A Truncated Mixture Transition Model for Interval-valued Time Series

Yun Luo

Department of Economics, Finance, and Real Estate, Monmouth University

Gloria González-Rivera

Department of Economics, University of California, Riverside

Address correspondence to Gloria González-Rivera, Department of Economics, University of California, Riverside, CA 92521-0427. Telephone: (951) 827-1590. Email: gloria.gonzalez@ucr.edu.

Abstract

We propose a model for interval-valued time series that specifies the conditional joint distribution of the upper and lower bounds as a mixture of truncated bivariate normal distributions. It preserves the interval natural order and provides great flexibility on capturing potential conditional heteroscedasticity and non-Gaussian features. The standard EM algorithm applied to truncated mixtures does not provide a closed-form solution in the M step. A new EM algorithm solves this problem. The model applied to the interval-valued IBM daily stock returns exhibits superior performance over competing models in-sample and out-of-sample evaluation. A trading strategy showcases the usefulness of our approach.

Key Words: interval-valued data, mixture transition model, EM algorithm, truncated normal distribution.

JEL Classification: C01, C32, C34.

1 Introduction

Interval data refers to data sets where the observation is an interval in contrast to a single point. Intervals arise in a variety of situations. There are instances when the data is directly collected in interval format. A standard example is survey design that avoids asking participants about private or sensitive information, e.g. income, and the answer is provided in interval format, e.g. [$50K, $100K]. In these cases, interval data is the only data format available to the researchers. In other instances, intervals arise as a result of aggregating data. The data may be collected at the individual level, e.g., gas prices in a gas station, but the research question deals with a larger unit, e.g., gas prices at the county level. Rather than providing an average of gas station prices, aggregating the data in interval format for each county is more informative because it preserves the internal price variation of each county. Aggregating the data into intervals may also provide information on volatility, which is particularly useful in financial markets e.g. daily max/min price interval provides information on both the price level and the daily price volatility. Finally, intervals can also arise because there is uncertainty on the measurement of the variable of interest. Regardless of the data generation mechanism of intervals, we define an interval-valued time series (ITS) as a collection of interval data observed over time as opposed to the classical point-valued time series (PTS) where the observations are scalars ordered over time.

The defining feature of an interval is the order of its bounds, i.e., the upper bound cannot be smaller than the lower bound. A formal modeling of ITS with the bound restriction was introduced by González-Rivera and Lin (2013), who propose a constrained regression model (GL) that preserves the natural order of the interval. They assume that the bivariate errors of the system of bounds follow a bivariate truncated normal distribution, where the truncation encloses the constraint that the upper bound is not smaller than the lower bound. However, this distributional assumption is restrictive as the consistency of the estimators heavily depends on it.

Conditional heteroscedasticity and non-Gaussian behavior such as flat stretches, bursts, outliers, and change points (see Le *et al.*, 1996; Wong and Li, 2000) are also important features

that, to the best of our knowledge, have not been explicitly modeled in ITS. [1] These features open the field for models capable of generating more flexible predictive densities. Non-Gaussian features have been extensively considered in the PTS literature. Particularly, Le *et al.* (1996) propose a Mixture Transition Distribution (MTD) model for univariate PTS that accounts for non-Gaussian features. Their idea is to specify the conditional distribution of the variable of interest as a mixture distribution. The fact that MTD is able to handle conditional heteroscedasticity is noted and discussed by Berchtold and Raftery (2002). MTD is further generalized by Wong and Li (2000) under the name of Mixture Autoregressive (MAR), and by Hassan and Lii (2006), who extend MTD to marked point processes under a bivariate setting.

In this paper, we propose a model for ITS in the spirit of the MTD model and its extensions. We specify the joint conditional distribution of the upper bound (x_t) and lower bound (y_t) as a mixture of truncated bivariate normal distributions, where for each component the bivariate normal distribution is truncated at $x_t \geq y_t$. For each component, the pseudo location of the truncated bivariate normal distribution is a linear function of the information set. [2] This model provides several advantages. First, it preserves the natural order of ITS, that is, the upper bound is not smaller than the lower bound for all the observations in the ITS. Second, the model captures conditional heteroskedasticity as the covariance matrix of the process becomes time-varying due to the dynamic truncation and the mixture framework. Third, the mixture distribution provides great flexibility to approximate the underlying true conditional bivariate distribution of the lower/upper bounds, and hence improving the quality of density forecast.[3]

For mixture models, the maximum likelihood estimator (MLE) does not have a closed-form solution because of the complexity of the likelihood function. The standard approach to find the MLE is to implement the EM algorithm due to its simplicity and monotonicity in the likelihood (Dempster *et al.*, 1977). The EM algorithm is based on the idea of data augmentation.[4] Specifically, it finds the MLE that maximizes the target likelihood function by maximizing a pseudo complete likelihood function derived from data augmentation. By construction, the pseudo complete likelihood function is easier to maximize (usually it has

closed form solutions) than the target likelihood function. However, when the components of the mixture are subject to truncation, the data augmentation techniques in the standard EM algorithm to estimate mixture models (see Hamilton, 1990; Le *et al.*, 1996; Hassan and Lii, 2006) do not provide closed-form solutions when maximizing the pseudo complete likelihood function in the M step. To overcome this problem, we propose a new EM algorithm that considers two data augmentation processes. [5] A first augmentation brings latent variables that will suggest from which component of the mixture the observation will truly come and, conditional on this step, a second augmentation provides additional latent variables that will suggest whether the observation to be generated is invalid and then falls into the truncated region.

Monte Carlo simulations indicate that the new EM algorithm performs well in finite samples. Even with a small sample size ($T = 200$), the parameter estimates are precise. As expected, the standard errors of the parameter estimates decrease when the sample size increases. The standard errors also differ across components of the mixture. Standard errors in components with large weights tend to be smaller than those in components with smaller weights. This is also expected because there is less information available for components with smaller weights. Hence, a larger sample size would be desirable to estimate precisely the parameters of those components. The Monte Carlo simulations also provide some evidence on the asymptotic normality of the MLE. For a restricted version of the TMT model, simulation results show that the density of the ML estimator departs from normality in small samples, but as the sample size increases, we observe a gradual approach towards normality.

We apply the model to the ITS of the IBM daily stock returns. In sample, the model fits well the data. We identify four components with the first two explaining 75% of the dynamics of the series and capturing periods of low volatility. It is the fourth component, which has the smaller weight, the one to capture high volatility periods. Out-of-sample, the proposed mixture model outperforms the best competing model (VAR(7)-DCC-t density) on approximating the underlying conditional distribution. To demonstrate the usefulness of the model, we apply a trading strategy developed by González-Rivera et al. (2020) that exploits the probability distribution of high/low return forecasts. Based on the density forecasts of

the proposed mixture model, the trading strategy delivers higher profits or smaller losses than those based on the density forecasts of the best competing model.

As an alternative procedure to our modeling strategy to avoid the truncation of the density, we could model the upper bound (x_t) and the log of the range of the interval ($\log(x_t - y_t)$), by employing a bivariate Gaussian mixture distribution (e.g., the MAR model by Wong and Li (2000)). However, this procedure creates new econometric problems, when we need to recover the range of the interval. The predicted value of range will require some bias corrections that depend on the assumed range distribution. Furthermore, it is not trivial to obtain the joint distribution of the upper and lower bounds and to build the prediction regions from the joint distribution of the upper bound and log of the range. See González-Rivera et al. (2020) for a bootstrap approach that accomplishes such transformation.

The organization of the paper is as follows. In Section 2, we introduce the truncated mixture transition model and discuss some properties. In Section 3, we estimate the model by MLE and provide a new EM algorithm. In Section 4, we perform Monte Carlo simulations, and in Section 5, we apply our model to ITS of the IBM daily stock returns and implement a trading strategy to show the usefulness of our proposed model. We conclude in Section 6. In Appendix, we provide the technical details of the EM algorithm, a discussion of the stationary condition, and the consistency of the ML estimator.

2 The Truncated Mixture Transition Model

2.1 Definition

Interval time series data has the following format

$$\{ \quad (x_t, y_t), t = 1, \ldots T \quad \},$$

where x_t is the upper bound and y_t the lower bound of the interval observed at time t and it is required that $x_t \geqslant y_t$. Denote the vector $Y_t = (x_t, y_t)'$. We say that Y_t is generated by a truncated mixture transition ($TMT(P, Q)$) model if the conditional density function of the

process can be written as

$$f(Y_t|\mathcal{F}^{t-1}) = \sum_{j=1}^{P} \alpha_j f_j(Y_t|Y_{t-Q}^{t-1}), \tag{2.1}$$

$$\sum_{j=1}^{P} \alpha_j = 1, \alpha_j > 0, j = 1, \ldots, P,$$

where $\mathcal{F}^{t-1}$ is the information set up to time $t-1$, P is the number of components, assumed to be fixed, Q is the number of lags in each component,[6] and $Y_{t-Q}^{t-1} = (Y_{t-Q}, Y_{t-Q+1}, ..., Y_{t-1})$. The function $f_j(Y_t|Y_{t-Q}^{t-1})$ is a truncated bivariate normal probability density, truncated at $x_t \geqslant y_t$ so that the upper bound is not smaller than the lower bound. The truncated density has the following functional form (see e.g. Nath, 1972)

$$f_j(Y_t|Y_{t-Q}^{t-1}) = \frac{1}{2\pi\sqrt{|\Sigma_j|}F_{t,j}} \exp[-\frac{1}{2}(Y_t - \mu_{t,j})'\Sigma_j^{-1}(Y_t - \mu_{t,j})], \tag{2.2}$$

where the pseudo location is a linear function of the information set, i.e., $\mu_{t,j} = C_j + B_{j,1}Y_{t-1} + ... + B_{j,Q}Y_{t-Q}$, with constant vector C_j (2×1) and matrices $B_{j,r}$ (2×2) $(r = 1, ..., Q)$; Σ_j (2×2) is a positive semi-definite matrix, and $|\Sigma_j|$ is the determinant of Σ_j. We denote $A_j = (C_j, B_{j,1}, ..., B_{j,Q})$, and hence the parameter set of the model is $\Psi = \{\alpha_j, A_j, \Sigma_j | \forall j\}$. The functional form (2.2) differs from a bivariate normal distribution in the normalization term $F_{t,j} = 1 - \Phi(\frac{-w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})$, which represents the cumulative distribution of the truncated area $(x_t \geq y_t)$, with Φ being the standard normal cumulative distribution function and $w = (1, -1)'$.

According to Extreme Value Theory, extreme processes, i.e. max/min, asymptotically follow a generalized extreme value (GEV) distribution. In Lin and González-Rivera (2019), the marginal conditional means of the maximum and minimum returns are modeled based on results involving GEV distributions. However, as they do not consider the joint modeling of the extremes, the interval constraint cannot be guaranteed. The TMT model provides a framework to accommodate these issues. The finite mixture of truncated normal densities provides a flexible approximation to the GEV distribution because the data, via information selection criteria, determines the number of components and the weight of each component in the mixture as well as the dynamic truncation in each component. The idea of using finite

mixture distributions to approximate certain distributions is not uncommon, particularly for Bayesian inference (see Shephard, 1994; Chib *et al.*, 2002; and Nakajima *et al.*, 2017).

2.2 Conditional moments

From (2.1) and (2.2), we obtain the conditional mean of Y_t as:

$$E(Y_t|\mathcal{F}^{t-1}) = \sum_{j=1}^{P} \alpha_j (M^1_{o,t,j} + \mu_{t,j}), \tag{2.3}$$

where

$$M^1_{o,t,j} = \frac{\Sigma_j w}{\sqrt{w'\Sigma_j w}} \frac{\phi(\frac{-w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}{1 - \Phi(\frac{-w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}, \tag{2.4}$$

and ϕ is the standard normal density function. The term $M^1_{o,t,j}$ represents the mean shift after the truncation takes place (see Nath, 1972, for moments of truncated normal distribution). If there is no truncation, the term $M^1_{o,t,j} = 0$. Then for the component j, the conditional mean is no longer $\mu_{t,j}$ but a nonlinear function of Y^{t-1}_{t-Q}. The natural order of interval time series is also preserved at the conditional mean level, i.e., $w'E(Y_t|\mathcal{F}^{t-1})) \geq 0$ (see Appendix A.1).

An important feature of TMT model (2.1) is that captures conditional heteroscedasticity. The conditional variance is time varying and is calculated as follows

$$\begin{aligned} & V(Y_t|\mathcal{F}^{t-1}) \\ =& E(Y_t Y_t'|\mathcal{F}^{t-1}) - E(Y_t|\mathcal{F}^{t-1})E(Y_t|\mathcal{F}^{t-1})' \\ =& \sum_{j=1}^{P} \alpha_j (M^2_{o,t,j} + \mu_{t,j}(M^1_{o,t,j})' + M^1_{o,t,j}\mu'_{t,j} + \mu_{t,j}\mu'_{t,j}) \\ & -(\sum_{j=1}^{P} \alpha_j (M^1_{o,t,j} + \mu_{t,j}))(\sum_{j=1}^{P} \alpha_j (M^1_{o,t,j} + \mu_{t,j}))', \end{aligned} \tag{2.5}$$

where

$$M^2_{o,t,j} = \Sigma_j + \frac{\Sigma_j w w' \Sigma_j}{w'\Sigma_j w} \frac{-w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}} \frac{\phi(\frac{-w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{-w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}. \tag{2.6}$$

If there is no truncation, in addition to $M^1_{o,t,j} = 0$, we have $M^2_{o,t,j} = \Sigma_j$ and, for each component j, its variance becomes constant.

3 Maximum Likelihood Estimation: EM algorithm

We discuss the estimation of the TMT model (2.1) by maximum likelihood (ML).[7] Our goal is to estimate Ψ by maximizing the likelihood function:

$$L(\Psi) \quad = \frac{1}{T-Q} \quad \sum_{t=Q+1}^{T} \log[\sum_{j=1}^{P} \alpha_j f_j(Y_t | Y_{t-Q}^{t-1}, A_j, \Sigma_j)]. \tag{3.1}$$

It is obvious that a closed-form solution is not achievable by maximizing (3.1). The likelihood functions of mixture models are usually non-concave, and often have several local maxima (see e.g. Redner and Walker, 1984). Dempster *et al.* (1977) propose the expectation maximization (EM) algorithm, which has been widely applied to find the ML estimators for mixture models due to its simplicity and monotonicity property. The statistical properties of EM algorithm have been studied extensively (see Wu, 1983; Meng, 1994; McLachlan and Krishnan, 2007; and Balakrishnan *et al.*, 2017).

Lee and Scott (2010) apply the EM algorithm to a truncated normal mixture model with each component truncated in a rectangular fashion, e.g., $s \leq Y_t \leq k$, where s and k are vectors with the same dimension as Y_t. Although our model has a different type of truncation ($x_t \geq y_t$, or $w'Y_t \geq 0$) , their arguments can be adapted to derive an EM algorithm. However, this modified EM algorithm will not provide a closed-form solution in the M step, mainly due to the truncation term ($\frac{\phi(.)}{1-\Phi(.)}$) in the density function (see Appendix A.2 for details). As a result, numerical maximization is needed in the M step (see Lange, 1995) and thus, the simplicity of the EM algorithm is lost. In the following section, we propose a new EM algorithm that solves this problem.

3.1 A new EM algorithm

As with any EM algorithm, we begin with the data augmentation procedure. Unlike the data generating process specified by the model, where only the observation Y_t is generated at time t, the data augmentation involves generating additional latent data.[8] To obtain the observation Y_t, we first generate a latent variable z_t from a multinomial distribution that will indicate the component of the mixture distribution from which the observation would be coming from. Specifically, $z_t = (z_{t1}, z_{t2}, ..., z_{tP})$, where $z_{tj} \in \{0, 1\}$ is the indicator variable such that $z_{tj} = 1$ if Y_t is generated from component j and 0 otherwise. Next, conditional on z_t, we generate another latent variable n_t from a geometric distribution that indicates the number of invalid draws ($x_t < y_t$) from the respective component before a valid draw ($x_t \geq y_t$) arrives. The valid $(n_t + 1)^{th}$ draw is then treated as the t^{th} observation (Y_t). Clearly, the data (Y_t) is augmented by introducing z_t, n_t, and all the invalid draws. Denote $Y_t^A = \{Y_{t,1}, Y_{t,2}, ..., Y_{t,n_t}, Y_{t,n_t+1}\}$ as all the draws at time t. We now formalize the above data augmentation process (i.e., pseudo complete data generating process thereafter).

Let z_t follow a multinomial distribution:

$$g(z_t|\Psi) \quad = \quad \prod_{j=1}^{P} \alpha_j^{z_{tj}}. \tag{3.2}$$

where $\prod_{j=1}^{P} \alpha_j^{z_{tj}} = \alpha_1^{z_{t1}} \alpha_2^{z_{t2}} \cdots \alpha_P^{z_{tP}}$. Given the role n_t plays in the above pseudo complete data generating process, it is natural to specify its distribution, conditional on z_t, as a geometric distribution, a discrete probability distribution that describes the number of failures before the first occurrence of success, i.e.,

$$q(n_t|z_t, \Psi) \quad = \quad \prod_{j=1}^{P} \Big[(1 - F_{t,j})^{n_t} F_{t,j}\Big]^{z_{tj}}, \tag{3.3}$$

where $F_{t,j} = 1 - \Phi(\frac{-w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})$ is the cumulative distribution function of the truncated area ($x_t \geq y_t$) for component j at time t, and represents the probability of obtaining a valid draw from the bivariate normal distribution. Then, the conditional density of Y_t^A is specified as

follows,

$$h(Y_t^A|z_t,n_t,\Psi) = \prod_{j=1}^{P}\left[\frac{f_{t,j}^N(Y_{t,n_t+1})}{F_{t,j}}\prod_{k=1}^{n_t}(\frac{f_{t,j}^N(Y_{t,k})}{1-F_{t,j}})\right]^{z_{tj}}, \tag{3.4}$$

where $f_{t,j}^N(.)$ is the bivariate normal density of component j at time t.

Next, we construct the joint density function of the pseudo complete data ($\{Y_t^A, z_t, n_t\}$), i.e.,

$$\begin{aligned} l(Y_t^A,z_t,n_t|\Psi) &= g(z_t|\Psi)q(n_t|z_t,\Psi)h(Y_t^A|z_t,n_t,\Psi) \\ &= \prod_{j=1}^{P}\left[\alpha_j f_{t,j}^N(Y_{t,n_t+1})\prod_{k=1}^{n_t} f_{t,j}^N(Y_{t,k})\right]^{z_{tj}}, \end{aligned} \tag{3.5}$$

so that we write the pseudo complete log-likelihood function as follows

$$L^C(\Psi) = \frac{1}{T-Q}\sum_{t=Q+1}^{T}\sum_{j=1}^{P} z_{tj}[\log\alpha_j + \log f_{t,j}^N(Y_{t,n_t+1}) + \sum_{k=1}^{n_t}\log f_{t,j}^N(Y_{t,k})]. \tag{3.6}$$

E Step. The above likelihood (3.6) is replaced with its conditional expectation (see Appendix A.3 for details),

$$\begin{aligned} &Q(\Psi|\Psi^l) \\ =&E[L^C(\Psi)|Y,\Psi^l] \\ =&\frac{1}{T-Q}\sum_{t=Q+1}^{T}\sum_{j=1}^{P}\tilde{z}_{tj}[\log\alpha_j + \log f_{t,j}^N(Y_{t,n_t+1}) + \tilde{n}_{t,j}(\int \log f_{t,j}^N(Y_{t,k})(\frac{f_{t,j}^{N,l}(Y_{t,k})}{1-F_{t,j}^l})dY_{t,k})], \end{aligned} \tag{3.7}$$

where $\tilde{n}_{t,j} = E(n_t|z_{tj}=1,Y,\Psi^l) = \frac{1-F_{t,j}^l}{F_{t,j}^l}$, $f_{t,j}^{N,l}(.)$ and $F_{t,j}^l$ are respectively $f_{t,j}^N(.)$ and $F_{t,j}$ conditional on Ψ^l (the parameter set of the previous (l^{th}) iteration).

$$\begin{aligned} \tilde{z}_{tj} &= P(z_{tj}=1|Y,\Psi^l) \\ &= \frac{P(z_{tj}=1,Y_t|\Psi^l)}{P(Y_t|\Psi^l)} \\ &= \frac{\alpha_j^l f_{t,j}^l(Y_t)}{\sum_{r=1}^{P}\alpha_r^l f_{t,r}^l(Y_t)}. \end{aligned} \tag{3.8}$$

M Step. By maximizing $Q(\Psi|\Psi^l)$, we obtain the iterated rules for Ψ (see Appendix A.4 for

details)

$$\alpha_j^{l+1} = \frac{\sum_{t=Q+1}^{T} \tilde{z}_{tj}}{T-Q}, \tag{3.9}$$

$$A_j^{l+1} = (\bar{X}_j' \bar{Y}_j + \tilde{X}_j' \tilde{M}_{d',\bar{T},j}^1)'(\bar{X}_j' \bar{X}_j + \tilde{X}_j' \tilde{X}_j)^{-1}, \tag{3.10}$$

$$\Sigma_j^{l+1} = \frac{\sum_{t=Q+1}^{T} \tilde{z}_{tj}[(Y_t - A_j^{l+1} X_{t-1})(Y_t - A_j^{l+1} X_{t-1})' + \tilde{n}_{t,j} M_{d',t,j}^2]}{\sum_{t=Q+1}^{T} \tilde{z}_{tj}(1+\tilde{n}_{t,j})}, \tag{3.11}$$

where $\tilde{M}_{d',\bar{T},j}^1 = (\tilde{M}_{d',Q+1,j}^1, ..., \tilde{M}_{d',T,j}^1)'$, and $\tilde{M}_{d',t,j}^1 = \sqrt{\tilde{z}_{tj}\tilde{n}_{t,j}}(M_{d,t,j}^1 + \mu_{t,j}^l)$.

$M_{d',t,j}^2 = M_{d,t,j}^{2,l} + (\mu_{t,j}^l - \mu_{t,j}^{l+1})(M_{d,t,j}^{1,l})' + (M_{d,t,j}^{1,l})(\mu_{t,j}^l - \mu_{t,j}^{l+1})' + (\mu_{t,j}^l - \mu_{t,j}^{l+1})(\mu_{t,j}^l - \mu_{t,j}^{l+1})'$.

$M_{d,t,j}^{1,l}$ and $M_{d,t,j}^{2,l}$ are respectively $M_{d,t,j}^1$ and $M_{d,t,j}^2$ conditional on Ψ^l,

$$M_{d,t,j}^1 = \frac{-\Sigma_j w}{\sqrt{w'\Sigma_j w}} \frac{\phi(\frac{w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}, \tag{3.12}$$

$$M_{d,t,j}^2 = \Sigma_j + \frac{\Sigma_j w w' \Sigma_j}{w'\Sigma_j w} \frac{w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}} \frac{\phi(\frac{w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{w'\mu_{t,j}}{\sqrt{w'\Sigma_j w}})}. \tag{3.13}$$

Furthermore, $\mu_{t,j}^l = C_j^l + B_{j,1}^l Y_{t-1} + ... + B_{j,Q}^l Y_{t-Q}$.

$\bar{X}_j = \sqrt{\tilde{z}_j \tau_1^{1+2Q}} \odot X$, and $X = (\tau_{T-Q}^1, (Y_Q^{T-1})', ..., (Y_1^{T-Q})')$, where τ_a^b is a vector of ones with dimension $a \times b$.

$\tilde{X}_j = \sqrt{(\tilde{z}_j \odot \tilde{n}_j)\tau_1^{1+2Q}} \odot X$, $\tilde{z}_j = (\tilde{z}_{Q+1,j}, ..., \tilde{z}_{T,j})'$, $\tilde{n}_j = (\tilde{n}_{Q+1,j}, ...\tilde{n}_{T,j})'$, $\bar{Y}_j = \sqrt{\tilde{z}_j \tau_1^2} \odot (Y_{Q+1}^T)'$, and $X_{t-1}' = (1, Y_{t-1}', ..., Y_{t-Q}')$. The operator $\odot$ is the Hadamard product.

Repeat E step and M step until convergence. Clearly, the new EM algorithm provides a closed-form solution. Furthermore, the constraints on parameters are satisfied by construction, i.e., Σ^{l+1} is positive semi-definite, $\sum_{j=1}^{P} \alpha_j^{l+1} = 1$, and $\alpha_j^{l+1} > 0$. Note that A_j has an iterated rule that resembles the formula of the maximum likelihood estimates of the parameters of a vector autoregressive model (VAR). When truncation is not in present, it becomes $A_j^{l+1} = (\bar{X}_j' \bar{Y}_j)'(\bar{X}_j' \bar{X}_j)^{-1}$. Therefore, (3.10) can be viewed as applying VAR estimation to the pseudo complete sample.

4 Monte Carlo Simulations

4.1 Finite Sample Performance of the EM Algorithm

We perform Monte Carlo simulations to evaluate the finite sample performance of the proposed EM algorithm to estimate the parameters of the TMT model.

In Table 1, we show the design of three data generating processes: DGP 1 and DGP 2 are $TMT(2,1)$ and DGP 3 is $TMT(3,1)$). Specifically, DGP 1 considers two components with the binding constraint ($x_t \geq y_t$) in one component but not in the other component. The constraint is not binding at time t if $w'\mu_{t,j} = w'(C_j + B_{j,1}Y_{t-1} + ... + B_{j,Q}Y_{t-Q}) \gg 0$ and Σ_j is relatively small, while it is binding otherwise.[9] Intuitively, when a significant portion of the bivariate density is truncated, i.e., substantial areas of the bivariate contours are above the 45-degree line, the constraint is said to be binding. DGP 2 considers the case where the constraint is binding in both components. DGP 3 considers three components with a non-binding restriction in the first component while binding in the other two components. For these two components, we also consider a low persistence process in one component and a high persistence process in the other. To visualize the constraint, we plot the truncations in DGP 3 in Figure 1. For each component at time t, we are re-centering the component's probability density to the origin (i.e., shifting the density by $\mu_{t,j}$, where $\mu_{t,j} = C_j + B_{j,1}Y_{t-1} + ... + B_{j,Q}Y_{t-Q}$). The truncation line, previously the 45 degree line, becomes time-varying as it has also been shifted by $\mu_{t,j}$. In Figure 1, the grey lines are the 45-degree truncation lines that were shifted at each time t. In panels (a) and (b), we show how the binding constraints significantly truncate the bivariate densities while is panel (c) there is no truncation because the constraint is not binding.

For each DGP, the data are generated as follows. First, we set the parameters as in Table 1. Second, we define $\eta_{t,j}$ to be the weight for component j at time t before the truncation is imposed such that the truncation delivers a new component weight α_j.[10] The relationship between α_j and $\eta_{t,j}$ is the following: $\alpha_j = \frac{\eta_{t,j}F_{t,j}}{\sum_{j=1}^{P}\eta_{t,j}F_{t,j}}$. Notice that α_j is fixed while $\eta_{t,j}$ changes with time. Third, independent random draws (e.g. 1000 draws) are extracted from

the bivariate normal mixture distribution (with component weight $\eta_{t,j}$). Fourth, we keep the draws that satisfy the constraint $x_t \geq y_t$, from which one is randomly selected as the actual observation at time t. Repeat the above steps until we generate a sample with desired size.[11]

We initialize the EM algorithm by randomly choosing 50 initial values of the parameter vector. [12] For each initial value of the parameter vector, we run the EM algorithm separately. We choose the values that achieve the highest likelihood. We consider two sample sizes ($T = 200$ and $T = 1,000$) and we run 100 Monte Carlo replications.

We summarize the Monte Carlo results in Tables 2, 3 and 4. Across all the experiments, the EM algorithm performs satisfactorily. Even in small samples, the estimates are rather precise. As the sample size increases, the point estimates are closer to the true parameter values and the standard errors become smaller, as we expected. Whether the constraint is binding or not does not seem to affect the estimation results in any particular fashion. The same can be said whether there is low or high persistence in the conditional means.

Across Tables 2-4, it is interesting to observe that the standard errors of the estimates are smaller in those components whose weight α is the larger. For example, in Table 4, the standard errors for C, B, and Σ are the smallest for the first component that has the largest component weight of 0.5, and they increase for the second and third component that have smaller weights, 0.3 and 0.2 respectively. A possible explanation is that with a smaller component weight, relatively fewer observations would be generated from this component and hence less information is available to accurately estimate this component. As the sample size becomes larger, such differences in standard errors across components shrink since all the standard errors decrease with a large sample size. This suggests that a relatively large sample size would be desirable for more accurate estimation when small components are present.

In summary, Monte Carlo results seem to suggest that the proposed EM algorithm works very efficiently identifying the component weights, the dynamics of the conditional means in each component, and the dynamic truncation (binding or not binding constraints) in the bivariate density of each component. The two factors that contribute to more efficient estimates are the sample size and the weight of each component in the mixture. A larger

sample size and a large weight provide more information and consequently, we obtain smaller standard errors.

4.2 Asymptotic Normality of the MLE

We aim to provide evidence of the asymptotic normality of the ML estimator based on the proposed EM algorithm. This is particularly challenging in a simulation setting because we are facing the "label switching" issue, that is, in the mixture model it is not possible to identify to which component the parameter estimates belong. Though Yao (2013) proposed some methods to mitigate the label switching issue in simulation experiments, it is not fully eliminated and, to the best of our knowledge, there is not a satisfactory solution in the current literature. Therefore, we only consider a special case of the TMT model where the label switching issue does not exist.

We introduce a restricted version of the TMT model, $RTMT(P)$, where we impose the restriction that in each component, there is only one lag, that is, the pseudo location looks like $\mu_{t,j} = C_j + B_{j,j}Y_{t-j}$ and the matrix B is restricted such that $B_{j,r} = 0$ for $r \neq j$. Obviously, the restricted model does not suffer from the label switching issue as each component has different lags. We consider a $RTMT(2)$, with parameter values set as DGP 1 in Table 1 with the discussed restriction. The first component that only includes regressor Y_{t-1} in the pseudo location has a weight of 0.6. The second component only includes regressor Y_{t-2} in the pseudo location and has a weight of 0.4. We perform 500 Monte Carlo simulations and we obtain the ML estimators by implementing the proposed proposed EM algorithm.[13] We also consider small and large samples ($T = 50$ and $T = 500$).

We present the simulation results in Figures 2 to 5 where we plot histograms of the parameter estimates and QQ plots. [14] The common feature in the four figures is that in small samples the estimators depart from normality but, as the sample size increases, we observe a gradual approach towards normality. In Figure 2, the histogram and QQ Plot of the estimate of the constant vectors C_j display fatter right and left tails than those of a normal density. However, in Figures 3 to 5, the approach towards normality when $T = 500$ is more evident,

and even in the small sample $T = 50$ environment, the normal density seems to be a good approximation for the density of the ML estimators of the pseudo location parameters and the weight parameter.

5 Empirical Application

We model the interval-valued time series of the IBM daily stock returns. The high/low returns are calculated as the percentage change of the highest/lowest daily price with respect to the closing price of the previous day. The high return at time t is $r_{high,t} = 100 \times (P_{high,t} - P_{close,t-1})/P_{close,t-1}$ and similarly the low return $r_{low,t} = 100 \times (P_{low,t} - P_{close,t-1})/P_{close,t-1}$. Consequently, the interval-valued time series satisfies $r_{high,t} \geq r_{low,t}$. In Figure 6, we plot the time series from 2004/1/1 to 2018/4/1 (3584 observations); in blue, the high returns and in red, the low returns. As in any financial time series, heteroscedasticity is a very salient feature with high and low volatility periods in both bounds and low returns that tend to be more volatile than high returns.

5.1 In-sample Evaluation

We estimate and evaluate the in-sample performance of the models with the entire sample from 1/1/2014 to 4/1/2018. We start by considering a TMT model with a maximum of seven components and four lags. That is, $P = \{2, ..., 7\}$, and $Q = \{1, 2, 3, 4\}$, for a total of 28 specifications.[15] We select the best models by the BIC. The selected model is a $TMT(4, 2)$ and we report the estimation results in Table 5.[16]. We observe that the first two components account for 75% of the dynamics of the series and the first three components for 90% and are components with relative small volatility (small Σ) while the fourth component has a lower weight (about 9%) but captures periods of high volatility (large Σ). Components 1 and 2 seem to have less persistence than components 3 and 4. In most cases, the upper bound is positively affected by its own lags while negatively affected by the lags of lower bound. Similarly, the lower bound is positively affected by its own lags while negatively affected by

the lags of upper bound. The standard errors for the parameter estimates (C, B, and Σ) increase for components with smaller weights. This aligns with the observation from Section 4.

In summary, the importance of the estimation results shown in Table 5 lies on their contributions to our understanding of the dynamic truncations in the conditional density warranted by the data as well as the estimation of the conditional means, variances, and correlation. We proceed to analyze these features.

In Figure 7, we show the time-varying truncations (as explained in the simulation section 4.1) in the bivariate density of each component after re-centering (shifted by $\mu_{t,j}$ for each t and each j). The truncations are very different across components; the interval constraint is not binding for components 1 and 2 but it is for components 3 and 4, which means that the time-varying heteroscedasticity is driven mainly by these last two components.

In Figure 8, we plot the fitted conditional means (2.3) together with the realized data. The persistency in the data seems to be well captured by the model. We also plot the fitted conditional variances and correlation coefficients (2.5) of the high/low returns. The conditional variances are capturing the volatility clustering in the data very well. The contemporaneous conditional correlations between low and high returns tend to be very high and positive, around 0.8, in low volatility periods and substantially lower in high volatility times.

In Figure 9, we plot two estimated conditional densities of the bivariate process, one in Dec. 18, 2008 and the other in Dec. 29, 2017, to illustrate the flexibility of the truncated normal mixture distribution. The shapes are rather different. In 2008, the density seems to be bimodal and very asymmetric; in contrast in 2017, the density is unimodal and mostly symmetric. The number of components and the truncations provide enough flexibility to adapt to the time-varying conditional density of the data.

We also consider a more parsimonious specification of the model, the restricted model $RTMT(P)$, as described in Section 4.2. We consider up to seven components ($P = \{1, ..., 7\}$) for $RTMT$. We also compare the TMT and $RTMT$ models with five other models. For

all models, the number of lags in the conditional means is selected by BIC. We set up a linear vector autoregressive, VAR, model as the benchmark. We consider two multivariate $GARCH$ models to account for conditional heteroskedasticity in the data, one with a conditional normal density for the errors ($VAR-DCC-N$) and the other with a Student-t density ($VAR-DCC-t$). We also estimate the one-component model proposed by GL. Notice that VAR, $VAR-DCC-N$, and $VAR-DCC-t$ models cannot preserve the natural order of the ITS. In-sample comparison of the six models is summarized in Table 6.

The worst performer is the $VAR(7)$, with the largest BIC and the smallest log-likelihood, though it is the most parsimonious. It is clear that modeling the heteroscedasticity in the data with either of the two multivariate GARCH models improves the performance, in particular when we fit the Student-t to the errors. Neither of these three models considers the natural order of the high/low interval. Among the models that satisfy the interval restriction, the $GL(7)$, one-component model, does not seem to capture enough heteroscedasticity and suggests the need for the introduction of more components. The $RTMT(5)$ and $TMT(4,2)$ are the best performers with the smallest BIC and the largest log-likelihood though there is an increase in the number of parameters to estimate.

5.2 Out-of-sample Evaluation

We split the data into an in-sample period (from 1/1/2004 to 12/31/2013) for model estimation and an out-of-sample period (from 1/1/2014 to 4/1/2018) for model evaluation. We focus on comparing two of the best models, $TMT(4,2)$ and $VAR(7)-DCC-t$, according to the analysis in Section 5.1. As we mention before, the TMT model considers the restriction in the interval bounds but the $VAR-DCC-t$ does not.[17]

Both models are estimated recursively. We construct the one-step-ahead density forecasts in the out-of-sample period. First, we evaluate the density forecasts following Diebold *et al.* (1998) and Diebold *et al.* (1999) by obtaining the corresponding probability integral transformations (PITs) of the densities associated with $r_{high,t}$, $r_{low,t}$, and $r_{high,t}|r_{low,t}$. [18] If the density forecasts coincide with the underlying true conditional densities, then the

PITs should be i.i.d. (i.e., identically and independently distributed) uniformly distributed, $U(0,1)$. In Figures 10 and 11, we plot the PITs for $TMT(4,2)$ and $VAR(7)-DCC-t$.[19] The $TMT(4,2)$ model seems to generate density forecasts that better approximate the underlying true conditional densities when compared with those from the $VAR(7)-DCC-t$ model because its PITs are closer to the uniform distribution. For the $VAR(7)-DCC-t$, there is a clear rejection of uniformity. To assess the dependence of the PITs, we plot the autocorrelation functions of the PITs and those of their squares, third, and fourth powers in Figures 12 and 13.[20] The main difference between these two figures lies in the slightly significant autocorrelations in plots (b) and (d) of the TMT model compared to those generated by the $VAR-DCC-t$ model. It seems that there is slight advantage of the $VAR-DCC-t$ specification on explicitly modeling the heteroscedasticity in the data. For both models, plots (a) and (c) do not exhibit any autocorrelation statistically different from zero.

We also evaluate the density forecasts by a battery of powerful tests, called the Generalized AutoContour (G-ACR) tests, introduced by Gonzalez-Rivera, Senyuz, and Yoldas (2011) and generalized to multivariate densities by Gonzalez-Rivera and Sun (2015). In Figure 14, we plot the PITs of $r_{low,t}$ against those of $r_{high,t}|r_{low,t}$ for both models. If the density forecasts coincide with the underlying true conditional densities, the points in the plots should be uniformly distributed across the area of the unit square and show no dependency. It is clear that comparing the unit squares from the two models, the unit square from the $VAR(7)-DCC-t$ does not exhibit a uniform distribution of the PITs in the square as there are missing points in the lower region of the square. This is not the case for the $TMT(4,2)$ model. We also conduct formal G-ACR tests to test the null hypothesis that the density forecasts coincide with the underlying true conditional densities. We report the results in Tables 7 and Table 8. The null hypothesis cannot be rejected for the TMT model at the 1% significance level except for a few lags and autocontour levels while it is strongly rejected for the $VAR(7)-DCC-t$ model. Such results are consistent with those observed in the plots of Figure 14.

5.3 A Trading Strategy

To demonstrate the usefulness of our model, particularly the density forecasts, we apply a trading strategy developed by González-Rivera et al. (2020). The trading strategy exploits the probability distribution of high/low return forecasts. Consider the ratio $s_t = \frac{\left|O_t - \hat{r}_{low,t,1}\right|}{\left|\hat{r}_{high,t,1} - O_t\right|}$, where O_t is the opening return at day t, calculated using the opening price at day t with respect to the closing price at day $t-1$, and $\hat{r}_{high,t,1}$ and $\hat{r}_{low,t,1}$are the one-step ahead high and low return forecasts, respectively. If $s_t < 1$, then the return is more likely to go up than down in the next day. If this is observed for several days, it is reasonable to believe that the market is forming an upward trend and a "buy alert signal" should be generated. A similar argument can be applied to the "sell alert signal." Figure 15 illustrates the proposed trading strategy. Note that s_t is the absolute value of the slope of any line that connects point $A \equiv (O_t, O_t)$ and any other point below the 45° line. The slope of line AB is equal to (minus) one and it is perpendicular to the 45° line. Hence the area under the 45° line is divided into two areas by the line AB: $s_t > 1$ to the left of line AB, and $s_t < 1$ to the right of line AB. With the predicted probability distribution of high/low return forecasts, the probabilities of $s_t < 1$ and $s_t > 1$ can be computed through numerical integration. Note that González-Rivera et al. (2020) compute the probabilities of $s_t < 1$ and $s_t > 1$ using the Bootstrap method. They count the number of Bootstrap realizations within the prediction region to approximate the probabilities, while we directly integrate over the predicted probability distribution. The trading strategy then consists of the following steps:

- At day t, plot Figure 1 based on O_t. Given the one-step-ahead predictive density of high and low returns, calculate $Prob(s_t < 1)$ and $Prob(s_t > 1)$. If $Prob(s_t < 1) > Prob(s_t > 1)$, a "buy alert signal" is generated.

- If the "buy alert signal" is observed for m consecutive days beginning with day t, buy the asset on day $t + m - 1$ using the closing price on that day.

- After buying the asset, on any other day d,watch for the "sell alert signal"; that is $Prob(s_t < 1) < Prob(s_t > 1)$. If the "sell alert signal" is observed for m consecutive days from day d, sell the asset on day $d + m - 1$ using the closing price on that day.

Otherwise, hold the asset.

We evaluate this trading strategy over the out-of-sample period (January 1, 2014, to April 1, 2018) for $TMT(4,2)$ and $VAR(7)-DCC-t$ models.. For the implementation, the choice of m should not be too small because it will introduce substantial noise in trading, but it should not be too large either because we could miss profitable trades. We consider m = 4 and 5. We apply a transaction cost of 0.1%, and we annualize the profit/loss for each trade because each trade will have a different holding period. The annualized return is calculated as $AR_t = \left(\frac{P_{close,t+j}-P_{close,t}}{P_{close,t}} - 0.001\right)\left(\frac{365}{j}\right)$, where $P_{close,t+j}$ $(j>0)$ and $P_{close,t}$ are the closing prices for the selling and buying days, respectively. The investor can buy the asset again before the previous bought asset is sold. At the end of the evaluation period, if there are still assets that have not been sold, these assets will not be considered when calculating the profits.

Table 9 reports the average of AR_t. TMT model achieves on average higher profit than $VAR-DCC-t$ for $m=4$. For $m=5$, although both models incur losses, loss from the TMT model is on average less than $VAR-DCC-t$. Figure 16 plots the histograms of profits/losses over trades for two models. It is interesting to see that the TMT model is able to pick up a few very profitable trades on the right tails of the histograms while the $VAR-DCC-t$ model misses them.

6 Conclusions

We have proposed a truncated mixture transition model for the interval-valued time series that guarantees the natural order of the data (upper bound greater than lower bound).The model enjoys great flexibility in terms of both parameter and density specifications and captures data features such as heteroscedasticity and non-Gaussianity. However, the standard EM algorithm to estimate truncated mixture models does not provide closed-form solutions in the M step. Therefore, we have proposed a new EM algorithm with a novel data augmentation process that encloses a closed-form solution in the M step. We prove the consistency of the maximum likelihood estimator and simulation results indicate good convergence prop-

erties of the estimator in small and large samples. We have illustrated the performance of the model with an application to the IBM daily high/low stock returns and provided evaluation metrics in-sample and out-of-sample. We have also offered a comparison with several competing specifications and have shown the advantages of the truncated mixture transition model in generating the best density forecasts among the models considered and implemented a trading strategy that delivers better results when density forecasts are based on those generated by the truncated mixture model.

Figure Legends

Figure 1: Trading Strategy Comparison for IBM average annualized returns over the out-of-sample period from January 1, 2014 to April 1, 2018.

Figure 2: Histogram and QQ plot of the first element of C_1 (true value is -2). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 3: Histogram and QQ plot of the first element of $B_{1,1}$ (true value is 0.7). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 4: Histogram and QQ plot of α_1 (true value is 0.6). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 5: Histogram and QQ plot of the first element of Σ_1 (true value is 0.4). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 6: Daily IBM high/low stock returns (2004/1/1 to 2018/4/1).

Figure 7: Truncations in the bivariate density of each component of the model $TMT(4, 2)$.

Figure 8: Estimated conditional mean, variance and correlation of daily IBM high/low stock returns (2004/1/1 to 2018/4/1).

Figure 9: Estimated conditional bivariate density contours.

Figure 10: PITs from $TMT(4, 2)$ density forecasts.

Figure 11: PITs from VAR(7)-DCC-t density forecasts.

Figure 12: ACF of functions of PITs extracted from the $r_{low,t}$ densities generated by $TMT(4, 2)$ model. p_t is the PIT and $\bar{p}$ is the sample mean of p_t.

Figure 13: ACF of functions of PITs extracted from the $r_{low,t}$ densities generated by VAR(7)-DCC-t model. p_t is the PIT and $\bar{p}$ is the sample mean of p_t.

Figure 14: G-ACR plots for TMT(4,2) and VAR(7)-DCC-t models.

Figure 15: Buy and sell signals from trading strategy.

Figure 16: Histograms of the annualized trading returns over the out-of sample period from January 1, 2014 to April 1, 2018.

Figure 17: Inverse Mill Ratio.

References

[1] Albert, J.H. and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88: 669–679.

[2] Amemiya, T. 1973. Regression analysis when the dependent variable is truncated normal. *Econometrica* 41: 997–1016.

[3] Barndorff-Nielsen, O. 1965. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications* 12: 115–121.

[4] Bauwens, L., S. Laurent, and J. V. K. Rombouts. 2006. Multivariate GARCH models: a survey. *Journal of Applied Econometrics* 21: 79–109.

[5] Berchtold, A. and A. Raftery. 2002. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science* 17: 328–356.

[6] Chang, S.-H., P. C. Cosman, and L. B. Milstein. 2011. Chernoff-Type bounds for the Gaussian error function. *IEEE Transactions on Communications.* 59: 2939–2944.

[7] Chib, S. 1992. Bayes inference in the Tobit censored regression model. *Journal of Econometrics* 51: 79–99.

[8] Chib, S., F. Nardari, and N. Shephard. 2002. Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics* 108: 281–316.

[9] Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39: 1–38.

[10] Diebold, F.X., T. A. Gunther, and A. S. Tay. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39: 863–883.

[11] Diebold, F.X., J. Hahn, and A. S. Tay. 1999. Multivariate density forecast evaluation and calibration In financial risk management: high-frequency returns on foreign exchange. *The Review of Economics and Statistics* 81: 661–673.

[12] Fong, P.W., W. K. Li, C. W. Yau, and C. S. Wong. 2007. On a mixture vector autoregressive model. *The Canadian Journal of Statistics* 35: 135–150.

[13] González-Rivera, G. and W. Lin. 2013. Constrained regression for interval-valued data. *Journal of Business and Economic Statistics* 31: 473–490.

[14] González-Rivera, G. and Y. Sun. 2015. Generalized autocontours: evaluation of multivariate density models. *International Journal of Forecasting* 31: 799–814.

[15] González-Rivera, G., Z. Senyuz, and E. Yoldas. 2011. Autocontours: dynamic specification testing. *Journal of Business & Economic Statistics* 29: 186–200.

[16] González-Rivera, G., Y. Luo, and E. Ruiz. 2020. Prediction regions for interval-valued time series. *Journal of Applied Econometrics* 35: 373–390.

[17] Hamilton, J. D. 1990. Analysis of time series subject to changes in regime. *Journal of Econometrics* 45: 39–70.

[18] Hassan, M.Y. and K.-S. Lii. 2006. Modeling marked point processes via bivariate mixture transition distribution models. *Journal of the American Statistical Association* 101: 1241–1252.

[19] Kalliovirta, L., M. Meitz, and P. Saikkonen. 2016. Gaussian mixture vector autoregression. *Journal of Econometrics,* 192: 485–498.

[20] Krengel, U. 1985. *Ergodic Theorems*. Berlin: de Gruyter.

[21] Le, N.D., R. D. Martin, and A. E. Raftery. 1996. Modeling flat stretches, bursts and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association* 91: 1504–1515.

[22] Lee, G. and C. Scott. 2012. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis* 56: 2816–2829.

[23] Leroux, B. 1992. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications* 40: 127–143.

[24] Lin, W., and G. González-Rivera. 2019. Extreme returns and intensity of trading. *Journal of Applied Econometrics* 34: 1121–1140.

[25] McNeil, A.J. and R. Frey. 2000. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance* 7: 271–300.

[26] Nakajima, J., T. Kunihama, and Y. Omori. 2017. Bayesian modeling of dynamic extreme values: extension of generalized extreme value distributions with latent stochastic processes. *Journal of Applied Statistics* 44: 1248–1268.

[27] Nath, G. B. 1972. Moments of a linearly truncated bivariate normal distribution. *Australian Journal of Statistics* 14: 97–102.

[28] Polson, N. G., J. G. Scott, and J. Windle. 2013. Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association* 108: 1339–1349.

[29] Potscher, B. M. and I. R. Prucha. 1991. Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part i: consistency and approximation concepts. *Econometric Reviews* 10: 125–216.

[30] Rao, R. R. 1962. Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics* 33: 659–680.

[31] Shephard, N. 1994. Partial non-Gaussian state space. *Biometrika* 81: 115–131.

[32] Straumann, D. and T. Mikosch. 2006. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics* 34: 2449–2495.

[33] Tanner, M. A. and W.H. Wong. 2010. From EM to data augmentation: the emergence of MCMC Bayesian computation in the 1980s. *Statistical Science* 25: 506–516.

[34] Teicher, H. 1961. Identifiability of mixtures. *The Annals of Mathematical Statistics* 32: 244–248.

[35] Teicher, H. 1963. Identifiability of finite mixtures. *The Annals of Mathematical Statistics* 34: 1265–1269.

[36] Wong, C., and W. Li. 2000. On a mixture autoregressive model. *Journal of the Royal Statistical Society. Series B (Methodological)* 62: 95–115.

[37] Wu, C. F. J. 1983. On the convergence properties of the EM Algorithm. *The Annals of Statistics* 11: 95–103.

[38] Yakowitz, S. and J. Spragins. 1968. On the Identifiability of finite mixtures. *The Annals of Mathematical Statistics* 39: 209–214.

[39] Yao, W. 2015. Label switching and its solutions for frequentist mixture models. *Journal of Statistical Computation and Simulation* 85: 1000–1012.

[40] Yao, J.-F. and J.-G. Attali. 2000. On stability of nonlinear AR processes with Markov switching. *Advances in Applied Probability* 32: 394–407.

Appendix

A.1 Proof of $w'E(Y_t|\mathcal{F}^{t-1})) \geq 0$

It is sufficient to show that $w'M^1_{o,t,j} + w'\mu_{t,j} \geq 0$ for all j. Thus, it suffices to prove that

$$\frac{\phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})} \quad \geq \quad \frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}.$$

Let $\lambda = \frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}$. When $\lambda \leq 0$, the above inequality obviously holds.

When $\lambda > 0$, $1-\Phi(\lambda) = \frac{1}{2}\text{erfc}(\frac{\lambda}{\sqrt{2}})$, where erfc is the complementary error function defined as $\text{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^\infty \exp(-t^2)dt$. In addition, $\phi(\lambda) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{\lambda^2}{2})$. The inequality becomes

$$\frac{1}{\sqrt{2\pi}}\exp(-\frac{\lambda^2}{2}) \quad \geq \quad \frac{1}{2}\text{erfc}(\frac{\lambda}{\sqrt{2}})\lambda.$$

Using the property of erfc function: $\text{erfc}(z) \leq \frac{2}{\sqrt{\pi}}\frac{\exp(-z^2)}{z+\sqrt{z^2+\frac{4}{\pi}}}$, when $z > 0$, we have

$$\frac{1}{\sqrt{\pi}}\frac{\exp(-\frac{\lambda^2}{2})\lambda}{\frac{\lambda}{\sqrt{2}}+\sqrt{\frac{\lambda^2}{2}+\frac{4}{\pi}}} \geq \frac{1}{2}\text{erfc}(\frac{\lambda}{\sqrt{2}})\lambda.$$

With this upper bound of $\frac{1}{2}\text{erfc}(\frac{\lambda}{\sqrt{2}})\lambda$, and it suffices to show that

$$\begin{aligned} &\frac{1}{\sqrt{2\pi}}\exp(-\frac{\lambda^2}{2}) \geq \frac{1}{\sqrt{\pi}}\frac{\exp(-\frac{\lambda^2}{2})\lambda}{\frac{\lambda}{\sqrt{2}}+\sqrt{\frac{\lambda^2}{2}+\frac{4}{\pi}}} \\ \iff &1 \geq \frac{1}{\frac{1}{2}+\sqrt{\frac{1}{4}+\frac{2}{\pi\lambda^2}}}, \end{aligned}$$

which obviously holds when $\lambda > 0$.

A.2 The EM algorithm for truncated normal mixture model

Lee and Scott (2010) apply the EM algorithm to the multivariate truncated normal mixture model with each component truncated by a rectangle, e.g., $s \leq Y \leq k$, where s and k are vectors with the same dimension as Y. We adapt their arguments to derive the EM algorithm as below. To demonstrate, the derivations are made without specifying the dynamics of μ_j. The idea remains the same when such dynamics are added. The parameter set to be estimated is denoted as $\Theta = \{\alpha_j, \mu_j, \Sigma_j | \forall j\}$.

It is not difficult to derive the pseudo complete log-likelihood function for Θ:

$$L^C(\Theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} z_{tj} \log \alpha_j + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} z_{tj} \log f_j(Y_t|\mu_j, \Sigma_j), \tag{6.1}$$

where T is the sample size. The EM algorithm begins by initializing the parameter set, Θ^0, followed by the E and M steps.

E Step: Because z_{tj} is not observed, $L^C(\Theta)$ is replaced with its conditional expectation ($Q(\Theta|\Theta^l)$) conditional on the the observed data (Y) and the parameter set from the previous iteration (Θ^l).

$$Q(\Theta|\Theta^l) = E(L^C(\Theta)|Y, \Theta^l) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} \tilde{z}_{tj} \log \alpha_j + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} \tilde{z}_{tj} \log f_j(Y_t|\mu_j, \Sigma_j), \tag{6.2}$$

$$\begin{aligned} \tilde{z}_{tj} &\equiv E(z_{tj}|Y_t, \Theta^l) \\ &= P(z_{tj}|Y_t, \Theta^l) \\ &= \frac{P(z_{tj}, Y_t, \Theta^l)}{P(Y_t, \Theta^l)} \\ &= \frac{\alpha_j^l f_j(Y_t|\mu_j^l, \Sigma_j^l)}{\sum_{k=1}^{P} \alpha_k^l f_k(Y_t|\mu_j^l, \Sigma_j^l)}. \end{aligned} \tag{6.3}$$

M Step:

$$\alpha_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj}}{T}, \tag{6.4}$$

$$\mu_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj} Y_t}{\sum_{t=1}^{T} \tilde{z}_{tj}} - v_j(\mu_j^{l+1}, \Sigma_j^{l+1}), \tag{6.5}$$

$$\Sigma_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj}(Y_t - \mu_j^{l+1})(Y_t - \mu_j^{l+1})'}{\sum_{t=1}^{T} \tilde{z}_{tj}} + I_j(\mu_j^{l+1}, \Sigma_j^{l+1}), \tag{6.6}$$

where $v_j(\mu_j^{l+1}, \Sigma_j^{l+1})$ and $I_j(\mu_j^{l+1}, \Sigma_j^{l+1})$are nonlinear functions of μ_j^{l+1} and Σ_j^{l+1}. Details are discussed in appendix A.2.1.

A.2.1 Derivation of the EM algorithm

Let Y follows a truncated bivariate normal distribution:

$$f(Y) = \frac{1}{2\pi\sqrt{|\Sigma|}[1 - \Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})]} \exp[-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)], \tag{6.7}$$

Denote $Y^o = Y - \mu$, and its first and second moments are given as (Nath 1972):

$M_o^1 = \frac{\Sigma w}{\sqrt{w'\Sigma w}} \frac{\phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})}{1-\Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})}$,

$M_o^2 = \Sigma + \frac{\Sigma w w'\Sigma}{w'\Sigma w} \frac{-w'\mu}{\sqrt{w'\Sigma w}} \frac{\phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})}{1-\Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})}$.

In E step, the conditional expectation of the pseudo complete log-likelihood function can be obtained:

$$Q(\Theta|\Theta^l) = E(L^C(\Theta)|Y, \Theta^l) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} \tilde{z}_{tj}\Bigg[\log\alpha_j - \log 2\pi - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(Y_t-\mu_j)'\Sigma_j^{-1}(Y_t-\mu_j) - \log(1 - \Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}))\Bigg],$$

where $1 - \Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}) = \frac{1}{\sqrt{\pi}} \int_{\frac{-w'\mu_j}{\sqrt{2w'\Sigma_j w}}}^{\infty} \exp(-t^2)dt$.

First, take the derivative of $\log(1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}))$ with respect to μ_j

$$\begin{aligned}\frac{\partial}{\partial \mu_j}\left[\log(1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}))\right] &= \frac{1}{1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}\left\{\frac{1}{\sqrt{\pi}}\frac{w}{\sqrt{2}\sqrt{w'\Sigma_j w}}\exp(-(\frac{-w'\mu_j}{\sqrt{2}\sqrt{w'\Sigma_j w}})^2)\right\} \\ &= \frac{\phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}\frac{w}{\sqrt{w'\Sigma_j w}} \\ &= \frac{ww'M^1_{o,j}}{w'\Sigma_j w},\end{aligned}$$

where $M^1_{o,j}$ is M^1_o with $\mu=\mu_j$ and $\Sigma=\Sigma_j$.

Next, take the derivative of $Q(\Theta|\Theta^l)$ with respect to μ_j

$$\frac{\partial}{\partial \mu_j}[Q(\Theta|\Theta^l)] = \frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{tj}\left[\Sigma_j^{-1}Y_t-\Sigma_j^{-1}\mu_j-\frac{ww'M^1_{o,j}}{w'\Sigma_j w}\right]=0.$$

rearrange the above equation gives:

$$\mu_j = \frac{\sum_{t=1}^{T}\tilde{z}_{tj}Y_t}{\sum_{t=1}^{T}\tilde{z}_{tj}}\text{-}v_j(\mu_j,\Sigma_j),$$

where $v_j(\mu_j,\Sigma_j)=\frac{\Sigma_j ww'M^1_{o,j}}{w'\Sigma_j w}$.

Now, take derivative of $Q(\Theta|\Theta^l)$ with respect to Σ_j to obtain:

$$w'M^2_{o,j}w = w'\Sigma_j w + w'\Sigma_j w(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})[\frac{\phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}],$$

where $M^2_{o,j}$ is M^2_o with $\mu=\mu_j$ and $\Sigma=\Sigma_j$.

Next, take derivative of $\log(1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}))$ with respect to Σ_j

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_j}[\log(1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}))] &= \frac{1}{1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}\left\{\frac{1}{\sqrt{\pi}}[\frac{w'\mu_j}{2\sqrt{2}(w'\Sigma_j w)^{\frac{3}{2}}}ww'\exp(-(\frac{-w'\mu_j}{\sqrt{2}\sqrt{w'\Sigma_j w}})^2)]\right\} \\
&= \frac{1}{2}(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})(\frac{\phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})})(\frac{-ww'}{w'\Sigma_j w}) \\
&= \frac{1}{2}\frac{w'M_{o,j}^2 w - w'\Sigma_j w}{w'\Sigma_j w}(\frac{-ww'}{w'\Sigma_j w}) \\
&= \frac{1}{2}w[\frac{1}{w'\Sigma_j w} - \frac{w'M_{o,j}^2 w}{(w'\Sigma_j w)^2}]w'.
\end{aligned}$$

Then, take the derivative of $Q(\Theta|\Theta^l)$ with respect to Σ_j

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_j}[Q(\Theta|\Theta^l)] &= \frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{tj}\Big\{-\frac{1}{2}\Sigma_j^{-1} + \frac{1}{2}\Sigma_j^{-1}(Y_t-\mu_j)(Y_t-\mu_j)'\Sigma_j^{-1} \\
&\quad -\frac{1}{2}w[\frac{1}{w'\Sigma_j w} - \frac{w'M_{o,j}^2 w}{(w'\Sigma_j w)^2}]w'\Big\} \\
&= 0.
\end{aligned}$$

Some linear algebra properties were used: $\frac{\partial \log|A|}{\partial A} = (A')^{-1}$ and $\frac{\partial x'A^{-1}x}{\partial A} = -A^{-1}xx'A^{-1}$. Finally, it can be shown that:

$$\Sigma_j = \frac{\sum_{t=1}^{T}\tilde{z}_{tj}(Y_t-\mu_j)(Y_t-\mu_j)'}{\sum_{t=1}^{T}\tilde{z}_{tj}} + I_j(\mu_j,\Sigma_j),$$

where $I_j(\mu_j,\Sigma_j) = \Sigma_j w[\frac{1}{w'\Sigma_j w} - \frac{w'M_{o,j}^2 w}{(w'\Sigma_j w)^2}]w'\Sigma_j$.

A.3 E step of the new EM algorithm

$$
\begin{aligned}
&E[L^C(\Psi)|Y,\Psi^l] \\
=&E_{z,n|Y,\Psi^l}\{E[L^C(\Psi)|z,n,Y,\Psi^l]\} \\
=&E_{z,n|Y,\Psi^l}\{E[\frac{1}{T}\sum_{t=Q+1}^{T}\sum_{j=1}^{P} z_{tj}(\log\alpha_j+\log f^N_{t,j}(Y_{t,n_t+1})+\sum_{k=1}^{n_t}\log f^N_{t,j}(Y_{t,k}))|z,n,Y,\Psi^l]\} \\
=&E_{z,n|Y,\Psi^l}\{\frac{1}{T}\sum_{t=Q+1}^{T}\sum_{j=1}^{P} z_{tj}(\log\alpha_j+\log f^N_{t,j}(Y_{t,n_t+1})+n_tE[\log f^N_{t,j}(Y_{t,k})|z,n,Y,\Psi^l])\} \\
=&E_{z|Y,\Psi^l}\{\frac{1}{T}\sum_{t=Q+1}^{T}\sum_{j=1}^{P} z_{tj}(\log\alpha_j+\log f^N_{t,j}(Y_{t,n_t+1})+E(n_t|z,Y,\Psi^l)E[\log f^N_{t,j}(Y_{t,k})|z,n,Y,\Psi^l])\} \\
=&E_{z|Y,\Psi^l}\{\frac{1}{T}\sum_{t=Q+1}^{T}\sum_{j=1}^{P} z_{tj}(\log\alpha_j+\log f^N_{t,j}(Y_{t,n_t+1})+ \\
&(\sum_{n_t=0}^{\infty} n_t\prod_{h=1}^{P}[(1-F^l_{t,h})^{n_t}F^l_{t,h}]^{z_{th}})(\int\log f^N_{t,j}(Y_{t,k})\prod_{m=1}^{P}(\frac{f^{N,l}_{t,m}(Y_{t,k})}{1-F^l_{t,m}})^{z_{tm}}dY_{t,k}))\} \\
=&\frac{1}{T}\sum_{t=Q+1}^{T}\sum_{j=1}^{P} E_{z|Y,\Theta^l}\{z_{tj}(\log\alpha_j+\log f^N_{t,j}(Y_{t,n_t+1})+ \\
&(\sum_{n_t=0}^{\infty} n_t\prod_{h=1}^{P}[(1-F^l_{t,h})^{n_t}F^l_{t,h}]^{z_{th}})(\int\log f^N_{t,j}(Y_{t,k})\prod_{m=1}^{P}(\frac{f^{N,l}_{t,m}(Y_{t,k})}{1-F^l_{t,m}})^{z_{tm}}dY_{t,k}))\} \\
=&\frac{1}{T}\sum_{t=Q+1}^{T}\sum_{j=1}^{P} P(z_{tj}|Y,\Psi^l)[\log\alpha_j+\log f^N_{t,j}(Y_{t,n_t+1})+ \\
&\frac{1-F^l_{t,j}}{F^l_{t,j}}(\int\log f^N_{t,j}(Y_{t,k})(\frac{f^{N,l}_{t,j}(Y_{t,k})}{1-F^l_{t,j}})dY_{t,k})] \\
=&\frac{1}{T}\sum_{t=Q+1}^{T}\sum_{j=1}^{P}\tilde{z}_{tj}[\log\alpha_j+\log f^N_{t,j}(Y_{t,n_t+1})+\tilde{n}_{t,j}(\int\log f^N_{t,j}(Y_{t,k})(\frac{f^{N,l}_{t,j}(Y_{t,k})}{1-F^l_{t,j}})dY_{t,k})],
\end{aligned}
$$

where $E_{z,n|Y,\Psi^l}(.)$ takes the joint expectation of z and n conditional on Y and Ψ^l. Law of iterated expectation $E(Y|X)=E[E(Y|Z,X)|X]$ was used.

A.4 M step of the new EM algorithm

To begin with, we derive the first two moments for Y coming from the invalid truncation area ($x < y$), whose density has the following form:

$$f(Y,\mu,\Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}[1-\Phi(\frac{w'\mu}{\sqrt{w'\Sigma w}})]}\exp[-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)]. \tag{6.8}$$

Let $Y^d = Y - \mu$. Then, the first and second moments of $Y^d = \begin{pmatrix} x^d \\ y^d \end{pmatrix}$ are:

$M_d^1 = \frac{-\Sigma w}{\sqrt{w'\Sigma w}}\frac{\phi(\frac{w'\mu}{\sqrt{w'\Sigma w}})}{1-\Phi(\frac{w'\mu}{\sqrt{w'\Sigma w}})}$,

$M_d^2 = \Sigma + \frac{\Sigma w w'\Sigma}{w'\Sigma w}\frac{w'\mu}{\sqrt{w'\Sigma w}}\frac{\phi(\frac{w'\mu}{\sqrt{w'\Sigma w}})}{1-\Phi(\frac{w'\mu}{\sqrt{w'\Sigma w}})}$.

- It is not difficult to take the derivative of (3.7) with respect to α_j, subject to the restriction that $\sum_{j=1}^P \alpha_j = 1$. One can obtain $\alpha_j^{l+1} = \frac{\sum_{t=Q+1}^T \tilde{z}_{tj}}{T-Q}$.

- Next, take the derivative of (3.7) with respect to Σ_j^{-1}.

$$\frac{\partial Q(\Psi|\Psi^l)}{\partial \Sigma_j^{-1}} = \frac{1}{T-Q}\sum_{t=Q+1}^T \tilde{z}_{tj}[\frac{1}{2}\Sigma_j - \frac{1}{2}(Y_t-\mu_{t,j}^{l+1})(Y_t-\mu_{t,j}^{l+1})' +$$
$$\tilde{n}_{t,j}\int(\frac{1}{2}\Sigma_j - \frac{1}{2}(Y_{t,k}-\mu_{t,j}^{l+1})(Y_{t,k}-\mu_{t,j}^{l+1})')(\frac{f_j^{N,l}(Y_{t,k})}{1-F_j^l})dY_{t,k}] = 0$$
$$\Rightarrow \sum_{t=Q+1}^T \tilde{z}_{tj}\Sigma_j - \sum_{t=Q+1}^T \tilde{z}_{tj}(Y_t-\mu_{t,j}^{l+1})(Y_t-\mu_{t,j}^{l+1})' + \sum_{t=Q+1}^T \tilde{z}_{tj}\tilde{n}_{t,j}\Sigma_j - \sum_{t=Q+1}^T \tilde{z}_{tj}\tilde{n}_{t,j}M_{d',t,j}^2 = 0$$
$$\Rightarrow \Sigma_j^{l+1} = \frac{\sum_{t=Q+1}^T \tilde{z}_{tj}[(Y_t-\mu_{t,j}^{l+1})(Y_t-\mu_{t,j}^{l+1})' + \tilde{n}_{t,j}M_{d',t,j}^2]}{\sum_{t=Q+1}^T \tilde{z}_{tj}(1+\tilde{n}_{t,j})},$$

where $\mu_{t,j}^{l+1} = A_j^{l+1}X_{t-1}$, $M_{d',t,j}^2 = M_{d,t,j}^{2,l} + (\mu_{t,j}^l - \mu_{t,j}^{l+1})(M_{d,t,j}^{1,l})' + (M_{d,t,j}^{1,l})(\mu_{t,j}^l - \mu_{t,j}^{l+1})' + (\mu_{t,j}^l - \mu_{t,j}^{l+1})(\mu_{t,j}^l - \mu_{t,j}^{l+1})'$, $M_{d,t,j}^{1,l}$ and $M_{d,t,j}^{2,l}$ are M_d^1 and M_d^2 with $\mu = \mu_{t,j}^l$, $\Sigma = \Sigma_j^l$.

- Finally, take the derivative of (3.7) with respect to A_j.

Notice that maximizing $Q(\Psi|\Psi^l)$ is equivalent to minimizing the following expression for the

purpose of taking derivative with respect to A_j:

$$\begin{aligned}
L(A) &= \sum_{t=Q+1}^{T}\sum_{j=1}^{P}\tilde{z}_{tj}[(Y_t - A_jX_{t-1})'\Sigma_j^{-1}(Y_t - A_jX_{t-1}) + \\
&\quad \tilde{n}_{t,j}\int^{Tr}((Y_{t,k} - A_jX_{t-1})'\Sigma_j^{-1}(Y_{t,k} - A_jX_{t-1}))(\frac{f_{t,j}^{N,l}(Y_{t,k})}{1-F_{t,j}^{l}})dY_{t,k}] \\
&= \sum_{j=1}^{P}\{[\mathrm{vec}(\bar{Y}_j) - (I_2\otimes\bar{X}_j)\mathrm{vec}(A_j')]'(\Sigma_j^{-1}\otimes I_{T-Q})[\mathrm{vec}(\bar{Y}_j) - (I_2\otimes\bar{X}_j)\mathrm{vec}(A_j')] + \\
&\quad \int[\mathrm{vec}(\tilde{Y}_j) - (I_2\otimes\tilde{X}_j)\mathrm{vec}(A_j')]'(\Sigma_j^{-1}\otimes I_{T-Q})[\mathrm{vec}(\tilde{Y}_j) - (I_2\otimes\tilde{X}_j)\mathrm{vec}(A_j')]f_j^l(\tilde{Y}_j)d\tilde{Y}_j\},
\end{aligned}$$

where $\tilde{Y}_j = \sqrt{(\tilde{z}_j\odot\tilde{n}_j)\tau}\odot Y_k$, and $Y_k = (Y_{Q+1,k},...,Y_{T,k})'$. Take the derivative of $L(A)$ with respect to $\mathrm{vec}(A_j')$:

$$\begin{aligned}
&\frac{\partial L(A)}{\partial\mathrm{vec}(A_j')} \\
&= -2(I_2\otimes\bar{X}_j)(\Sigma_j^{-1}\otimes I_{T-Q})\mathrm{vec}(\bar{Y}_j) + 2(I_2\otimes\bar{X}_j)'(\Sigma_j^{-1}\otimes I_{T-Q})(I_2\otimes\bar{X}_j)\mathrm{vec}(A_j') + \\
&\quad \int[-2(I_2\otimes\tilde{X}_j)(\Sigma_j^{-1}\otimes I_{T-Q})\mathrm{vec}(\tilde{Y}_j) + 2(I_2\otimes\tilde{X}_j)'(\Sigma_j^{-1}\otimes I_{T-Q})(I_2\otimes\tilde{X}_j)\mathrm{vec}(A_j')]f(\tilde{Y}_j)d\tilde{Y}_j \\
&= -(I_2\otimes\bar{X}_j)'(\Sigma_j^{-1}\otimes I_{T-Q})\mathrm{vec}(\bar{Y}_j) + (I_2\otimes\bar{X}_j)'(\Sigma_j^{-1}\otimes I_{T-Q})(I_2\otimes\bar{X}_j)\mathrm{vec}(A_j') - \\
&\quad (I_2\otimes\tilde{X}_j)'(\Sigma_j^{-1}\otimes I_{T-Q})\mathrm{vec}(\tilde{M}_{d',\bar{T},j}^1) + (I_2\otimes\tilde{X}_j)'(\Sigma_j^{-1}\otimes I_{T-Q})(I_2\otimes\tilde{X}_j)\mathrm{vec}(A_j') \\
&= -[(\Sigma_j^{-1}\otimes\tilde{X}_j')\mathrm{vec}(\tilde{M}_{d',\bar{T},j}^1) + (\Sigma_j^{-1}\otimes\bar{X}_j')\mathrm{vec}(\bar{Y}_j)] + [(\Sigma_j^{-1}\otimes\bar{X}_j'\bar{X}_j) + (\Sigma_j^{-1}\otimes\tilde{X}_j'\tilde{X}_j)]\mathrm{vec}(A_j') \\
&= -[\mathrm{vec}(\tilde{X}_j'\tilde{M}_{d',\bar{T},j}^1\Sigma_j^{-1}) + \mathrm{vec}(\bar{X}_j'\bar{Y}_j\Sigma_j^{-1})] + [\Sigma_j^{-1}\otimes(\bar{X}_j'\bar{X}_j + \tilde{X}_j'\tilde{X}_j)]\mathrm{vec}(A_j') \\
&= -(\Sigma_j^{-1}\otimes I_2)\mathrm{vec}(\tilde{X}_j'\tilde{M}_{d',\bar{T},j}^1 + \bar{X}_j'\bar{Y}_j) + [\Sigma_j^{-1}\otimes(\bar{X}_j'\bar{X}_j + \tilde{X}_j'\tilde{X}_j)]\mathrm{vec}(A_j') \\
&= 0.
\end{aligned}$$

Then, we can write down $\mathrm{vec}(A_j')$ as:

$$\begin{aligned}
&\mathrm{vec}(A_j') \\
&= [\Sigma_j^{-1}\otimes(\bar{X}_j'\bar{X}_j + \tilde{X}_j'\tilde{X}_j)]^{-1}(\Sigma_j^{-1}\otimes I_2)\mathrm{vec}(\tilde{X}_j'\tilde{M}_{d',\bar{T},j}^1 + \bar{X}_j'\bar{Y}_j) \\
&= (I_2\otimes(\bar{X}_j'\bar{X}_j + \tilde{X}_j'\tilde{X}_j)^{-1})\mathrm{vec}(\tilde{X}_j'\tilde{M}_{d',\bar{T},j}^1 + \bar{X}_j'\bar{Y}_j) \\
&= \mathrm{vec}[(\bar{X}_j'\bar{X}_j + \tilde{X}_j'\tilde{X}_j)^{-1}(\tilde{X}_j'\tilde{M}_{d',\bar{T},j}^1 + \bar{X}_j'\bar{Y}_j)].
\end{aligned}$$

Therefore, we have

$$A_j^{l+1} = (\tilde{X}_j'\tilde{M}^1_{d',\bar{T},j} + \bar{X}_j'\bar{Y}_j)'(\bar{X}_j'\bar{X}_j + \tilde{X}_j'\tilde{X}_j)^{-1}.$$

A.5 Discussion of Stationarity Conditions

Deriving the stationarity conditions is not straightforward as the model is highly nonlinear and the errors (if we write the conditional mean of the model in a regression form) are dependent due to the time-varying truncations. We notice that the TMT model can be interpreted as a special case of a nonlinear autoregressive model with Markov switching, for which the stability problem has been studied by Yao and Attali (2000). However, their results cannot be directly applied as the identically and independently distributed assumption does not hold for the TMT model. The truncation imposed on the distribution of the error term varies over time as a function of the information set, rendering heteroscedastic and dependent errors. Nevertheless, we provide some heuristic and theoretical reasoning to understand stationarity conditions of the model.

Suppose the TMT model has only one component and no truncation (i.e., it becomes a standard bivariate VAR model). If the process generated from this model is stationary, we would expect the process generated from the model with truncation (i.e., one component TMT) to be also stationary because imposing the truncation would not introduce any deterministic or stochastic trend to the process. The same reasoning could be extended to the multiple components case. Because the component weights are fixed, positive, and sum up to one, we have a convex combination of stationary processes, which would also be expected to be stationary. If the process generated from a bivariate mixture autoregressive model is stationary, it is reasonable to expect that the process generated from the TMT model is also stationary. The stationarity conditions for the mixture autoregressive model have been studied (see Wong and Li, 2000; Fong *et al.*, 2007).

Another perspective is to analyze the nonlinearity of the model to understand whether it can be approximated by a linear relationship under certain conditions. The nonlinearity of the TMT model comes from the truncation, which is the inverse Mills ratio (IMR), $\frac{\phi(x)}{1-\Phi(x)}$,

see equation (2.4). IMR is approximately linear for some range of x (see Figure 17 for a plot of IMR).

We analyze IMR into two parts: (a) when IMR goes towards zero on the left side, we have the case where the interval constraint is not binding. In this case, the model looks like a standard VAR. Equations 2.3 and 2.5 are linear. (b) when the interval constraint is binding, we move to right side of Figure 15 and IMR is mostly linear. This makes equations 2.3 and 2.5 linear in these components. Overall, when we have a mixture of components, some with binding constraints, equations 2.3 and 2.5 can be approximately linear. Hence, a stable solution can be found.

We now demonstrate the above reasoning with a simple example. For a TMT model with only one component and binding interval constraint, the regression form of the model can be written as follows

$$Y_t = \mu_t + \frac{\Sigma w}{\sqrt{w'\Sigma w}} \frac{\phi(\frac{-w'\mu_t}{\sqrt{w'\Sigma w}})}{1-\Phi(\frac{-w'\mu_t}{\sqrt{w'\Sigma w}})} + \varepsilon_t$$

where $\varepsilon_t = \begin{pmatrix} \varepsilon_{u,t} \\ \varepsilon_{l,t} \end{pmatrix}$ follows a truncated normal distribution such that $w'\varepsilon_t \geq w'G$, where $G = \mu_t + \frac{\Sigma w}{\sqrt{w'\Sigma w}} \frac{\phi(\frac{-w'\mu_t}{\sqrt{w'\Sigma w}})}{1-\Phi(\frac{-w'\mu_t}{\sqrt{w'\Sigma w}})}$, and $E(\varepsilon_t) = 0$. When the constraint is binding, we can approximate IMR with a Taylor expansion, expanded at $\frac{-w'\mu_t}{\sqrt{w'\Sigma w}} = 0$.

$$\frac{\phi(\frac{-w'\mu_t}{\sqrt{w'\Sigma w}})}{1-\Phi(\frac{-w'\mu_t}{\sqrt{w'\Sigma w}})} = \frac{\phi(0)}{1-\Phi(0)} + S(0)\left[\frac{-w'\mu_t}{\sqrt{w'\Sigma w}}\right]$$

where $S(0)$ is the first derivative of IMR evaluated at zero. $S(0) \simeq 0.6$, and $\frac{\phi(0)}{1-\Phi(0)} \simeq 0.8$. Then,

$$\begin{aligned} Y_t &= \mu_t + \frac{\Sigma w}{\sqrt{w'\Sigma w}}\left[0.8 + 0.6\left[\frac{-w'\mu_t}{\sqrt{w'\Sigma w}}\right]\right] + \varepsilon_t \\ &= C + BY_{t-1} + \frac{\Sigma w}{\sqrt{w'\Sigma w}}\left[0.8 + 0.6\left[\frac{-w'(C+BY_{t-1})}{\sqrt{w'\Sigma w}}\right]\right] + \varepsilon_t \\ &= \Lambda + \mathcal{H}Y_{t-1} + \varepsilon_t \end{aligned}$$

where $\Lambda = C + \frac{\Sigma w}{\sqrt{w'\Sigma w}}\left[0.8 - 0.6\frac{w'C}{\sqrt{w'\Sigma w}}\right]$ and $\mathcal{H} = B - \frac{0.6\Sigma w w'B}{\sqrt{w'\Sigma w}}$ are functions of the model

parameters. The stationary condition would be all values of z satisfying $|I - \mathcal{H}z|$ lying outside of the unit circle.

A.6 Consistency of the ML estimator

We discuss the consistency of the ML estimator. Let $\vartheta_j = \{A_j, \Sigma_j\}$ for $j = 1, ..., P$. The following parameter restrictions are necessary to ensure that the same TMT model cannot be obtained by relabeling the components.

$$\alpha_1 > \alpha_2 > ... > \alpha_P > 0 \text{ and } \vartheta_i = \vartheta_j \text{ only if } 1 \leq i = j \leq P. \tag{6.9}$$

The following theorem shows that under some regular conditions, the MLE is consistent. We begin by imposing the following assumptions:

Assumption 1. The process $\{Y_t\}$ is generated from (2.1) and is strictly stationary and ergodic.

Assumption 2. The true parameter set, Ψ_0, is an interior point of Ξ, where Ξ is a compact subset of $\{\Psi \in (0,1)^{P-1} \times \mathbb{R}^{(5+4Q)P} : 6.9 \text{ holds and } \Sigma_j \text{ are positive definite } \forall j\}$.

Assumption 3. $E(\|Y_t\|^2) < \infty$, where $\|.\|$ is the Euclidean norm.

These assumptions are fairly regular in the literature. Assumption 1 is the most challenging because the model is highly nonlinear, nevertheless we provide an extensive discussion on the stationarity of the model in Appendix A.5. Assumptions 2 and 3 are sufficient to ensure the uniform convergence of the likelihood function.

The following theorem establishes the strong consistency of ML estimator.

Theorem 1. *Under Assumptions 1,2 and 3, the maximum likelihood estimator* $\hat{\Psi} = \underset{\Psi \in \Xi}{argmax} L(\Psi)$ *is strongly consistent, that is* $\hat{\Psi} \to \Psi_0$ *a.s.*

Proof of Theorem 1.

To proof consistency, we will need to show that the finite mixtures of truncated normal distributions are identifiable. The identification of mixture distributions and models have been extensively studied in the literature (see e.g., Teicher (1963), Yakowitz and Spragins (1968), Leroux (1992)). We introduce a lemma that shows the finite mixtures of truncated normal distributions are identifiable up to the label switching.

Lemma 1. *Let $\nu = (\mu, \Sigma)$, and suppose that $\Lambda = \{F(Y,\nu); \nu \in \mathbb{R}^6, Y \in \mathbb{R}^2\}$ is the family of cumulative distribution functions whose density is given by*

$$f(Y,\nu) \quad = \quad \frac{1}{2\pi\sqrt{|\Sigma|}[1-\Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})]} \exp[-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)]. \tag{6.10}$$

Then, the family of finite mixtures of Λ is identifiable up to label switching. That is, $\sum_{i=1}^{P} \alpha_i f(Y,\nu_i) = \sum_{j=1}^{P} \alpha_j f(Y,\nu_j)$ implies that for each $1 \leq i \leq P$, there is some j such that $\alpha_i = \alpha_j$ and $v_i = v_j$ assuming that $\alpha_i's$ and $v_i's$ are respectively distinct.

Proof of Lemma 1.

First, we define distributions that belong to the exponential family for later use.

If, for some $\sigma-$finite measure μ,

$$dG(Y,\tau) \quad = \quad a(\tau)b(Y)\exp[\tau' h(Y)]d\mu(Y), \tag{6.11}$$

for $Y \in \mathbb{R}^n$, $\tau(m \times 1)$, and $h(Y)$ $(m \times 1)$, where $a(\tau) > 0, b(Y) \geq 0$ and a, b, h_j, for $j = 1, 2, \ldots, m$ are all measurable, then G is called an exponential family member.

Let $H(Y) = \sum_{i=1}^{P} \alpha_i G(Y, \tau_i)$ be the finite mixtures. Denote $\mathbb{G}$ the class of all n-dimensional cdf's G and $\mathbb{H}$ the induced class of mixtures of H. Barndorff-Nielsen (1965, Corollary 3) shows that $\mathbb{H}$ is identifiable up to label switching if (a) μ is n-dimensional Lebesgue measure, (b) functions h_j, $j = 1, 2, \ldots, m$, are all continuous, and (c) the set $\{y : y = h(Y), b(Y) > 0, Y \in \mathbb{R}^n\}$ contains a nonempty open set.[21]

The truncated normal distribution $F(Y,\nu)$ whose density is given by 6.10 belongs to expo-

nential family as after re-parameterized it can be written as:

$$\begin{aligned}\frac{dF(Y,\tau)}{d\mu(Y)} &= \frac{1}{2\pi\sqrt{|\Sigma|}[1-\Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})]}\exp[-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)] \\ &= a(\tau)b(Y)\exp[\tau' h(Y)],\end{aligned}$$

where μ is two-dimensional Lebesgue measure. $\tau = \left(\Sigma^{-1}\mu, \ -\frac{1}{2}\vec{(}\Sigma^{-1})\right)$, $a(\tau) = \left\{\sqrt{|\Sigma|}[1-\Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})]\exp(\frac{1}{2}\mu'\Sigma^{-1}\mu)\right\}^{-1}$, $b(Y) = \frac{1}{2\pi}$, and $h(Y) = \left(Y, \ \vec{(}YY')\right)'$.

The image of the mapping h: $\mathbb{R}^2 \to \mathbb{R}^6$, for $x \geq y$ is the set $\Omega = \{h(Y), x \geq y\}$, which contains an open set $\Omega' = \{h(Y), x > y\}$. In addition, the map from τ to ν is unique. Lemma 1 follows. □

Now, we proceed to prove Theorem 1. It is straightforward to see that $L(\Psi)$ is a measurable function of data for each $\Psi \in \Xi$, and continuous in Ψ. Therefore, it suffices to show that (a) the log-likelihood follows a uniform strong law of large numbers: $\underset{\Psi\in\Xi}{sup} \mid L(\Psi) - E[L(\Psi)] \mid \to 0$ a.s. as $T \to \infty$; (b) the identification condition: $E[L(\Psi)] \leq E[L(\Psi_0)]$, and $E[L(\Psi)] = E[L(\Psi_0)]$ implies $\Psi = \Psi_0$. (see Amemiya (1973, Lemma 3)).

Let $L(\Psi) = \frac{1}{T-P}\sum_t l(\Psi)$. By Assumption 1 and continuity of $l(\Psi)$, $l(\Psi)$ is stationary and ergodic (see Krengel (1985, Proposition 4.3)), and hence $E[L(\Psi)] = E[l(\Psi)]$. To verify (a), it suffices to show that $E[\underset{\Psi\in\Xi}{sup} \mid l(\Psi) \mid] < \infty$ (see Rao (1962) or Straumann and Mikosch (2006 Theorem 2.7)). Kalliovirta *et al.* (2016) prove that the above inequality holds for the likelihood in their model. We adapt similar procedures here. It can be obtained that

$$l(\Psi) = \log\{\sum_{j=1}^{P}\alpha_j(2\pi)^{-1}|\Sigma_j|^{-1/2}$$
$$\exp[-\frac{1}{2}(Y_t - A_jX_{t-1})'\Sigma_j^{-1}(Y_t - A_jX_{t-1})]/[\frac{1}{2}\text{erfc}(-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w})]\},$$

where $w = (1,-1)'$. Assumption 2 implies that, $\Delta \geq |\Sigma_j| \geq \delta$, $\forall j$ for some $\delta > 0$, and $\Delta < \infty$, and that $w'\Sigma_j w \geq \gamma$, $\forall j$ for some $\gamma > 0$. Furthermore, $\exp[-\frac{1}{2}(Y_t - A_jX_{t-1})'\Sigma_j^{-1}(Y_t - A_jX_{t-1})] \leq 1$. In addition, when $-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w} \leq 0$, $\text{erfc}(-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w}) \geq 1$, and thus $l(\Psi) \leq \log(\pi^{-1}\delta^{-1/2})$. When $-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w} > 0$, using the inequality

$\text{erfc}(x) \geq \frac{1}{2}\exp(-2x^2)$ to get (see Chang *et al.* (2011, Theorem 2)):

$$\begin{aligned}
\text{erfc}(-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w}) &\geq \frac{1}{2}\exp(-w'A_jX_{t-1}X_{t-1}'A_j'w/w'\Sigma_j w) \\
&\geq \frac{1}{2}\exp[-\frac{1}{\gamma}tr(X_{t-1}X_{t-1}'A_j'ww'A_j)] \\
&\geq \frac{1}{2}\exp[-\frac{1}{\gamma}tr(X_{t-1}X_{t-1}')tr(A_j'ww'A_j)] \\
&\geq \frac{1}{2}\exp[-\frac{\kappa}{\gamma}X_{t-1}'X_{t-1}],
\end{aligned}$$

where the last inequality holds by compactness of Ξ (Assumption 2). That is, $tr(A_j'ww'A_j) \leq \kappa$, $\forall j$ for some $0 < \kappa < \infty$. Now, it can be seen that

$$\begin{aligned}
l(\Psi) &\leq \log\{\sum_{j=1}^{P}\alpha_j(2\pi)^{-1}\delta^{-1/2}4\exp[\frac{\kappa}{\gamma}X_{t-1}'X_{t-1}]\} \\
&= \log(2\pi^{-1}\delta^{-1/2}) + \frac{\kappa}{\gamma}X_{t-1}'X_{t-1}.
\end{aligned}$$

Therefore, regardless of the value of $-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w}$, $l(\Psi) \leq \log(2\pi^{-1}\delta^{-1/2}) + \frac{\kappa}{\gamma}X_{t-1}'X_{t-1}$.

On the other hand, it can be seen that

$$\begin{aligned}
&(Y_t - A_jX_{t-1})'\Sigma_j^{-1}(Y_t - A_jX_{t-1}) \\
=&tr[(Y_t - A_jX_{t-1})(Y_t - A_jX_{t-1})'\Sigma_j^{-1}] \\
\leq&tr[(Y_t - A_jX_{t-1})(Y_t - A_jX_{t-1})']tr(\Sigma_j^{-1}) \\
=&(Y_t - A_jX_{t-1})'(Y_t - A_jX_{t-1})tr(\Sigma_j^{-1}) \\
\leq&(1 + Y_t'Y_t + X_{t-1}'X_{t-1})\rho,
\end{aligned}$$

where the first inequality holds because both $(Y_t - A_jX_{t-1})(Y_t - A_jX_{t-1})'$ and Σ_j^{-1} are positive semi-definite. The second last inequality is implied by Cauchy-Schwarz inequality and Assumption 2 ($tr(\Sigma_j^{-1}) \leq \rho$, $\forall j$ for some $0 < \rho < \infty$). Furthermore, $\text{erfc}(-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w}) \leq 2$, thus

$$\begin{aligned}
l(\Psi) &\geq \log\{\sum_{j=1}^{P}\alpha_j(2\pi)^{-1}\Delta^{-1/2}\exp[-\frac{1}{2}(1 + Y_t'Y_t + X_{t-1}'X_{t-1})\rho]\} \\
&= G_1 - \frac{1}{2}\rho(1 + Y_t'Y_t + X_{t-1}'X_{t-1}),
\end{aligned}$$

for some finite G_1. Overall, $G_1 - \frac{1}{2}\rho(1 + Y_t'Y_t + X_{t-1}'X_{t-1}) \leq l(\Psi) \leq \log(2\pi^{-1}\delta^{-1/2}) +$

$\frac{\kappa}{\gamma}X'_{t-1}X_{t-1}$, from which $E[\underset{\Psi\in\Xi}{sup} \mid l(\Psi) \mid] < \infty$ holds because $X'_{t-1}X_{t-1} = 1 + Y'_{t-1}Y_{t-1} + ... + Y'_{t-Q}Y_{t-Q}$, and $E(Y'_tY_t) < \infty$ for all t by Assumption 3.

Now, we verify (b). Let $s(Y_{t-Q}^{t-1}, \Psi_0)$ be the stationary distribution of Y_{t-Q}^{t-1}, then

$$
\begin{aligned}
&E[L(\Psi)] - E[L(\Psi_0)] \\
&= \iint s(Y_{t-Q}^{t-1}, \Psi_0)[\sum_{j=1}^{P} \alpha_{j,0} f_j(Y_t|Y_{t-Q}^{t-1}, A_{j,0}, \Sigma_{j,0})] \log \frac{\sum_{j=1}^{P} \alpha_j f_j(Y_t|Y_{t-Q}^{t-1}, A_j, \Sigma_j)}{\sum_{j=1}^{P} \alpha_{j,0} f_j(Y_t|Y_{t-Q}^{t-1}, A_{j,0}, \Sigma_{j,0})} dY_t dY_{t-Q}^{t-1} \\
&= \int s(Y_{t-Q}^{t-1}, \Psi_0)\{\int [\sum_{j=1}^{P} \alpha_{j,0} f_j(Y_t|Y_{t-Q}^{t-1}, A_{j,0}, \Sigma_{j,0})] \log \frac{\sum_{j=1}^{P} \alpha_j f_j(Y_t|Y_{t-Q}^{t-1}, A_j, \Sigma_j)}{\sum_{j=1}^{P} \alpha_{j,0} f_j(Y_t|Y_{t-Q}^{t-1}, A_{j,0}, \Sigma_{j,0})} dY_t\} dY_{t-Q}^{t-1},
\end{aligned}
$$

where the inner integral is the negative Kullback-Leibler divergence between two mixture densities: $\sum_{j=1}^{P} \alpha_j f_j(Y_t|Y_{t-Q}^{t-1}, A_j, \Sigma_j)$ and $\sum_{j=1}^{P} \alpha_{j,0} f_j(Y_t|Y_{t-Q}^{t-1}, A_{j,0}, \Sigma_{j,0})$. Therefore, $E[L(\Psi)] - E[L(\Psi_0)] \leq 0$ and the equality holds if and only if

$$
\sum_{j=1}^{P} \alpha_j f_j(Y_t|Y_{t-Q}^{t-1}, A_j, \Sigma_j) = \sum_{j=1}^{P} \alpha_{j,0} f_j(Y_t|Y_{t-Q}^{t-1}, A_{j,0}, \Sigma_{j,0}).
$$

By the identification result from Lemma 1 and the parameter restrictions in equation 6.9, we have that $\alpha_j = \alpha_{j,0}$, $\Sigma_j = \Sigma_{j,0}$ and $A_jX_{t-1} = A_{j,0}X_{t-1}$ for all j, where $A_jX_{t-1} = A_{j,0}X_{t-1}$ implies either that $A_j = A_{j,0}$ or that X_{t-1} takes values only on a $2(Q-1)$ dimensional hyperplane. The latter is impossible as $\{X_{t-1}\}$ takes values on $H \subset \mathbb{R}^{2Q}$, where H has positive Lebesque measure. Therefore, $\alpha_j = \alpha_{j,0}$, $\Sigma_j = \Sigma_{j,0}$ and $A_j = A_{j,0}$ for all j.

Notes

1. Though the model proposed by GL produces conditional heteroskedasticity as a byproduct, its main focus is the modeling of the conditional mean of ITS.

2. The pseudo location parameter of a truncated bivariate normal distribution can be interpreted as the location parameter of the bivariate normal distribution without truncation. It is called pseudo because it no longer represents the mean (location) of the truncated distribution after the truncation is imposed.

3. In Lin and González-Rivera (2019), the extremes (min/max) are modeled following distributional results provided by the Extreme Value Theory. The authors show that the conditional mean of the extremes are non-linear functions of the moments of the underlying process and propose a non-parametric modeling strategy. In the same vein, here we are proposing a very flexible approach with a semi-parametric bent to approximate the true density of the extremes, which asymptotically falls within the family of bivariate Generalized Extreme Value (GEV) distributions. The flexibility comes from the ability of the data guiding the number of components and the weight of each component in the mixture and the dynamic truncation in each component.

4. The idea of data augmentation has also been explored extensively in Bayesian inference. For example, Albert and Chib (1993) introduced latent variables in the Probit model to facilitate the posterior sampling. Chib (1992) applied data augmentation techniques for Bayesian Tobit censored regression models. More recently, Polson *et al.* (2013) constructed a new data augmentation algorithm for Bayesian Logit model. It is worth noting that the EM algorithm was introduced earlier (1977) than the above Bayesian literature. Tanner and Wong (2010) attributed the widespread application of MCMC methods in Bayesian computation to the data augmentation idea in EM algorithm together with the Markov Chain simulation from the statistical physics literature.

5. Note that the data augmenting processes are different from the data generating process, which is specified by the model. See section 3 for details.

6. The analysis in this paper can be modified to accommodate the case where Q is allowed to be component specific.

7. The consistency of the ML estimator is discussed in Appendix A.6.

8. As pointed out in the introduction, the data augmentation process differs from the data generating process. The later is assumed to generate the data Y_t while the former is constructed to facilitate the ML estimation of the parameters of the model.

9. In our simulations, we fix B and play with the values of C to allow the restriction to be binding or not.

10. The objective is to recover the bivariate normal mixture distribution prior to the truncation, from which we can draw random samples by using the existing Matlab packages.

11. From the observations that satisfy the constraint, we start collecting from the 101^{th} observation on (the initial 100 observations are discarded, known as the burn-in period) until the completion of the desired sample size.

12. Elements of α are uniformly selected from $(0, 1)$ and sum up to one. Elements of B are uniformly selected from $(-1, 1)$. Elements of C and off-diagonal elements of L are uniformly selected from $(-3, 3)$, where L is the Cholesky decomposition lower triangle matrix of $\Sigma = LL'$. Diagonal elements of L are uniformly selected from $(0, 3)$. For DGP 3, we choose 200 initial points to account for a higher dimensional parameter space.

13. Differently from Section 5.1, here we use the parameters' true values as the initial values for EM algorithm instead of adopting the random initial value approach discussed in that section.

14. Because of space considerations, we present results for a few parameters of the model. Results for the rest of the parameters are similar and are available upon request.

15. When only one component is involved ($TMT(1, Q)$) turns out to be the same as that of GL, which will be discussed separately in the following discussion.

16. Standard errors are calculated using the block bootstrap (Politis and White, 2004)

17. Within the in-sample period, the best model selected by BIC is the $TMT(4, 2)$.

18. The PITs from the forecast densities of $r_{high,t}$ and $r_{low,t}$ are needed to evaluate their marginal densities and those of $r_{low,t}$ and $r_{high,t}|r_{low,t}$ for the evaluation of the joint densities.

19. The two horizontal lines represent the 95% confidence interval.

20. Let p_t be the PIT of the corresponding density forecast of $r_{low,t}$. Panels (a) to (d) show their sample autocorrelations of $(p_t - \bar{p})$, $(p_t - \bar{p})^2$,$(p_t - \bar{p})^3$,and $(p_t - \bar{p})^4$ respectively, where $\bar{p}$ is the sample mean of p_t. ACF plots for $r_{high,t}$ and $r_{high,t}|r_{low,t}$ for the two models considered provide similar information. These results are not reported but are available upon request.

21. Their results are built for the general mixtures of exponential families, and can be applied here for the finite mixtures.

Tables

Table 1

DGP		α	C	B		Σ	
1	NB	0.6	-2	0.7	-0.1	0.4	0.3
			-2	-0.1	0.7	0.3	0.4
	B	0.4	2	0.1	-0.8	0.4	0.3
			0	-0.8	0.1	0.3	0.4
2	B	0.6	2	0.2	-0.1	0.4	0.3
			2	-0.1	0.2	0.3	0.4
	B	0.4	0	0.1	-0.8	0.4	0.3
			0	-0.8	0.1	0.3	0.4
3	B	0.5	2	0.1	-0.8	0.4	0.3
			2	-0.8	0.1	0.3	0.4
	NB	0.3	2	0.3	-0.4	0.4	0.3
			0	-0.4	0.3	0.3	0.4
	B	0.2	-2	0.2	-0.1	0.4	0.3
			-2	-0.1	0.2	0.3	0.4

Table 1: Data Generating Processes (DGP 1 - DGP 3). B and NB denote binding and not binding interval constraint, respectively.

Table 2

DGP 1	α	C	B		Σ	
True	0.6	-2	0.7	-0.1	0.4	0.3
		-2	-0.1	0.7	0.3	0.4
	0.4	2	0.1	-0.8	0.4	0.3
		0	-0.8	0.1	0.3	0.4
EM (T=200)	0.6036	-1.9644	0.6939	-0.1061	0.3957	0.2997
	(0.0319)	(0.4446)	(0.0644)	(0.0766)	(0.0560)	(0.0476)
		-2.0041	-0.1023	0.6891	0.2997	0.4015
		(0.3235)	(0.0730)	(0.0595)	(0.0476)	(0.0661)
	0.3964	1.9385	0.1054	-0.7978	0.4177	0.3006
	(0.0319)	(0.7890)	(0.0801)	(0.0383)	(0.2974)	(0.0986)
		0.0510	-0.7941	0.1185	0.3006	0.4096
		(0.4026)	(0.0632)	(0.1738)	(0.0986)	(0.1867)
EM (T=1000)	0.6011	-2.0037	0.6995	-0.1023	0.4011	0.3006
	(0.0152)	(0.0625)	(0.0099)	(0.0141)	(0.0234)	(0.0212)
		-2.0038	-0.1012	0.6985	0.3006	0.3987
		(0.0615)	(0.0102)	(0.0144)	(0.0212)	(0.0261)
	0.3989	2.0073	0.0983	-0.8009	0.3937	0.2931
	(0.0152)	(0.0734)	(0.0127)	(0.0163)	(0.0253)	(0.0230)
		0.0038	-0.8016	0.0989	0.2931	0.3916
		(0.0785)	(0.0133)	(0.0170)	(0.0230)	(0.0280)

Table 2: Simulation results for DGP 1. Standard errors in parenthesis.

Table 3

DGP 2	α	C	B		Σ	
True	0.6	2	0.2	−0.1	0.4	0.3
		2	−0.1	0.2	0.3	0.4
	0.4	0	0.1	−0.8	0.4	0.3
		0	−0.8	0.1	0.3	0.4
EM (T=200)	0.6012	1.9462	0.2349	−0.1332	0.4023	0.2879
	(0.0415)	(0.2226)	(0.1950)	(0.1854)	(0.0723)	(0.0568)
		2.0131	−0.0790	0.1753	0.2879	0.3944
		(0.2178)	(0.2044)	(0.1960)	(0.0568)	(0.0728)
	0.3988	−0.0130	0.1034	−0.8003	0.3748	0.2805
	(0.0415)	(0.2122)	(0.2689)	(0.2610)	(0.0747)	(0.0718)
		0.0219	−0.7720	0.0607	0.2805	0.3878
		(0.2326)	(0.2643)	(0.2704)	(0.0718)	(0.0945)
EM (T=1000)	0.5990	1.9644	0.2187	−0.1178	0.4085	0.3002
	(0.0177)	(0.1349)	(0.0863)	(0.0838)	(0.0353)	(0.0306)
		2.0605	−0.1280	0.2271	0.3002	0.4203
		(0.1935)	(0.1189)	(0.1173)	(0.0306)	(0.0523)
	0.4010	−0.0088	0.0967	−0.7971	0.3978	0.2940
	(0.0177)	(0.1269)	(0.1268)	(0.1127)	(0.0386)	(0.0322)
		0.0208	−0.8237	0.1233	0.2940	0.3983
		(0.1076)	(0.1111)	(0.1003)	(0.0322)	(0.0462)

Table 3: Simulation results for DGP 2. Standard errors in parenthesis.

Table 4

DGP 3	α	C	B		Σ	
True	0.5	2	0.1	−0.8	0.4	0.3
		2	−0.8	0.1	0.3	0.4
	0.3	2	0.3	−0.4	0.4	0.3
		0	−0.4	0.3	0.3	0.4
	0.2	−2	0.2	−0.1	0.4	0.3
		−2	−0.1	0.2	0.3	0.4
EM (T=200)	0.5078	1.9985	0.0978	−0.8027	0.3910	0.2908
	(0.0427)	(0.1535)	(0.0551)	(0.0687)	(0.0665)	(0.0560)
		1.9867	−0.7980	0.0972	0.2908	0.3911
		(0.1433)	(0.0515)	(0.0642)	(0.0560)	(0.0594)
	0.2932	2.0114	0.2991	−0.3892	0.3665	0.2736
	(0.0378)	(0.1505)	(0.0640)	(0.0771)	(0.0887)	(0.0746)
		0.0055	−0.4047	0.3111	0.2736	0.3664
		(0.1390)	(0.0610)	(0.0753)	(0.0746)	(0.0792)
	0.1990	−2.0138	0.2002	−0.1047	0.3873	0.2917
	(0.0294)	(0.2691)	(0.0892)	(0.1025)	(0.0944)	(0.0820)
		−1.8540	−0.1382	0.2350	0.2917	0.4110
		(0.4928)	(0.1145)	(0.1285)	(0.0820)	(0.1459)
EM (T=1000)	0.5000	2.0026	0.0990	−0.8016	0.3966	0.2995
	(0.0178)	(0.0583)	(0.0223)	(0.0247)	(0.0272)	(0.0228)
		1.9920	−0.7969	0.0953	0.2995	0.3991
		(0.0574)	(0.0220)	(0.0249)	(0.0228)	(0.0253)
	0.3001	1.9968	0.3008	−0.3987	0.3907	0.2963
	(0.0167)	(0.0710)	(0.0273)	(0.0304)	(0.0334)	(0.0274)
		−0.0054	−0.3983	0.2990	0.2963	0.3986
		(0.0707)	(0.0263)	(0.0308)	(0.0274)	(0.0334)
	0.1999	−1.9983	0.2003	−0.0987	0.3886	0.2914
	(0.0115)	(0.0954)	(0.0298)	(0.0437)	(0.0483)	(0.0382)
		−2.0139	−0.0960	0.1996	0.2914	0.3906
		(0.1007)	(0.0310)	(0.0427)	(0.0382)	(0.0493)

Table 4: Simulation results for DGP 3. Standard errors in parenthesis.

Table 5

Component	α	C	B_1		B_2		Σ	
1	0.4184	0.3916	0.0681	−0.1033	−0.0276	0.0327	0.1838	0.1600
	(0.0428)	(0.0535)	(0.0331)	(0.0412)	(0.0368)	(0.0370)	(0.0230)	(0.0195)
		−0.2864	−0.0683	0.0801	−0.1285	0.1480	0.1600	0.1909
		(0.0688)	(0.0411)	(0.0501)	(0.0402)	(0.0397)	(0.0195)	(0.0204)
2	0.3635	0.3678	0.1758	−0.1563	0.0152	−0.0857	0.5367	0.5165
	(0.0450)	(0.0859)	(0.0781)	(0.0829)	(0.0442)	(0.0587)	(0.0883)	(0.0832)
		−0.4786	−0.0843	0.1641	−0.2135	0.1674	0.5165	0.7006
		(0.0886)	(0.0819)	(0.1001)	(0.0531)	(0.0856)	(0.0832)	(0.0840)
3	0.1323	0.4125	0.6549	−0.5425	0.1157	−0.2460	0.3476	0.1228
	(0.0508)	(0.1946)	(0.1715)	(0.1354)	(0.1214)	(0.0968)	(0.0819)	(0.0606)
		−0.1677	−0.1510	0.1473	0.1101	−0.2316	0.1228	0.1778
		(0.1054)	(0.0973)	(0.0821)	(0.0632)	(0.0693)	(0.0606)	(0.0617)
4	0.0857	0.1484	0.1015	−0.1265	0.5265	−0.3146	5.9263	5.4043
	(0.0189)	(0.3580)	(0.1948)	(0.1856)	(0.2271)	(0.1736)	(0.8068)	(0.7199)
		−0.9836	−0.1358	0.3614	−0.0414	0.2858	5.4043	6.2028
		(0.4077)	(0.1980)	(0.1778)	(0.2525)	(0.1805)	(0.7199)	(0.8251)

Table 5: Estimation results of the $TMT(4, 2)$. Standard errors in parentheses.

Table 6

Model	Log-likelihood	Number of parameters	BIC
$VAR(7)$	-8604	30	17,454
$VAR(7) - DCC - N$	-8155	39	15,991
$VAR(7) - DCC - t$	-7367	40	15,061
$GL(7)$	-8486	33	17,243
$RTMT(5)$	-6975	49	14,352
$TMT(4,2)$	-6833	55	14,117

Table 6: In-Sample Evaluation of Competing Models

Table 7

		G-ACR t-statistics				
		lag				
		1	2	3	4	5
	0.01	-0.47	0.10	-0.18	0.39	-1.03
	0.05	1.40	-0.16	2.02	2.38	0.46
	0.1	1.45	0.03	1.47	1.56	0.73
	0.2	0.88	0.53	1.51	1.22	1.41
	0.3	2.30	2.21	2.74	2.50	2.67
	0.4	2.68	2.14	2.44	2.23	2.62
alpha	0.5	2.47	2.45	2.79	2.27	2.61
	0.6	2.73	2.71	3.38	2.91	2.53
	0.7	1.99	1.83	2.64	2.14	1.83
	0.8	1.01	0.62	0.99	0.70	0.53
	0.9	1.50	1.42	1.93	1.56	1.40
	0.95	1.53	1.53	1.62	1.52	1.51
	0.99	1.45	1.45	1.45	1.45	1.45
C-statistic		17.69	17.39	24.56	22.30	17.47

Table 7: G-ACR Tests for $TMT(4,2)$ Model. The 1% and 5% critical values for the t-statistic are 2.58 and 1.96 respectively. The 1% and 5% critical values for the C-statistic are 27.69 and 22.36. Critical values are based on the asymptotic distributions of the corresponding statistic.

Table 8

		G-ACR t-statistics				
		lag				
		1	2	3	4	5
	0.01	-2.17	-1.60	-2.17	-1.60	-3.01
	0.05	-0.76	-1.00	-0.15	-0.14	0.82
	0.1	3.38	2.63	2.90	3.41	4.51
	0.2	6.90	6.49	7.65	7.18	8.16
	0.3	10.07	10.29	10.87	10.18	10.71
	0.4	10.94	11.20	11.50	11.10	11.22
alpha	0.5	11.39	11.82	11.85	11.74	11.68
	0.6	9.63	9.61	9.69	9.68	9.57
	0.7	8.70	8.64	8.83	8.67	8.52
	0.8	7.71	7.59	7.64	7.68	7.57
	0.9	4.99	5.06	5.05	4.98	4.97
	0.95	3.72	3.72	3.81	3.71	3.71
	0.99	1.45	1.45	1.45	1.45	1.45
C-statistic		177.59	186.69	190.22	175.75	185.23

Table 8: G-ACR Tests for VAR(7)-DCC-t Model. The 1% and 5% critical values for the t-statistic are 2.58 and 1.96. The 1% and 5% critical values for C-statistic are 27.69 and 22.36. Critical values are based on the asymptotic distributions of the corresponding statistic.

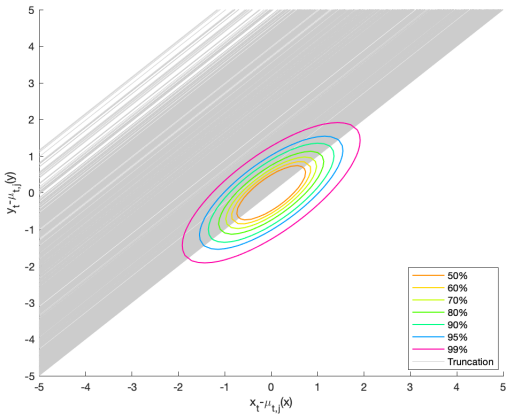
Table 9

m	4	5
TMT	17.23%	-7.50%
VAR-DCC-t	5.23%	-12.56%

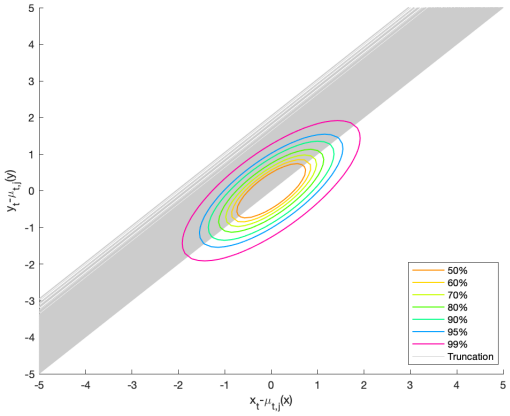
Table 9: Trading Strategy Comparison for IBM average annualized returns over the out-of-sample period from January 1, 2014 to April 1, 2018.

Figures

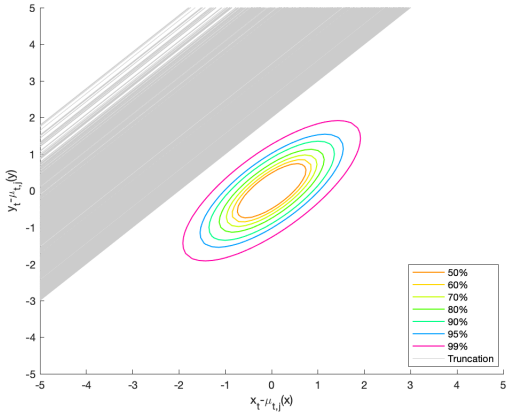
Figure 1



(a) Binding with high persistence



(b) Binding with low persistence



(c) Non-binding

Figure 1: Truncated areas in the conditional densities of DGP 3.

Figure 2

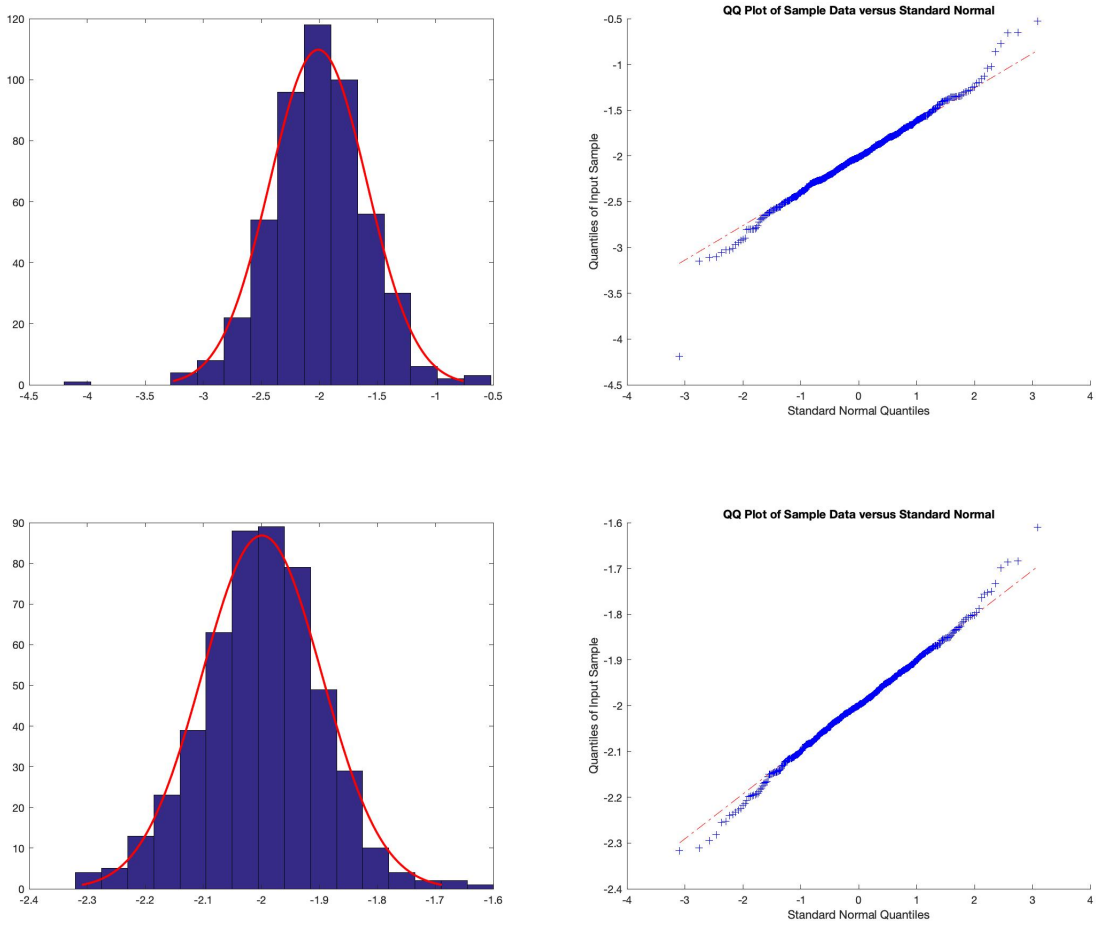


Figure 2: Histogram and QQ plot of the first element of C_1 (true value is -2). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 3

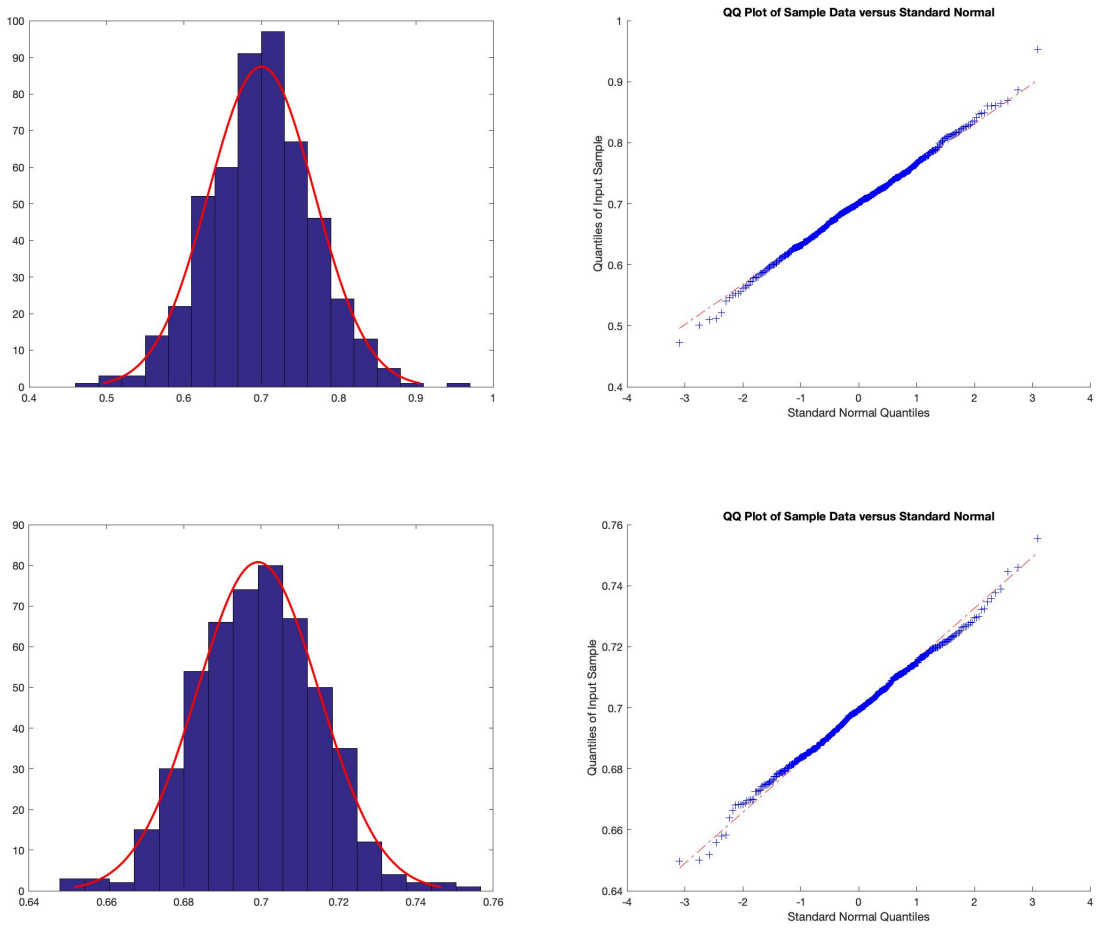


Figure 3: Histogram and QQ plot of the first element of $B_{1,1}$ (true value is 0.7). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 4

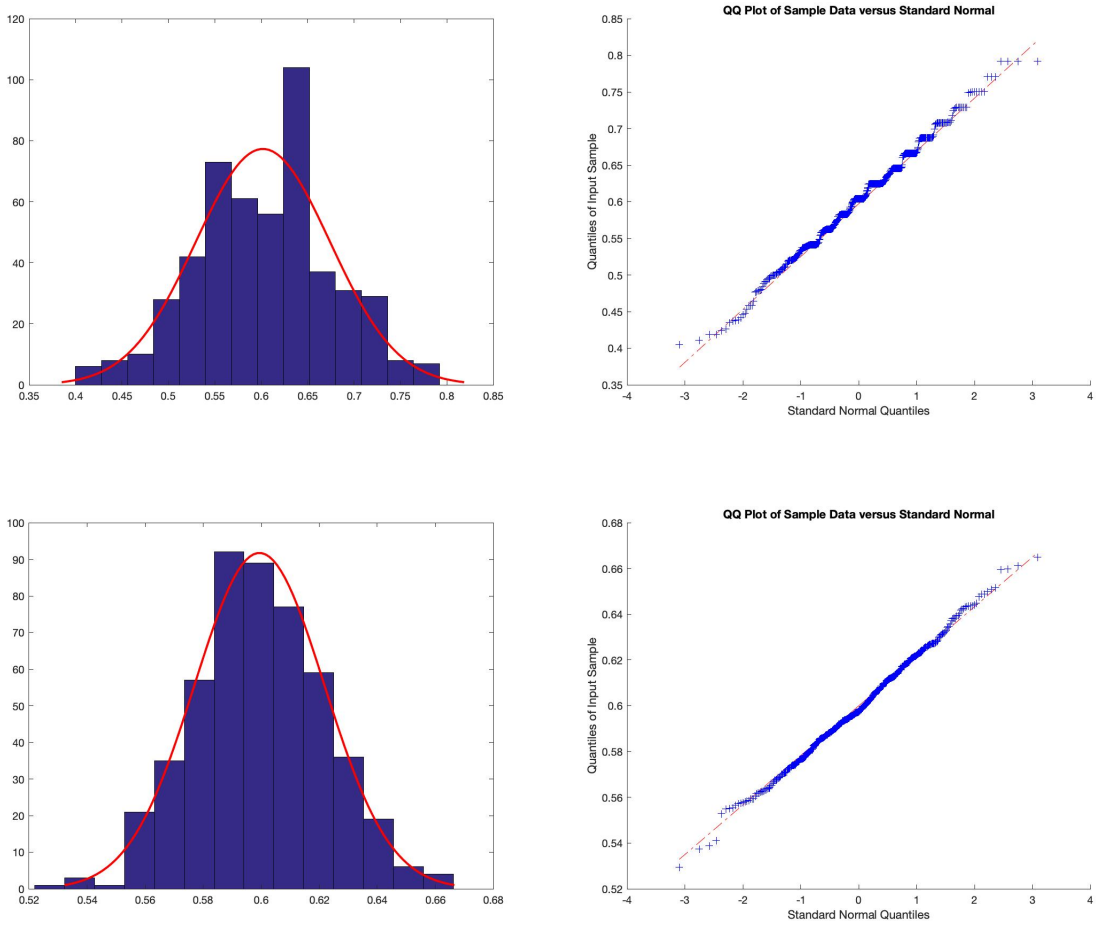


Figure 4: Histogram and QQ plot of α_1 (true value is 0.6). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 5

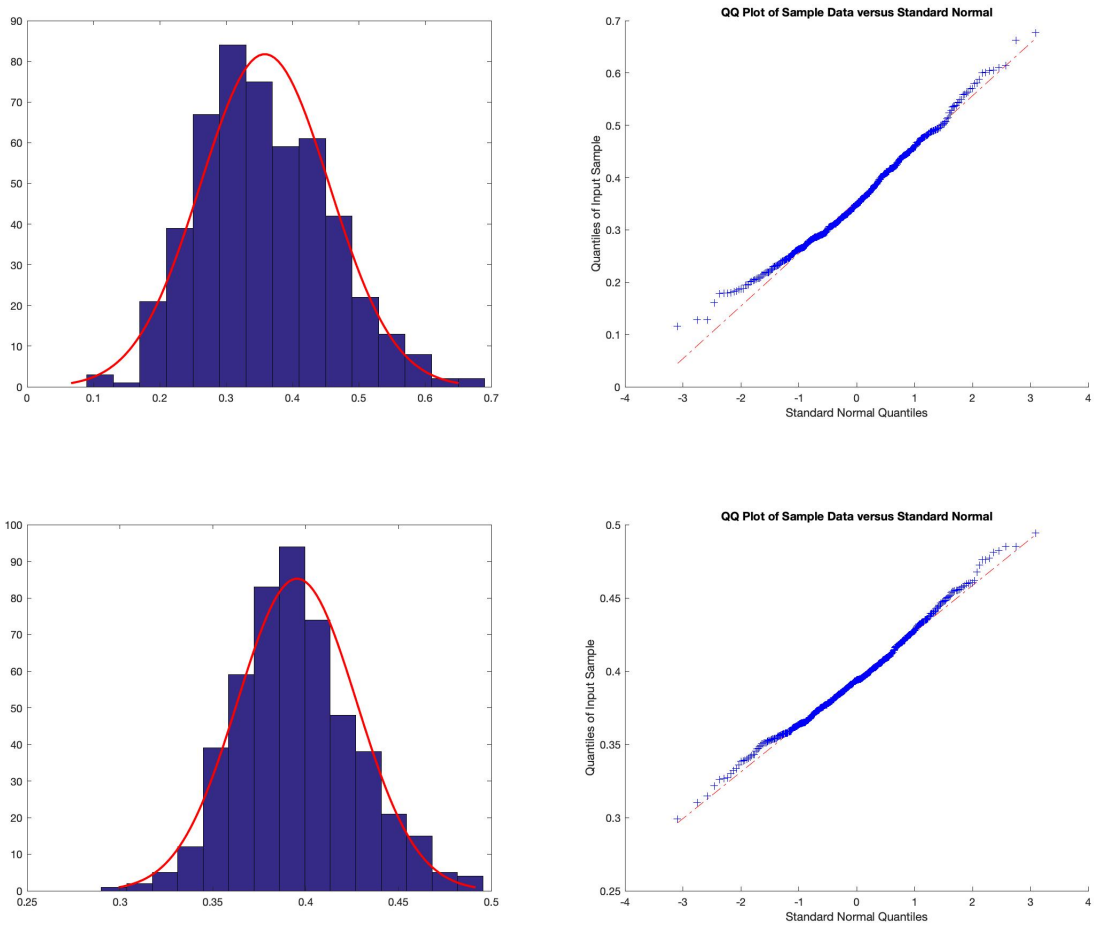


Figure 5: Histogram and QQ plot of the first element of Σ_1 (true value is 0.4). $T = 50$ in top panel and $T = 500$ in bottom panel. The solid curves in the histograms are normal densities.

Figure 6

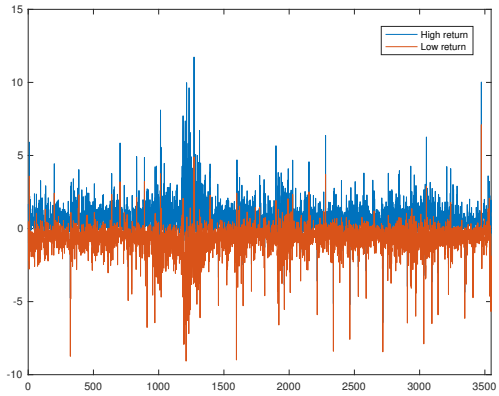
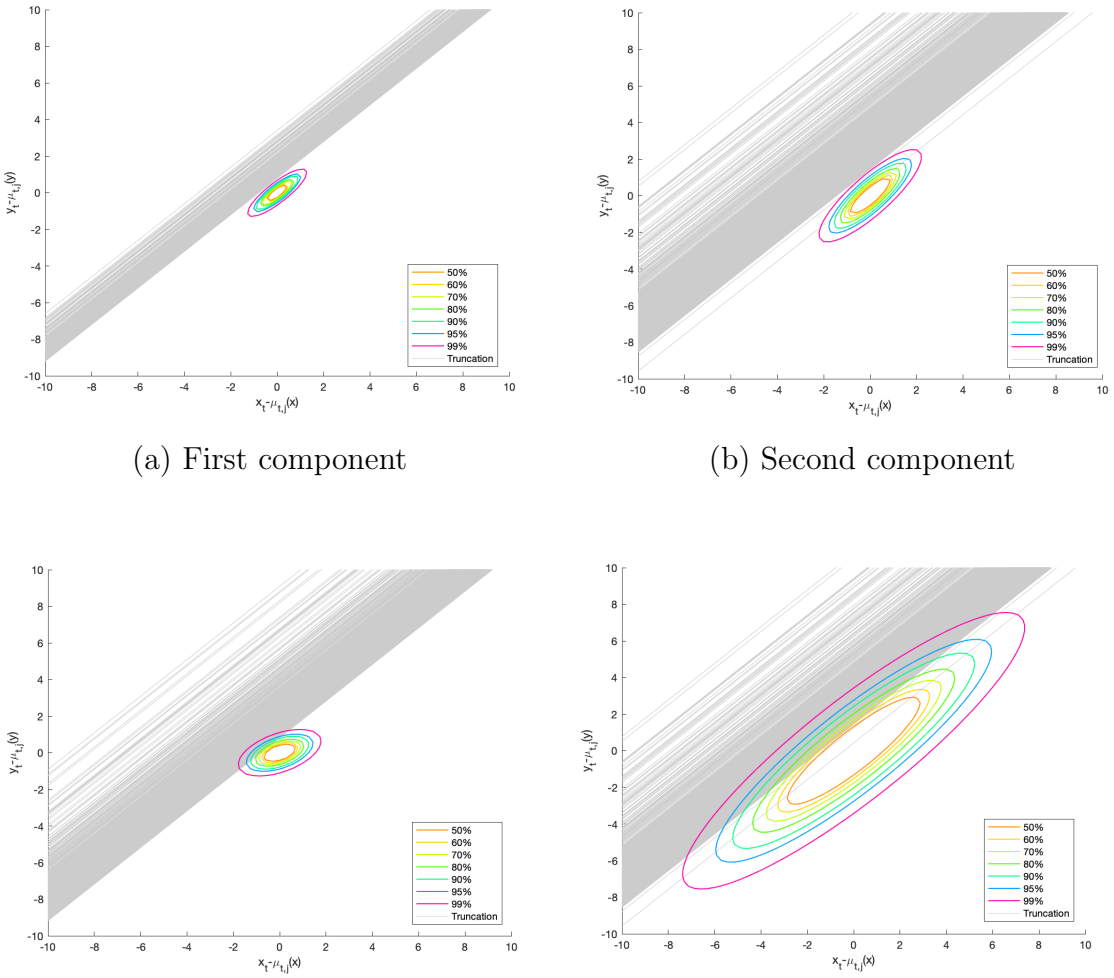


Figure 6: Daily IBM high/low stock returns (2004/1/1 to 2018/4/1).

Figure 7



(a) First component

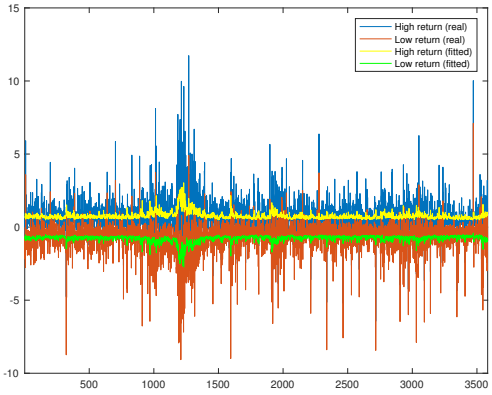
(b) Second component

(c) Third component

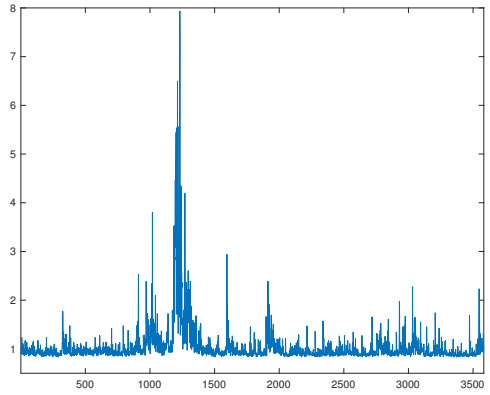
(d) Fourth component

Figure 7: Truncations in the bivariate density of each component of the model $TMT(4, 2)$.

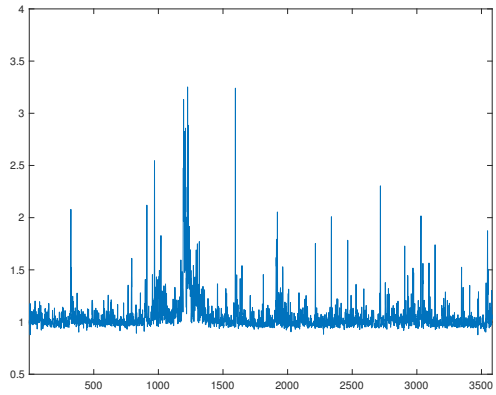
Figure 8



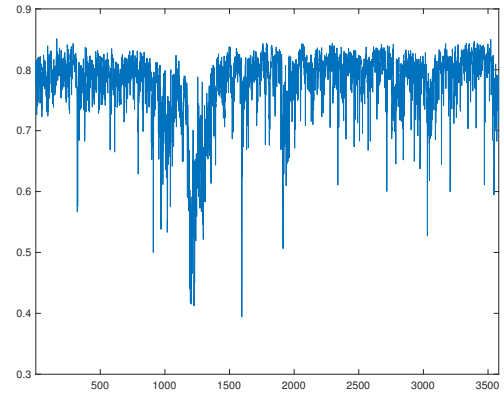
(a) Estimated Conditional Mean



(b) Estimated Variance (High Returns)



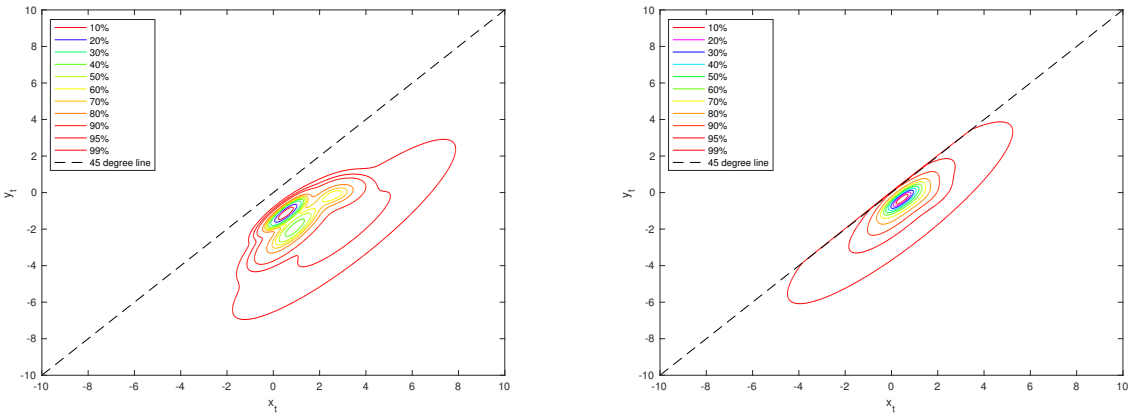
(c) Estimated Variance (Low Returns)



(d) Estimated Correlation

Figure 8: Estimated conditional mean, variance and correlation of daily IBM high/low stock returns (2004/1/1 to 2018/4/1).

Figure 9



(a) 12/18/2008 (b) 12/29/2017

Figure 9: Estimated conditional bivariate density contours.

Figure 10

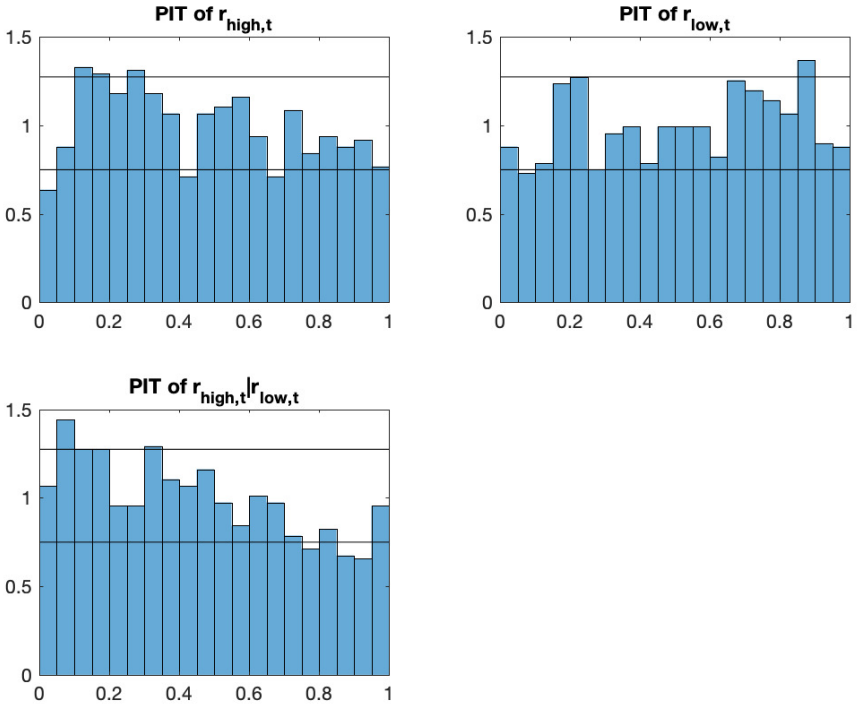


Figure 10: PITs from $TMT(4, 2)$ density forecasts.

Figure 11

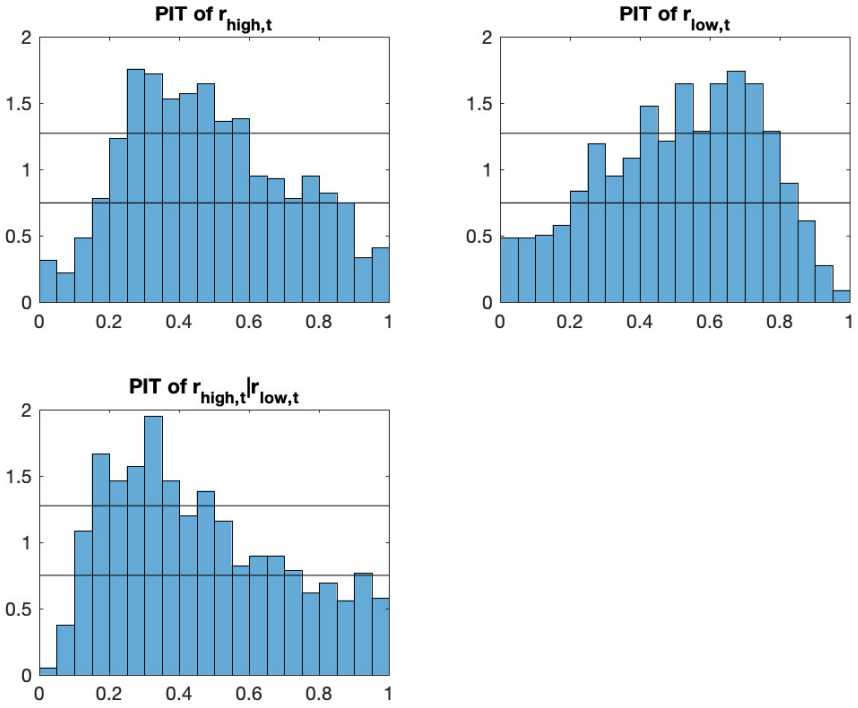


Figure 11: PITs from VAR(7)-DCC-t density forecasts.

Figure 12

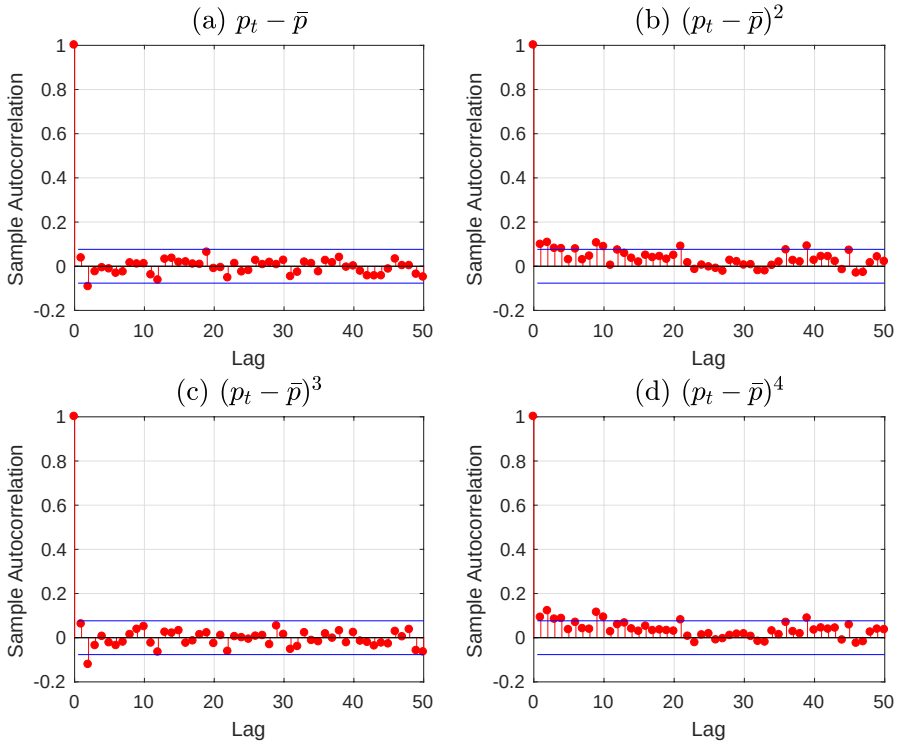


Figure 12: ACF of functions of PITs extracted from the $r_{low,t}$ densities generated by $TMT(4, 2)$ model. p_t is the PIT and $\bar{p}$ is the sample mean of p_t.

Figure 13

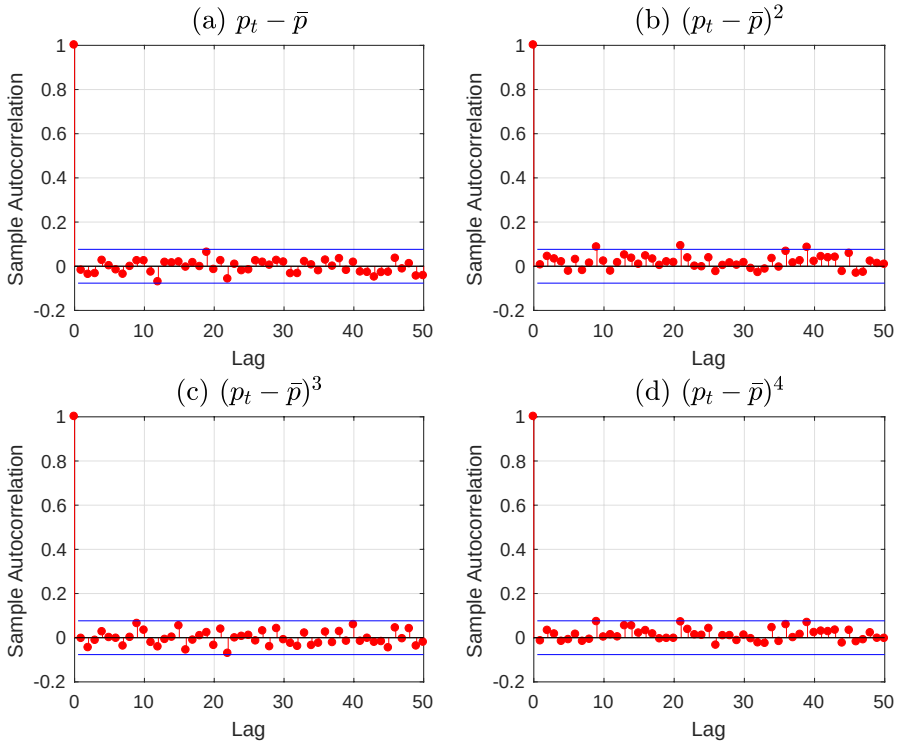


Figure 13: ACF of functions of PITs extracted from the $r_{low,t}$ densities generated by VAR(7)-DCC-t model. p_t is the PIT and $\bar{p}$ is the sample mean of p_t.

Figure 14

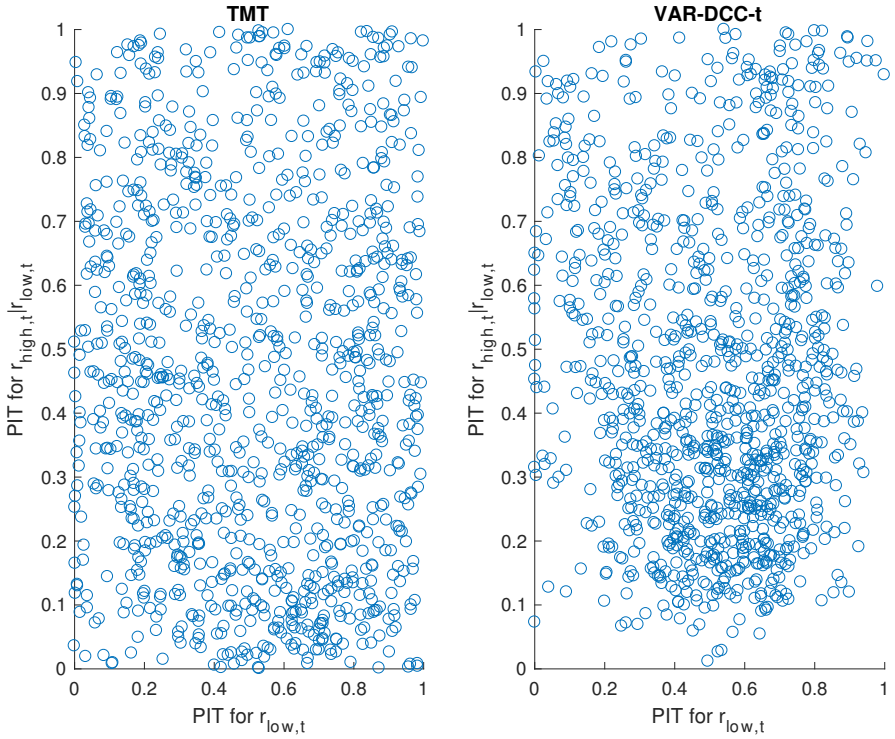


Figure 14: G-ACR plots for TMT(4,2) and VAR(7)-DCC-t models.

Figure 15

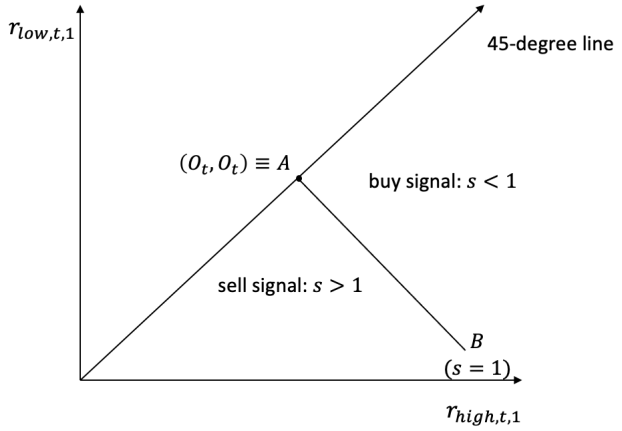
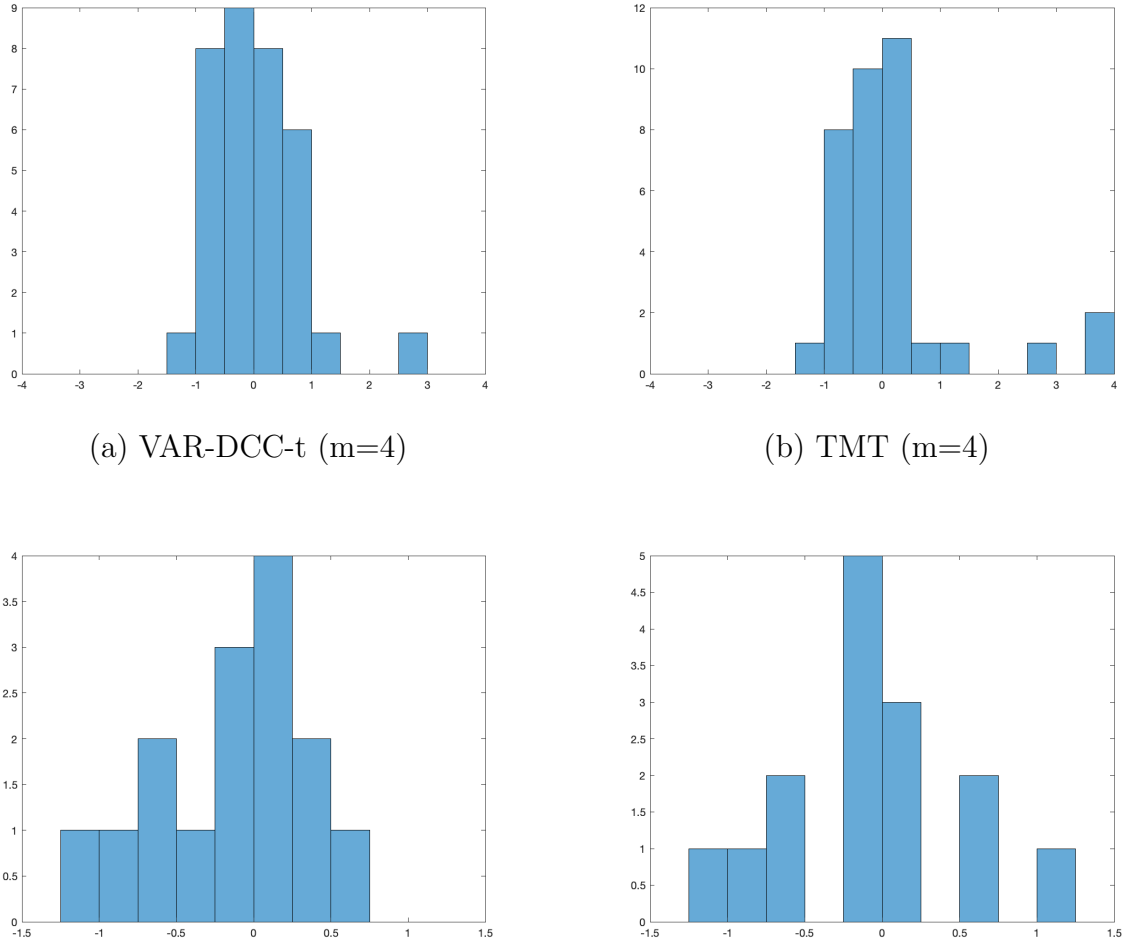


Figure 15: Buy and sell signals from trading strategy.

Figure 16



(a) VAR-DCC-t (m=4)

(b) TMT (m=4)

(c) VAR-DCC-t (m=5)

(d) TMT (m=5)

Figure 16: Histograms of the annualized trading returns over the out-of sample period from January 1, 2014 to April 1, 2018.

Figure 17

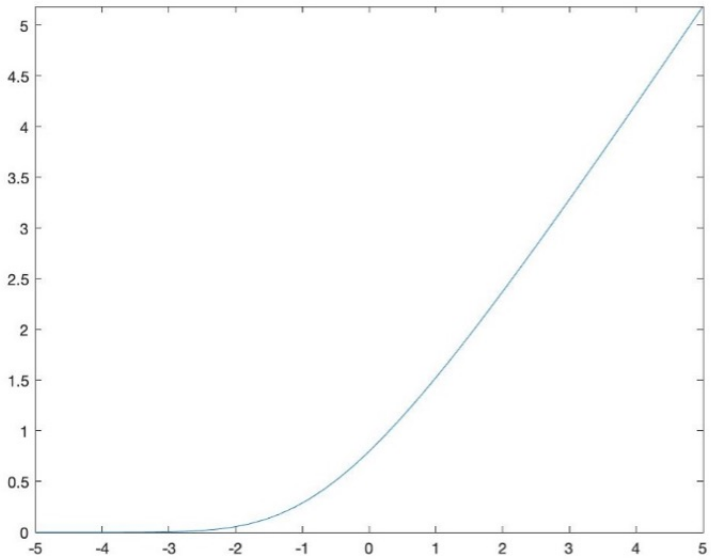


Figure 17: Inverse Mill Ratio.