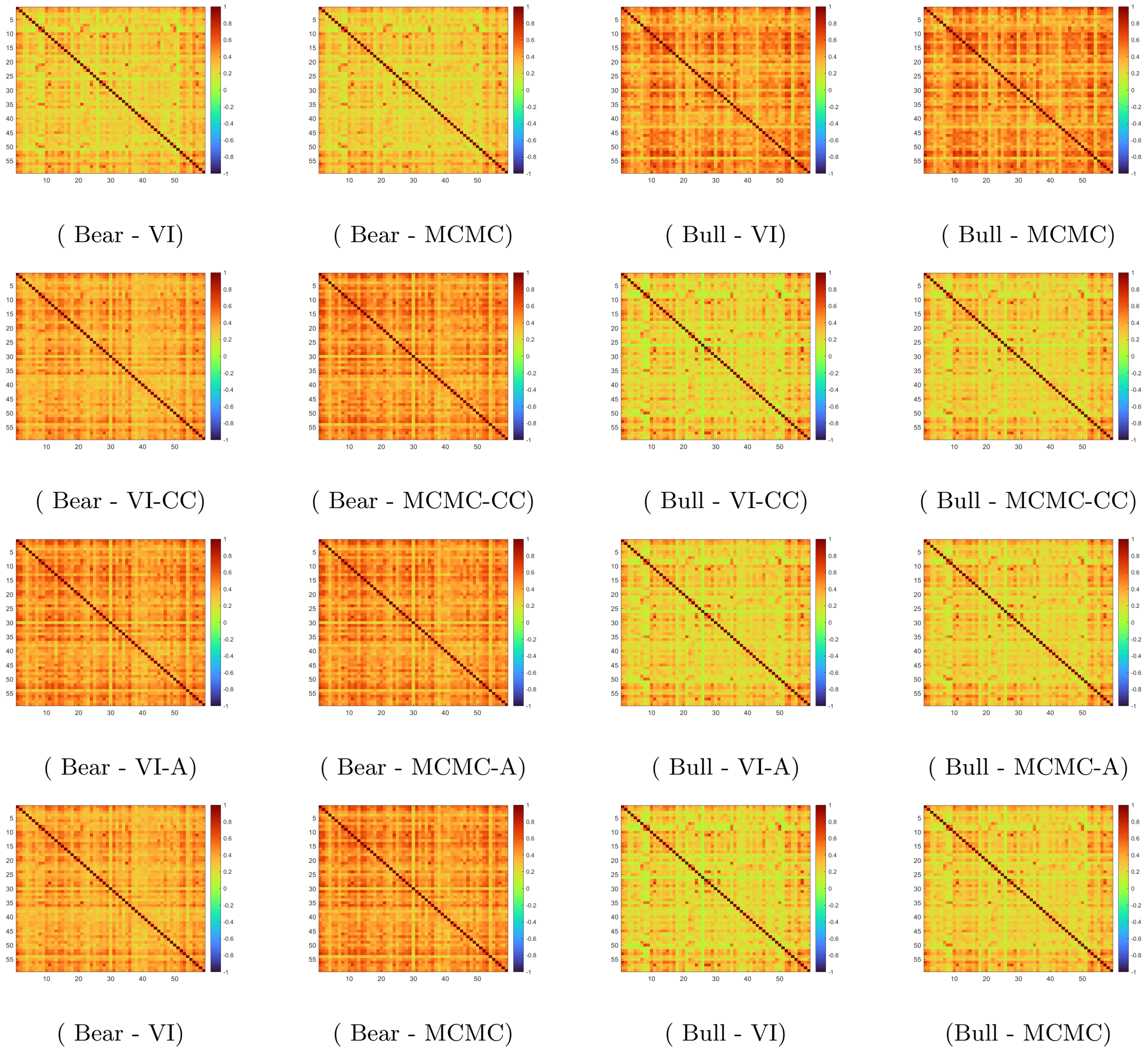
(Bear - VI-CC) (Bear - MCMC-CC) (Bull - VI-CC) (Bull - MCMC-CC)

(Bear - VI-A) (Bear - MCMC-A) (Bull - VI-A) (Bull - MCMC-A)

(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

Notes: Figure 14 consists of estimated correlation matrices from both VI and MCMC algorithms. Results are from a data set, EW34. The first row is the result of the Normal assumption. The next three rows are results of the Student t assumption.

Figure 15: Estimated correlation matrix of bear and bull state of EW59



(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

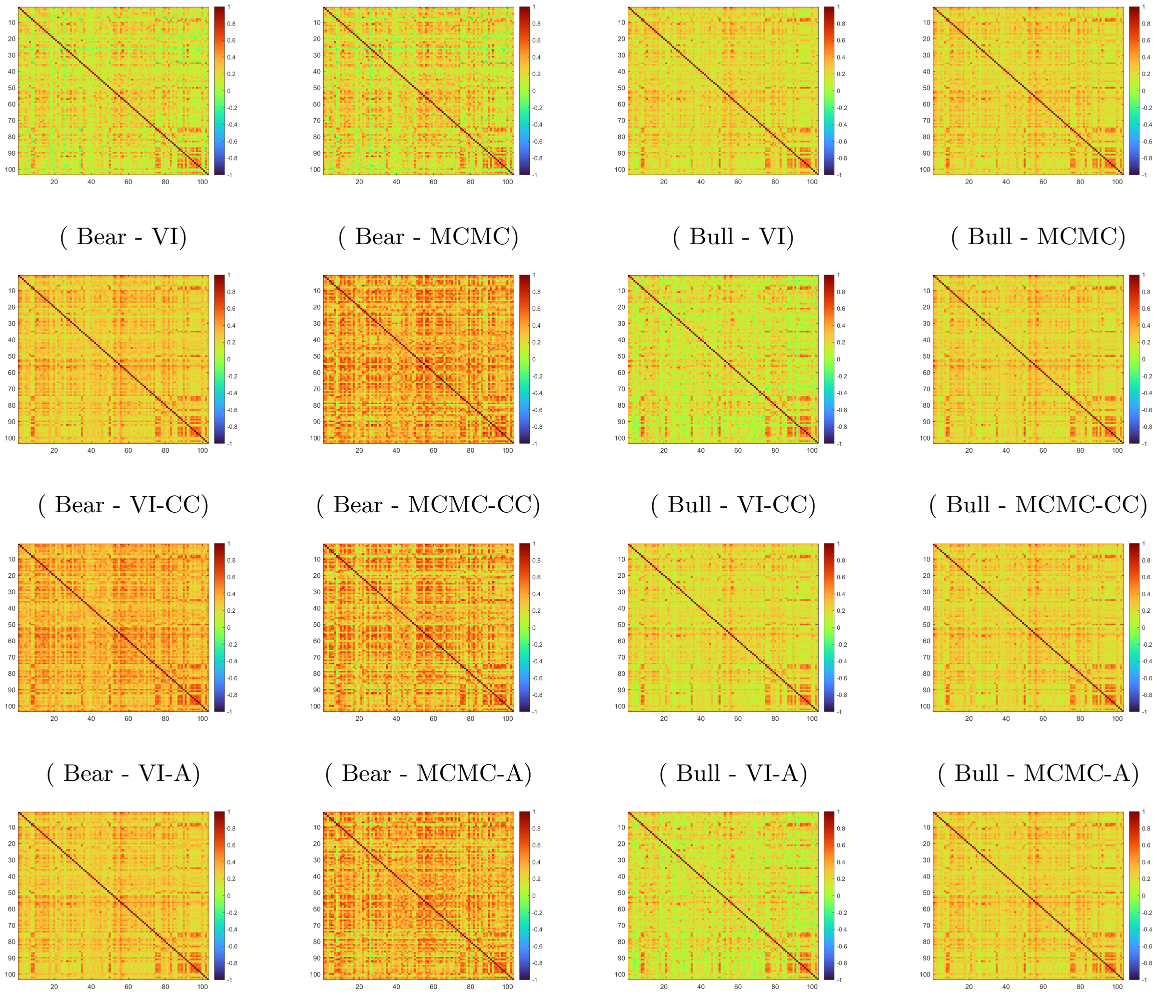
(Bear - VI-CC) (Bear - MCMC-CC) (Bull - VI-CC) (Bull - MCMC-CC)

(Bear - VI-A) (Bear - MCMC-A) (Bull - VI-A) (Bull - MCMC-A)

(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

Notes: Figure 15 consists of estimated correlation matrices from both VI and MCMC algorithms. Results are from a data set, EW59. The first row is the result of the Normal assumption. The next three rows are results of the Student t assumption.

Figure 16: Estimated correlation matrix of bear and bull state of EW103



(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

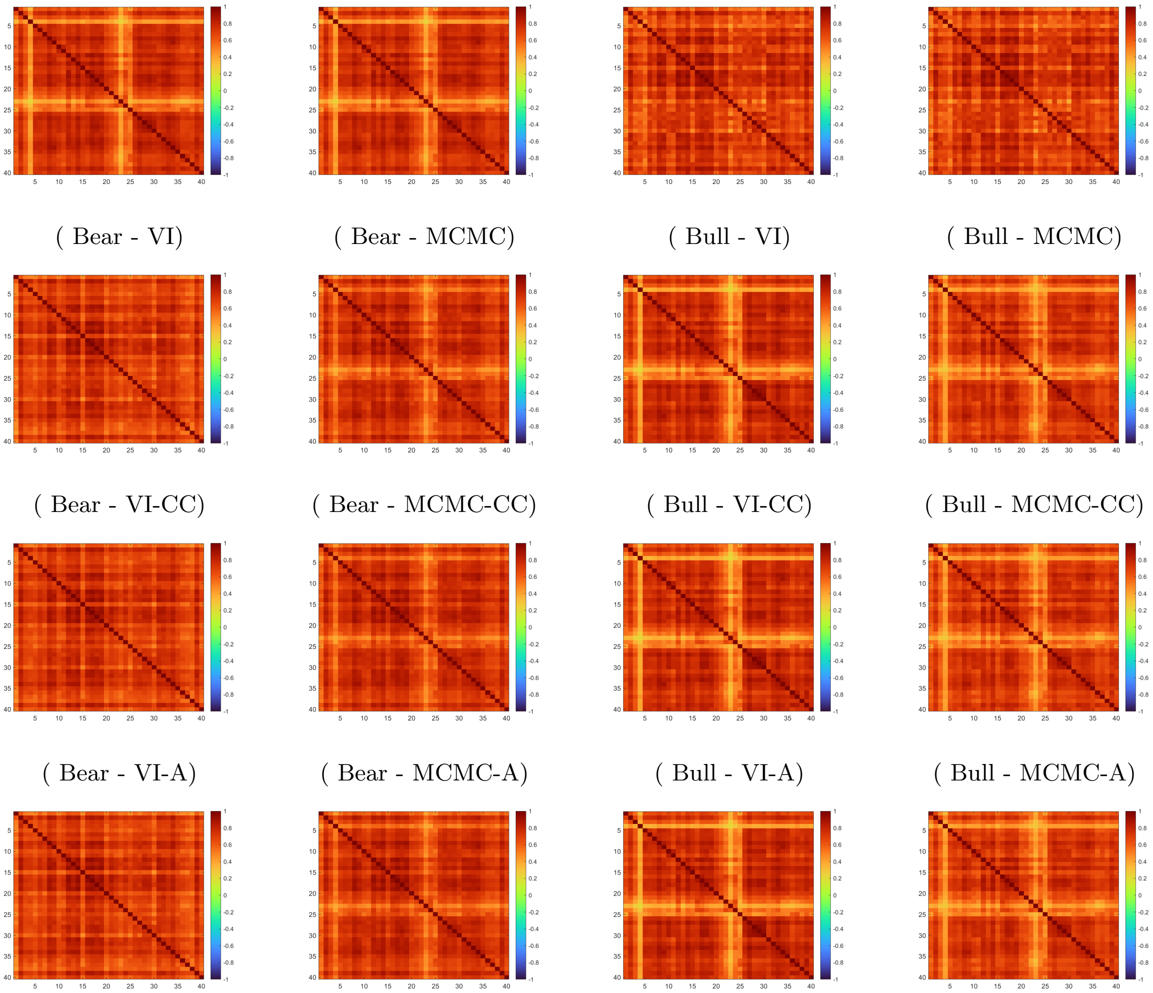
(Bear - VI-CC) (Bear - MCMC-CC) (Bull - VI-CC) (Bull - MCMC-CC)

(Bear - VI-A) (Bear - MCMC-A) (Bull - VI-A) (Bull - MCMC-A)

(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

Notes: Figure 16 consists of estimated correlation matrices from both VI and MCMC algorithms. Results are from a data set, EW103. The first row is the result of the Normal assumption. The next three rows are results of the Student t assumption.

Figure 17: Estimated correlation matrix of bear and bull state of IP40



(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

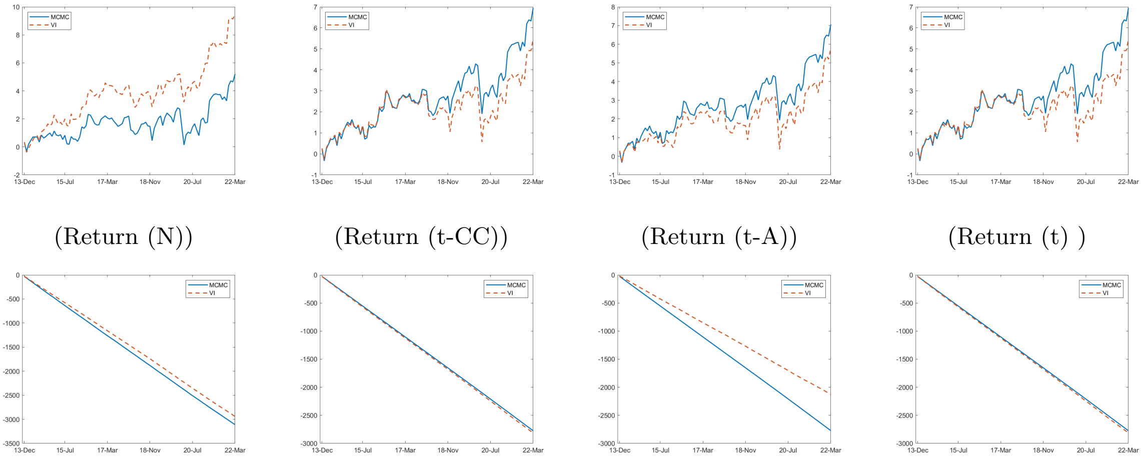
(Bear - VI-CC) (Bear - MCMC-CC) (Bull - VI-CC) (Bull - MCMC-CC)

(Bear - VI-A) (Bear - MCMC-A) (Bull - VI-A) (Bull - MCMC-A)

(Bear - VI) (Bear - MCMC) (Bull - VI) (Bull - MCMC)

Notes: Figure 17 consists of estimated correlation matrices from both VI and MCMC algorithms. Results are from a data set, IP40. The first row is the result of the Normal assumption. The next three rows are results of the Student t assumption.

Figure 18: Cumulative predictive returns and loglikelihood for EW59

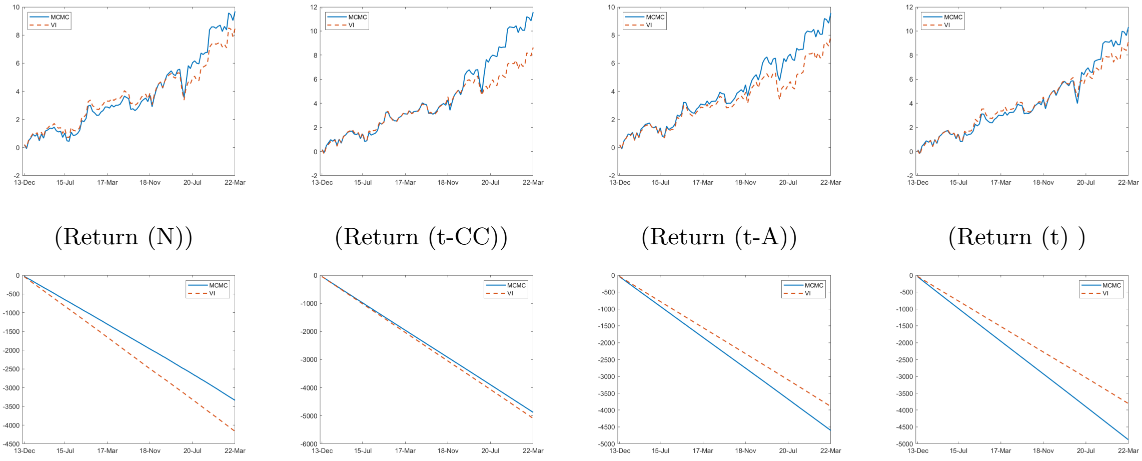


(Return (N)) (Return (t-CC)) (Return (t-A)) (Return (t))

(LLH (N)) (LLH (t-CC)) (LLH (t-A)) (LLH (t))

Notes: Figure 18 consists of out-of-sample results from both VI and MCMC algorithms for the portfolio of 59 equally-weighted assets. The first column on the left is the result of the Normal assumption. The next columns are the result of the Student t assumption. The upper row contains the cumulative predictive returns that are from the optimal weight calculated by the ex-ante Sharpe ratio. The second row includes the cumulative predictive loglikehood.

Figure 19: Cumulative predictive returns and loglikelihood for EW103

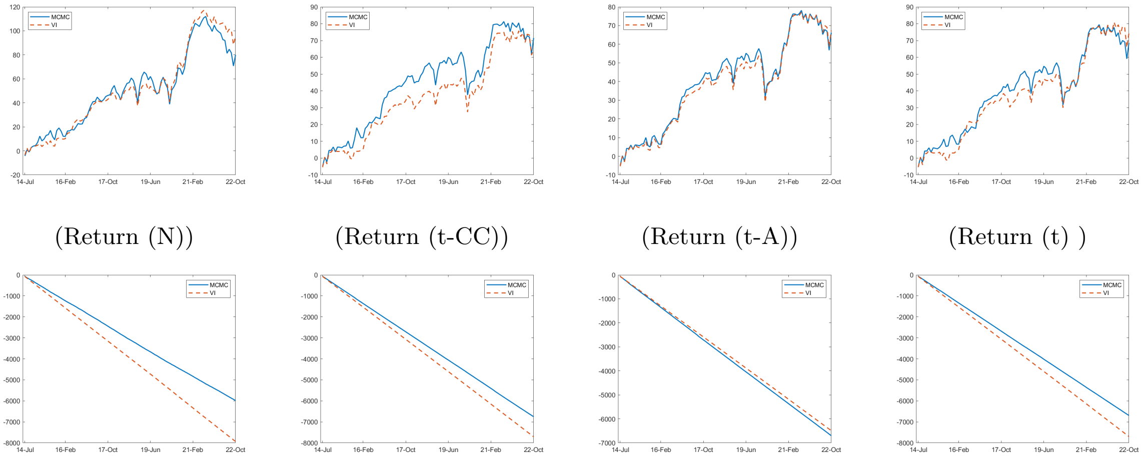


(Return (N)) (Return (t-CC)) (Return (t-A)) (Return (t))

(LLH (N)) (LLH (t-CC)) (LLH (t-A)) (LLH (t))

Notes: Figure 19 consists of out-of-sample results from both VI and MCMC algorithms for the portfolio of 103 equally-weighted assets. The first column on the left is the result of the Normal assumption. The next columns are the result of the Student t assumption. The upper row contains the cumulative predictive returns that are from the optimal weight calculated by the ex-ante Sharpe ratio. The second row includes the cumulative predictive loglikehood.

Figure 20: Cumulative predictive returns and loglikelihood for IP40



(Return (N))

(Return (t-CC))

(Return (t-A))

(Return (t))

(LLH)

(LLH (t-CC))

(LLH (t-A))

(LLH (t))

Notes: Figure 20 consists of out-of-sample results from both VI and MCMC algorithms for the portfolio of 40 equally-weighted industry portfolio. The first column on the left is the result of the Normal assumption. The next columns are the result of the Student t assumption. The upper row contains the cumulative predictive returns that are from the optimal weight calculated by the ex-ante Sharpe ratio. The second row includes the cumulative predictive loglikehood.