A Truncated Mixture Transition Model for Interval-valued Time Series

Yun Luo

Gloria González-Rivera

Department of Economics

University of California, Riverside

Abstract

We propose a model for interval-valued time series (ITS), e.g. the collection of daily intervals of high/low stock returns over time, that specifies the conditional joint distribution of the upper and lower bounds of the interval as a mixture of truncated bivariate normal distribution. This specification guarantees that the natural order of the interval (upper bound not smaller than lower bound) is preserved. The model also captures the potential conditional heteroscedasticity and non-Gaussian features in ITS. The standard EM algorithm, when applied to the estimation of mixture models with truncated distribution, does not provide a closed-form solution in M step. We propose a new EM algorithm that solves this problem. We establish the consistency of the maximum likelihood estimator. Monte Carlo simulations show the new EM algorithm has good convergence properties. We apply the model to the interval-valued IBM daily stock returns and it exhibits superior performance over competing methods.

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. Financial data, e.g., tick-by-tick stock transaction data recorded at the ultra-high frequency, generally collapses to a lower frequency single point, e.g. the daily closing price. Aggregating the data into intervals, e.g. daily max/min price interval, is more useful because it provides information on both the price level and the daily price volatility. A similar example is the interval of daily low/high temperature that provides relevant information for decision making. 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.

The literature on modeling interval data and ITS can be divided into two categories depending on the data representation: the center/range system (e.g. center and range are respectively the midpoint and the distance between the upper and lower bounds.) or the upper/lower bound system. In the center/range system, the interval constraint is that the range cannot be smaller than zero. Lima Neto and de Carvalho (2010) propose modeling center and range separately while imposing non-negative constraints on the parameters of the range equation, which are unnecessarily too restrictive and complicate the estimation of the system. Tu and Wang (2016) overcome this restriction by log-transforming the range. However, it requires bias correction for the conditional mean and can fail when zero is present in range data. In the upper/lower bound system, an equivalent interval constraint is that

the upper bound cannot be smaller than the lower bound. González-Rivera and Lin (2013) propose a constrained regression model (GL) that preserves this 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 assumption is restrictive as the consistency of the estimators heavily depends on it.

The previous literature explores a variety of ways to preserve the interval constraint, and mainly focuses on modeling the conditional mean of ITS. To the best of our knowledge, none of the existing work has considered modeling the potential conditional heteroskedasticity in ITS, a feature that has been widely recognized in point-valued time series (PTS). One exception is that GL may produce conditional heteroskedasticity as a byproduct. In addition, many PTS exhibit non-Gaussian features that may also appear in ITS, such as flat streches, burst of activities, outliers and changepoints (see e.g., Le et.al. 1996, Wong and Li 2000), opening a door for models capable of generating more flexible predictive densities, an issue that has not been addressed in the current ITS literature. By contrast, there is a vast amount of literature on modeling conditional heteroskedasticity and non-Gaussian behaviors for PTS. Particularly, Le et. al. (1996) propose a Mixture Transition Distribution (MTD) model for the univariate PTS that seeks to account for the non-Gaussian features. Their idea is to specify the conditional distribution for the variable of interest as a mixture distribution, where each component contains only one lag from the information set. The fact that MTD is able to handle conditional heteroskedasticity is noted and discussed by Berchtold and Raftery (2002). MTD is generalized by Wong and Li (2000) under the name of Mixture Autoregressive (MAR) model to entertain more flexibility by allowing each component to depend on the full information set. Hassan and Lii (2006) extend MTD for the marked point process under a bivariate setting.

In this paper, we propose a model for ITS in the upper/lower bound system in the spirit of the MTD model and its extensions. We specify the joint distribution of the upper bound (x_t) and lower bound (y_t) conditional on the information set as a mixture of truncated bivariate normal distribution, where for each component the bivariate normal distribution is

truncated at $x_t \geq y_t$. The information set enters the conditional distribution as a linear function through the pseudo location parameter of the truncated bivariate normal distribution for each component.[1] The model comes with several benefits. First, it is able to preserve the natural order of ITS, that is, the upper bound not smaller than the lower bound. Second, the model can capture conditional heteroskedasticity without modeling it explicitly, as the dynamics enter the covariance matrix via the truncation and the mixture framework. Third, the mixture distribution that the model based on provides great flexibility in terms of approximating the underlying true conditional distribution, and hence can improve the quality of density forecast.

It is well known that the maximum likelihood estimator (MLE) does not have a closed-form solution for mixture models resulting from the complexity of the likelihood. In the literature, EM algorithm is a standard device to find the MLE for mixture models due to its simplicity and monotonicity in the likelihood (see e.g. Hamilton 1990, Le et. al. 1996, Hassan and Lii 2006). However, such a standard EM algorithm fails in our model as no closed-form solution can be obtained in the M step. This is caused by the normalizing factor in the truncated normal distribution, which possesses a complex form after taking the derivative. To overcome this problem, we propose a new EM algorithm.[2] The innovation is made by constructing a high level pseudo complete data generating process that brings in more latent variables than the standard EM algorithm. Specifically, at each time the observation is generated in four steps. First, a membership variable (latent) is generated from a multinomial distribution that suggests which component the observation truly comes from. Second, conditional on the observation coming from the component indicated in the previous step, a variable (latent) is obtained from a geometric distribution that indicates the number of invalid observations ($x_t < y_t$) before the occurrence of a valid observation ($x_t \geq y_t$). Third, generate the corresponding number of invalid observations (latent) independently from the area of the bivariate normal distribution where $x_t < y_t$. Fourth, draw one observation from

[1]The pseudo location parameter of a truncated bivariate normal distribution can be interpreted as the location parameter of the bivariate normal distribution (before the truncation). It is called pseudo because it no longer represents the mean (location) of the truncated distribution after the truncation is imposed.

[2]The new EM algorithm can be applied generally to data sets with any kinds of truncation either in the time series setting or for cross-sectional probability clustering.

the area of the bivariate normal distribution where $x_t \geq y_t$, and treat it as the valid observation. The Monte Carlo simulations indicate that the new EM algorithm performs well with the finite sample. We show that the MLE is consistent under some regular assumptions. We apply the model to IBM daily stock return ITS and show that it outperforms the competing models.

The organization of the paper is as follows. In Section 2, we introduce the truncated mixture transition model and discuss its properties. In Section 3, we propose the new EM algorithm. In Section 4, we show the consistency of the MLE. In Section 5, we perform the Monte Carlo simulations. In Section 6, we apply our model to IBM daily stock return ITS. We conclude in Section 7.

2 The Truncated Mixture Transition Model

2.1 Definition

Let x_t be the upper bound, and y_t be the lower bound of the interval observed at time t. The interval time series data has the following format

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

where by construction $x_t \geqslant y_t$, and we denote $Y_t = (x_t, y_t)'$ hereafter. We say that Y_t is generated by a truncated mixture transition ($TMT(P,Q)$) model if its conditional density function given the past information set 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 P is the number of components and is assumed to be fixed, and Q is the number of lags in each component.[3] $\mathcal{F}^{t-1}$ is the information set up to time $t-1$, and $Y_{t-Q}^{t-1} = (Y_{t-Q}, Y_{t-Q+1}, ..., Y_{t-1})$. $f_j(Y_t|Y_{t-Q}^{t-1})$ is a truncated bivariate normal probability density function truncated at $x_t \geqslant y_t$. That is, for each component, the upper bound is not smaller than the lower bound. The truncated density has the following form (see e.g. Nath 1972)

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

where $\mu_{t,j} = C_j + B_{j,1}Y_{t-1} + ... + B_{j,Q}Y_{t-Q}$, C_j (2×1) is a constant vector, $B_{j,r}$ (2×2) $(r = 1, ..., Q)$ is a matrix, Σ_j (2×2) is a positive semi-definite matrix, and $|A|$ is the determinant of matrix A. (2.2) differs from a bivariate normal distribution in the extra 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 $(x_t \geq y_t)$. Φ is the standard normal cumulative distribution function, and $w = (1, -1)'$.

2.2 Theoretical properties

Given the definition above, we can write down the conditional mean of Y_t:

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

where

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

ϕ is the standard normal density function. Unlike the normal density, where $\mu_{t,j}$ is the mean

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

for component j, the additional term, $M^1_{o,t,j}$, represents the mean shift after the truncation (see Nath 1972 for moments of truncated normal distribution). As a result, the conditional mean is no longer $\mu_{t,j}$ but a nonlinear function of $\mathcal{F}^{t-1}$. We also show that the natural order of interval time series is preserved at the conditional mean level: $w'E(Y_t|\mathcal{F}^{t-1})) \geq 0$. The proof can be found in Appendix A.1.

A promising feature of TMT model is that it can produce a time-varying conditional variance to capture conditional heteroskedasticity. To see this, the conditional variance is given by:

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

where

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

3 Estimation

In this section, we discuss the estimation of the TMT model using maximum likelihood (ML). The goal is to estimate the set of parameters $\Psi = \{\alpha_j, A_j, \Sigma_j | \forall j\}$ by maximizing the likelihood:

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

where $A_j = (C_j, B_{j,1}, ..., B_{j,Q})$. We first consider an unconditional version of (3.1), where

$\mu_{t,j} = \mu_j$ doesn't depend on the information set. The corresponding log-likelihood function for $\Theta = \{\alpha_j, \mu_j, \Sigma_j | \forall j\}$ can be written as

$$L(\Theta) = \frac{1}{T}\sum_{t=1}^{T} log[\sum_{j=1}^{P} \alpha_j f_j(Y_t|\mu_j, \Sigma_j)] \tag{3.2}$$

Estimating Θ is easier than Ψ because the conditional distribution of Y_t doesn't depend on the information set and can be viewed as if the data is drawn i.i.d. from the mixture distribution. Therefore, we will first illustrate the ML estimation of (3.2) and then (3.1).

Clearly, no closed-form solution can be obtained from maximizing (3.2). In fact, 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, and it has been widely applied to find the ML estimators for mixture models due to its simplicity and monotonicity property (see Dempster et. al. 1977), e.g., Hamilton (1990) uses EM algorithm to estimate the regime switching model. The statistical properties of EM algorithm have been studied extensively in the literature (see e.g. Wu 1983, Meng 1994, McLachlan and Krishnan 2007, and Balakrishnan, et. al. 2017).

A review of the EM algorithm for normal mixture models in unconditional setting (each $f_j(.)$ in (3.2) represents a normal distribution) can be found in Appendix A.2. Lee and Scott (2010) apply the EM algorithm to a truncated normal mixture model with each component truncated by a rectangle, 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 EM algorithm fails to have a closed-form solution in the M step, mainly due to the truncation term ($\frac{\phi(.)}{1-\Phi(.)}$) in the density (see Appendix A.3 for details). As a result, numerical maximization is needed in M step (see e.g. Lange 1995), sacrificing the simplicity of the EM algorithm. In the following, we propose a new EM algorithm that solves this problem.

3.1 A new EM algorithm for truncated normal mixture model (unconditional case)

The new EM algorithm begins by transforming the data generating process into a missing data framework as follow. To obtain the observation Y_t, a latent variable z_t is generated from a multinomial distribution, indicating which component the observation truly comes from. Next, conditional on z_t, another latent variable n_t can be generated from a geometric distribution. n_t represents the number of invalid draws ($x_t < y_t$) from the respective component (a bivariate normal distribution) before a valid draw ($x_t \geq y_t$) arrives. The valid draw (the $(n_t + 1)^{th}$ draw) is then treated as the t^{th} observation (Y_t). In other words, only the valid draw can be observed while all the invalid draws (if any) are latent. Denote $Y_t^A = \{Y_{t,1}, Y_{t,2}, ..., Y_{t,n_t}, Y_{t,n_t+1}\}$ as all the draws for time t. We now formalize the above data generating process.

Let z_t follow a multinomial distribution:

$$g(z_t|\Theta) = \prod_{j=1}^{P} \alpha_j^{z_{tj}} \tag{3.3}$$

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.

$$q(n_t|z_t, \Theta) = \prod_{j=1}^{P} \Big[(1 - F_j)^{n_t} F_j\Big]^{z_{tj}} \tag{3.4}$$

where $F_j = 1 - \Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}})$ is the cumulative distribution for the truncated area ($x_t \geq y_t$) for component j, and represents the probability of getting a valid draw from the bivariate normal distribution. Then, the conditional density of Y_t^A is specified as below

$$h(Y_t^A|z_t, n_t, \Theta) = \prod_{j=1}^{P}\Big[\frac{f_j^N(Y_{t,n_t+1})}{F_j}\prod_{k=1}^{n_t}(\frac{f_j^N(Y_{t,k})}{1-F_j})\Big]^{z_{tj}} \tag{3.5}$$

where $f_j^N(.)$ is the bivariate normal density of component j. Next, the joint density function of the pseudo complete data ($\{Y_t^A, z_t, n_t\}$) can be constructed,

$$\begin{aligned} l(Y_t^A, z_t, n_t|\Theta) &= g(z_t|\Theta)q(n_t|z_t,\Theta)h(Y_t^A|z_t,n_t,\Theta) \\ &= \prod_{j=1}^{P}\Big[\alpha_j f_j^N(Y_{t,n_t+1})\prod_{k=1}^{n_t} f_j^N(Y_{t,k})\Big]^{z_{tj}} \end{aligned} \tag{3.6}$$

and we can write down the pseudo complete log-likelihood function.

$$L^C(\Theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} z_{tj}[log\alpha_j + log f_j^N(Y_{t,n_t+1}) + \sum_{k=1}^{n_t} log f_j^N(Y_{t,k})] \tag{3.7}$$

E Step: the above likelihood (3.7) is replaced with its conditional expectation. See Appendix A.4 for details.

$$\begin{aligned} &Q(\Theta|\Theta^l) \\ =&E(L^C(\Theta)|Y,\Theta^l) \\ =&\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P}\tilde{z}_{tj}[log\alpha_j + log f_j^N(Y_{t,n_t+1}) + \tilde{n}_{t,j}(\int log f_j^N(Y_{t,k})(\frac{f_j^{N,l}(Y_{t,k})}{1-F_j^l})dY_{t,k})] \end{aligned} \tag{3.8}$$

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

$$\begin{aligned}\tilde{z}_{tj} &= P(z_{tj}=1|Y,\Theta^l) \\ &= \frac{P(z_{tj}=1,Y_t|\Theta^l)}{P(Y_t|\Theta^l)} \\ &= \frac{\alpha_j^l f_j^l(Y_t)}{\sum_{r=1}^P \alpha_r^l f_r^l(Y_t)} \end{aligned} \tag{3.9}$$

M Step: We can obtain a closed-form solution by maximizing $Q(\Theta|\Theta^l)$. See Appendix A.5 for details.

$$\alpha_j^{l+1} = \frac{\sum_{t=1}^T \tilde{z}_{tj}}{N} \tag{3.10}$$

$$\mu_j^{l+1} = \frac{\sum_{t=1}^T \tilde{z}_{tj}(Y_t + \tilde{n}_{t,j}(M_{d,j}^{1,l} + \mu_j^l))}{\sum_{t=1}^T \tilde{z}_{tj}(1+\tilde{n}_{t,j})} \tag{3.11}$$

$$\Sigma_j^{l+1} = \frac{\sum_{t=1}^T \tilde{z}_{tj}[(Y_t - \mu_j^{l+1})(Y_t - \mu_j^{l+1})' + \tilde{n}_{t,j} M_{d',j}^2]}{\sum_{t=1}^T \tilde{z}_{tj}(1+\tilde{n}_{t,j})} \tag{3.12}$$

where $M_{d',j}^2 = M_{d,j}^{2,l} + (\mu_j^l - \mu_j^{l+1})(M_{d,j}^{1,l})' + (M_{d,j}^{1,l})(\mu_j^l - \mu_j^{l+1})' + (\mu_j^l - \mu_j^{l+1})(\mu_j^l - \mu_j^{l+1})'$. $M_{d,j}^{1,l}$ and $M_{d,j}^{2,l}$ are respectively $M_{d,j}^1$ and $M_{d,j}^2$ conditional on Θ^l.

$$M_{d,j}^1 = \frac{-\Sigma_j w}{\sqrt{w'\Sigma_j w}} \frac{\phi(\frac{w'\mu_j}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{w'\mu_j}{\sqrt{w'\Sigma_j w}})} \tag{3.13}$$

$$M_{d,j}^2 = \Sigma_j + \frac{\Sigma_j w w' \Sigma_j}{w'\Sigma_j w} \frac{w'\mu_j}{\sqrt{w'\Sigma_j w}} \frac{\phi(\frac{w'\mu_j}{\sqrt{w'\Sigma_j w}})}{1-\Phi(\frac{w'\mu_j}{\sqrt{w'\Sigma_j w}})} \tag{3.14}$$

It is interesting to notice that (3.11) and (3.12) are the first two moments of the pseudo complete sample weighted by z. For example, the numerator in (3.11) not only includes the observed valid draw (Y_t) but also imputes the sum of the latent invalid draws at time t with its conditional expectation that is feasible at the current iteration: $\tilde{n}_{t,j}(M_{d,j}^{1,l} + \mu_j^l)$. Similar pattern can also be observed in the denominator with $1+\tilde{n}_{t,j}$ being the total number of draws at time t. Moreover, our EM algorithm includes the standard EM algorithm for

normal mixture models (Appendix) as a special case. To see this, suppose no truncation is imposed, we have $F_j = 1$, and $\tilde{n}_{t,j} = 0$. Therefore, the E step and M step become the same as in Appendix.

Finally, repeat E step and M step until convergence. Clearly, the new EM algorithm provides a closed-form solution and is able to maintain the monotonicity property. Furthermore, the constraints on parameters are satisfied by construction, e.g., Σ^{l+1} is positive semi-definite, $\sum_{j=1}^{P} \alpha_j^{l+1} = 1$, and $\alpha_j^{l+1} > 0$.

3.2 A new EM algorithm for truncated normal mixture model (conditional case)

We now discuss the conditional case. The EM algorithm is applied to the likelihood (3.1) to estimate $\Psi = \{\alpha_j, A_j, \Sigma_j | \forall j \in P\}$. Similar to section 3.1, the pseudo complete log-likelihood function can be constructed:

$$L^C(\Psi) = \frac{1}{T-Q} \sum_{t=Q+1}^{T} \sum_{j=1}^{P} z_{tj} [log\alpha_j + log f_{t,j}^N(Y_{t,n_t+1}) + \sum_{k=1}^{n_t} log f_{t,j}^N(Y_{t,k})] \quad (3.15)$$

E Step: the conditional expectation of complete log-likelihood function can be written as

$$\begin{aligned} & Q(\Psi|\Psi^l) \\ =& E[L^C(\Psi)|Y, \Psi^l] \\ =& \frac{1}{T-Q} \sum_{t=Q+1}^{T} \sum_{j=1}^{P} \tilde{z}_{tj} [log\alpha_j + log f_{t,j}^N(Y_{t,n_t+1}) + \tilde{n}_{t,j} (\int log f_{t,j}^N(Y_{t,k}) (\frac{f_{t,j}^{N,l}(Y_{t,k})}{1 - F_{t,j}^l}) dY_{t,k})] \end{aligned} \quad (3.16)$$

where $\tilde{z}_{tj} = \frac{\alpha_j^l f_{t,j}^l(Y_t)}{\sum_{r=1}^{P} \alpha_r^l f_{t,r}^l(Y_t)}$. $\tilde{n}_{t,j} = E(n_t | z_{tj} = 1, Y, \Psi^l) = \frac{1 - F_{t,j}^l}{F_{t,j}^l}$. $f_{t,j}^{N,l}(.)$ and $F_{t,j}^l$ are respectively $f_j^N(.)$ and F_j conditional on Ψ^l and with μ_j replaced by $\mu_{t,j} = C_j + B_{j,1} Y_{t-1} + ... + B_{j,Q} Y_{t-Q}$.

M Step: maximizing $Q(\Psi|\Psi^l)$ gives the iterated rules for Ψ. See Appendix A.6 for details.

$$\alpha_j^{l+1} = \frac{\sum_{t=Q+1}^{T} \tilde{z}_{tj}}{T-P} \tag{3.17}$$

$$A_j^{l+1} = (\bar{X}_j'\bar{Y}_j + \tilde{X}_j'\tilde{M}_{d',\bar{T},j}^1)'(\bar{X}_j'\bar{X}_j + \tilde{X}_j'\tilde{X}_j)^{-1} \tag{3.18}$$

$$\Sigma_j^{l+1} = \frac{\sum_{t=Q+1}^{T} \tilde{z}_{tj}[(Y_t - A_j^{l+1}X_{t-1})(Y_t - A_j^{l+1}X_{t-1})' + \tilde{n}_{t,j}M_{d',t,j}^2]}{\sum_{t=P+1}^{T} \tilde{z}_{tj}(1+\tilde{n}_{t,j})} \tag{3.19}$$

where $\tilde{M}_{d',\bar{T},j}^1 = (\tilde{M}_{d',Q+1,j}^1, ..., \tilde{M}_{d',T,j}^1)'$, and $\tilde{M}_{d',t,j}^1 = \sqrt{\tilde{z}_{tj}\tilde{n}_{t,j}}(M_{d,t,j}^1 + \mu_{t,j}^l)$. $M_{d,t,j}^1$ is $M_{d,j}^1$ with μ_j^l replaced by $\mu_{t,j}^l = C_j^l + B_{j,1}^l Y_{t-1} + ... + B_{j,Q}^l Y_{t-Q}$, and $M_{d',t,j}^2$ is $M_{d',j}^2$ with μ_j^l and μ_j^{l+1} replaced by $\mu_{t,j}^l$ and $\mu_{t,j}^{l+1}$ respectively. Furthermore, $\bar{X}_j = \sqrt{\tilde{z}_j\tau_1^{1+2Q}} \odot X$, 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 $\odot$ represents Hadamard product.

The iterated rules for α_j and Σ_j remain similar to these in Section 3.1 with only minor changes. Note that A_j has an iterated rule that resembles the format of the maximum likelihood estimates for a vector autoregressive model (VAR). When truncation is not in presence, it becomes $A_j^{l+1} = (\bar{X}_j'\bar{Y}_j)'(\bar{X}_j'\bar{X}_j)^{-1}$. Therefore, (3.18) can be viewed as applying VAR to the pseudo complete sample.

4 Asymptotic theory

In this section, we discuss the asymptotic properties of the ML estimator. The following theorem shows that under some regular conditions, the MLE is consistent. We begin by imposing the following assumptions:

Assumption 1. $\{Y_t\}$ are generated from (2.1), and are strictly stationary and ergodic.

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

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

These assumptions are fairly regular in the literature. It may be challenging to verify As-

sumption 1 as the model is nonlinear. The necessary and sufficient conditions for stationarity and ergodicity that are imposed on parameters remain for future research. Notice that for the Gaussian MTD and MAR models, the sufficient and necessary conditions for first-order and second-order stationarity have been derived (see e.g., Le et. al. 1996, Wong and Li 2000). Assumption 2 and Assumption 3 are sufficient to ensure the uniform convergence of the likelihood function.

The following theorem establishes the strong consistency of ML estimator and the proof can be found in Appendix A.7.

Theorem 1. *Under Assumption 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.*

5 Monte Carlo Simulation

In this section, we perform Monte Carlo simulation to evaluate the finite sample performance of the proposed EM algorithm on the TMT model. Experiments are designed for both unconditional and conditional cases.

5.1 Unconditional case experiments

We consider two cases with the number of components being $P = 2$ (DGP 1) and $P = 3$ (DGP 2). The data generating process is as follow. First, we set the parameters according to the configurations in Table 1 and Table 2. Second, we calculate η_j for all j, which represents the corresponding component weight for each component before the truncation is imposed. The relationship between α_j and η_j can be described as: $\alpha_j = \frac{\eta_j F_j}{\sum_{j=1}^{P} \eta_j F_j}$. Third, a large enough sample is drawn from the bivariate normal mixture distribution (component weight η_j). Finally, only the observations that satisfy the constraint $x_t \geq y_t$ are kept.[4]

[4]From these observations that satisfy the constraint, start collecting from the 101^{th} observation (the initial 100 observations are discarded, known as the burn-in period) until the desired sample size is reached.

The initial values of parameters are estimated using K-means [5], from where the EM algorithm iterates until convergence to find the MLE. [6] We consider two sample sizes ($T = 200$ and $T = 1000$). The number of Monte Carlo replications is 100.

In Table 1 and Table 2, we report the means and standard errors of the estimated parameters across replications. The biases of parameters are small in both DGPs. As the sample size increases, the estimates get closer to the true values and the standard errors become smaller. One should bear in mind that in all the DGPs, it is not necessary to impose constraints on μ (e.g., $w'\mu \geq 0$) since μ is not the mean of the truncated normal distribution.

Two components	α	μ	Σ	
True	0.4	8 7	1 0.5	0.5 1
	0.6	4 3	2 0.3	0.3 2
EM (T=200)	0.3995 (0.0374)	7.9666 (0.1653) 7.0037 (0.1961)	1.0137 (0.2103) 0.5303 (0.2237)	0.5303 (0.2237) 1.0735 (0.3297)
	0.6005 (0.0374)	3.9657 (0.3028) 2.9665 (0.2932)	2.0160 (0.4444) 0.3161 (0.3399)	0.3161 (0.3399) 1.9763 (0.4059)
EM (T=1000)	0.3997 (0.0180)	7.9961 (0.0730) 7.0083 (0.0754)	1.0002 (0.1015) 0.4988 (0.0774)	0.4988 (0.0774) 1.0136 (0.1138)
	0.6003 (0.0180)	3.9994 (0.1122) 3.0082 (0.1301)	1.9831 (0.1873) 0.2975 (0.1224)	0.2975 (0.1224) 1.9995 (0.1994)

Note: the numbers in parentheses are standard errors.

Table 1: Simulation results for DGP 1

[5] K-means provides bias estimates because it doesn't account for the truncation. It treats the sample as if it comes from a bivariate normal mixture distribution. Nevertheless, in our experiments, these initial values are usually good enough for the EM algorithm to converge to the true parameters.

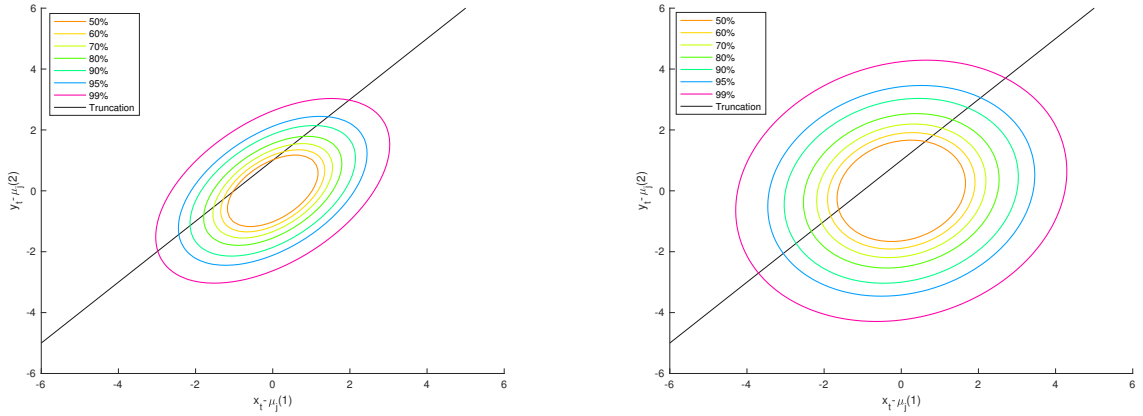
[6] The stopping criterium is set such that either 200 iterations are reached or the increase in log-likelihood (3.2) is less than e^{-10}.

Three components	α	μ	Σ	
True	0.2	10 9	1 0.5 0.5 1	
	0.3	2 2	3 1 1 3	
	0.5	−4 −6	5 2 2 5	
EM (T=200)	0.1971 (0.0280)	9.7305 (1.4018) 8.7987 (1.2179)	1.0930 (0.4596) 0.5093 (0.2354)	0.5093 (0.2354) 1.0291 (0.4231)
	0.2966 (0.0350)	2.2173 (1.5182) 2.1865 (1.3117)	2.9945 (0.9741) 1.0580 (0.5984)	1.0580 (0.5984) 2.9086 (1.1476)
	0.5063 (0.0349)	−4.1199 (0.4724) −5.9924 (0.3916)	5.2792 (1.3622) 2.0684 (0.7550)	2.0684 (0.7550) 4.8277 (0.9366)
EM (T=1000)	0.1991 (0.0144)	10.0053 (0.1045) 9.0070 (0.1022)	0.9981 (0.1107) 0.5074 (0.0882)	0.5074 (0.0882) 1.0219 (0.1183)
	0.3010 (0.0159)	2.0231 (0.2787) 2.0295 (0.2621)	2.9561 (0.4035) 0.9945 (0.2765)	0.9945 (0.2765) 3.0806 (0.5105)
	0.4999 (0.0176)	−3.9821 (0.1661) −5.9978 (0.1804)	4.9736 (0.4139) 2.0095 (0.3685)	2.0095 (0.3685) 4.9974 (0.5284)

Note: the numbers in parentheses are standard errors.

Table 2: Simulation results for DGP 2

To visualize the truncations on the mixture distribution, we plot in Figure 1 the truncations for two component in DGP 1. For a better comparison, each component is re-centered at the origin (shifted by μ_j) together with the truncation lines.

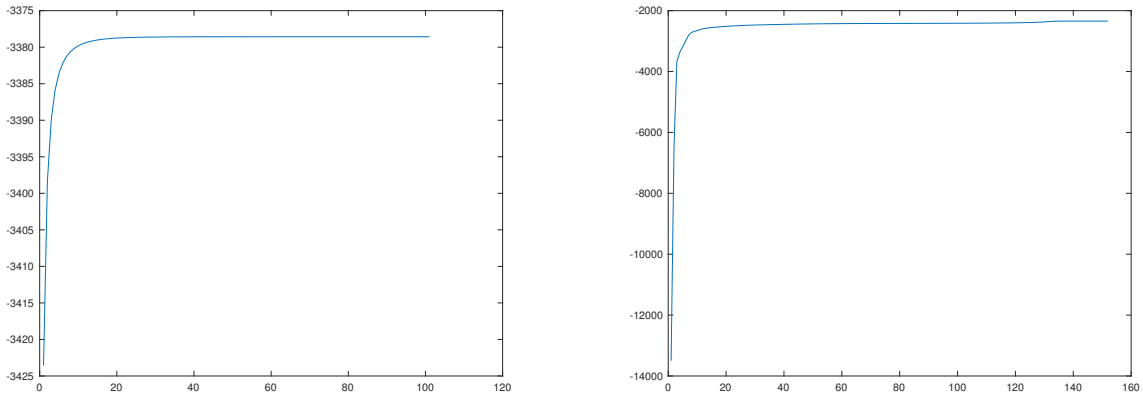


(a) Component 1 (b) Component 2

Note: $\mu_j(i)$ is the i^{th} element of μ_j.

Figure 1: Truncation of component density

In Figure 2(a), we plot the likelihood (3.2) for a one-time implementation of the EM algorithm in DGP 1. It provides the evidence that monotonicity in likelihood holds for the new EM algorithm. Moreover, the speed of convergence is fast with convergence achieved in about 20 iterations.



(a) Unconditional experiment (b) Conditional experiment

Figure 2: Log-likelihood

5.2 Conditional case experiments

Unlike the unconditional case where each observation is temporally independent, the conditional case carries time dependence in each observation. Hence, the data generating process is slightly different. First, we set the parameters as in Table 3. Second, at time t, we calculate $\eta_{t,j}$ for each j, where the subscript t comes from $F_{t,j}$ as μ_j is now replaced with $\mu_{t,j}$. Notice that α_j is fixed while $\eta_{t,j}$ changes with time. Third, independent random draws (e.g. 1000 draws) are made from the bivariate normal mixture distribution (component weight $\eta_{t,j}$). Fourth, we keep the draws that satisfy the constraint $x_t \geq y_t$, from which one is selected randomly as the observation at time t. Repeat the above steps until a sample with desired sample size is generated.[7]

In Table 3, we design three cases (DGP 3 and DGP 4 are $TMT(2,1)$, and DGP 5 is $TMT(3,1)$). Specifically, DGP 3 considers two components with the constraint ($x_t \geq y_t$) binding for one but not the other. [8] DGP 4 focuses on the case where the constraint is binding for both components. DGP 5 is a combination with the restriction not binding, binding with low persistency, and binding with high persistency. To visualize the constraint, we plot in Figure 3 the truncations for DGP 5. As the truncation is time varying, it cuts the density at different locations after re-centering (shifted by $\mu_{t,j}$ for each t and each j). The variation in truncations is smaller for the low persistency component because the location of truncation is more likely to be dominated by the constant C_j.

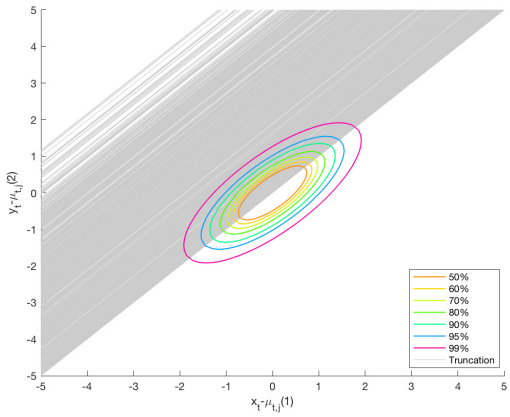
[7] Similar to section 5.1, the first 100 observations are discarded.

[8] The constraint will not be binding if $w'\mu_{t,j} = w'(C_j + B_{j,1}Y_{t-1} + ... + B_{j,Q}Y_{t-Q}) \gg 0$. In our simulation, we fix B and manipulate C to allow the restriction to be binding or not.

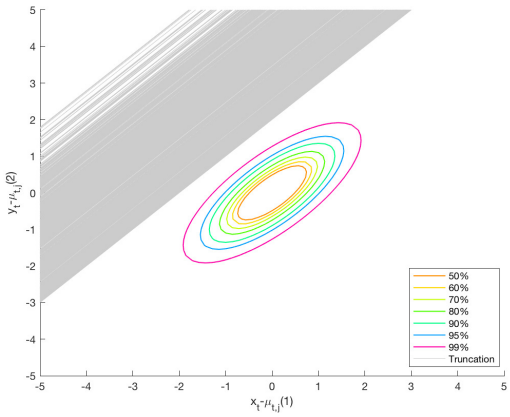
DGP		α	C	B		Σ	
3	NB	0.4	2	0.1	−0.8	0.4	0.3
			0	−0.8	0.1	0.3	0.4
	B	0.6	−2	0.7	−0.1	0.4	0.3
			−2	−0.1	0.7	0.3	0.4
4	B	0.4	0	0.1	−0.8	0.4	0.3
			0	−0.8	0.1	0.3	0.4
	B	0.6	2	0.2	−0.1	0.4	0.3
			2	−0.1	0.2	0.3	0.4
5	B	0.5	2	0.1	−0.8	0.4	0.3
			2	−0.8	0.1	0.3	0.4
	NB	0.3	2	0.3	−0.4	0.4	0.3
			0	−0.4	0.3	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

Note: B and NB denote binding and not binding respectively.

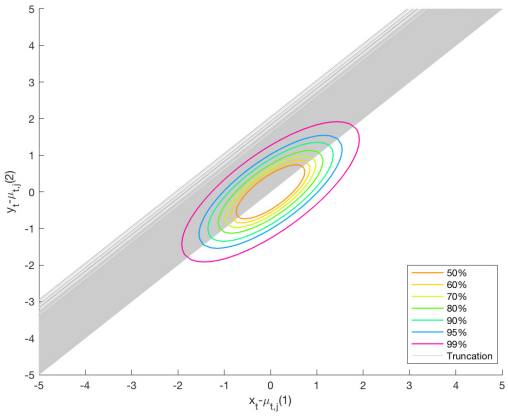
Table 3: Data Generating Process (DGP 3 - DGP 5)



(a) Binding (high persistency)



(b) Not Binding



(c) Binding (low persistency)

Figure 3: Truncations of DGP 5

We initialize the EM algorithm by randomly choosing 50 points from the parameter space.[9] Each point runs EM algorithm separately. The one that achieves the highest likelihood is chosen. We consider two sample sizes ($T = 200$ and $T = 1000$). The number of Monte Carlo replications is 100.

We summarize the average results across replications from Table 4 to Table 6. Standard errors are calculated over replications. In all cases, the EM algorithm performs satisfactory

[9]Elements of α is 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 9, 200 initial points are chosen to account for a higher dimensional parameter space.

in both small and large sample experiments. The standard error shrinks towards zero as the sample size increases. Last but not least, we can see in Figure 2(b) that the monotonicity of EM algorithm is preserved for the likelihood (3.1).

DGP 3	α	C	B		Σ	
True	0.4	2	0.1	−0.8	0.4	0.3
		0	−0.8	0.1	0.3	0.4
	0.6	−2	0.7	−0.1	0.4	0.3
		−2	−0.1	0.7	0.3	0.4
EM (T=200)	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)
	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)
EM (T=1000)	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)
	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)

Note: the numbers in parentheses are standard errors.

Table 4: Simulation results for DGP 3

DGP 4	α	C	B		Σ	
True	0.4	0	0.1	−0.8	0.4	0.3
		0	−0.8	0.1	0.3	0.4
	0.6	2	0.2	−0.1	0.4	0.3
		2	−0.1	0.2	0.3	0.4
EM (T=200)	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)
	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)
EM (T=1000)	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)
	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)

Note: the numbers in parentheses are standard errors.

Table 5: Simulation results for DGP 4

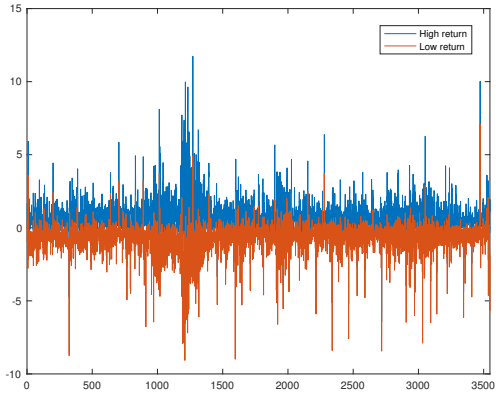
DGP 5	α	C	B		Σ	
True	0.5	2	0.1	−0.8	0.4	0.3
		2	−0.8	0.1	0.3	0.4
	0.3	2	0.3	−0.4	0.4	0.3
		0	−0.4	0.3	0.3	0.4
	0.2	−2	0.2	−0.1	0.4	0.3
		−2	−0.1	0.2	0.3	0.4
EM (T=200)	0.5078 (0.0427)	1.9985 (0.1535)	0.0978 (0.0551)	−0.8027 (0.0687)	0.3910 (0.0665)	0.2908 (0.0560)
		1.9867 (0.1433)	−0.7980 (0.0515)	0.0972 (0.0642)	0.2908 (0.0560)	0.3911 (0.0594)
	0.2932 (0.0378)	2.0114 (0.1505)	0.2991 (0.0640)	−0.3892 (0.0771)	0.3665 (0.0887)	0.2736 (0.0746)
		0.0055 (0.1390)	−0.4047 (0.0610)	0.3111 (0.0753)	0.2736 (0.0746)	0.3664 (0.0792)
	0.1990 (0.0294)	−2.0138 (0.2691)	0.2002 (0.0892)	−0.1047 (0.1025)	0.3873 (0.0944)	0.2917 (0.0820)
		−1.8540 (0.4928)	−0.1382 (0.1145)	0.2350 (0.1285)	0.2917 (0.0820)	0.4110 (0.1459)
EM (T=1000)	0.5000 (0.0178)	2.0026 (0.0583)	0.0990 (0.0223)	−0.8016 (0.0247)	0.3966 (0.0272)	0.2995 (0.0228)
		1.9920 (0.0574)	−0.7969 (0.0220)	0.0953 (0.0249)	0.2995 (0.0228)	0.3991 (0.0253)
	0.3001 (0.0167)	1.9968 (0.0710)	0.3008 (0.0273)	−0.3987 (0.0304)	0.3907 (0.0334)	0.2963 (0.0274)
		−0.0054 (0.0707)	−0.3983 (0.0263)	0.2990 (0.0308)	0.2963 (0.0274)	0.3986 (0.0334)
	0.1999 (0.0115)	−1.9983 (0.0954)	0.2003 (0.0298)	−0.0987 (0.0437)	0.3886 (0.0483)	0.2914 (0.0382)
		−2.0139 (0.1007)	−0.0960 (0.0310)	0.1996 (0.0427)	0.2914 (0.0382)	0.3906 (0.0493)

Note: the numbers in parentheses are standard errors.

Table 6: Simulation results for DGP 5

6 Empirical Application

We apply TMT to model the interval-valued IBM daily stock returns. The high/low return is calculated as the percentage change of the highest/lowest daily price with respect to the closing price of the previous day. For example, the high return at time t is: $r_{high,t} = 100(P_{high,t} - P_{close,t-1})/P_{close,t-1}$. The data is constructed as an interval-valued time series with $r_{high,t} \geq r_{low,t}$. To visualize the data, we plot a sample from 2004/1/1 to 2018/4/1 (3584 observations) in Figure 4. We can see that the volatility for the high and low returns is high in some periods while remaining quiet in others, suggesting potentially the presence of multiple regimes in the variance of the system.



(a) Real data

Figure 4: Daily IBM High/Low Stock Returns (2004/1/1 to 2018/4/1)

We consider TMT model with up to seven components and four lags. That is, $P = \{2, ..., 7\}$, and $Q = \{1, 2, 3, 4\}$, with total 28 specifications. [10] The best fitted model selected by BIC is $TMT(4, 2)$. The estimation results are reported in Table 7.[11] It is interesting to see that the fourth component has high volatility (big Σ) while only happens with a small probability (small α). Figure 5 shows the truncations for each component across time after re-centering (shifted by $\mu_{t,j}$ for each t and each j). The truncations vary by component: the first and second components have truncations almost not binding while for the last two components the truncations are binding.

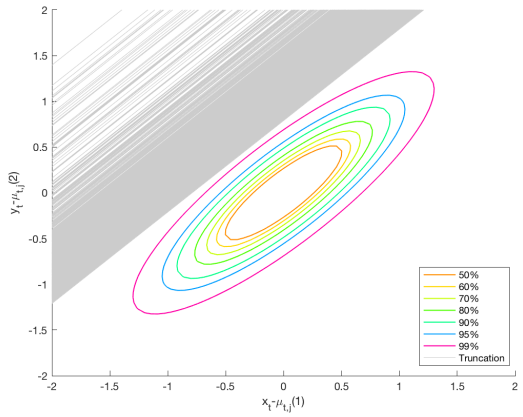
[10] The case when only one component is involved ($TMT(1, Q)$) turns out to be the same as GL, which, for a better comparison, will be discussed in the following separately.

[11] Standard errors are calculated using block bootstrap (Politis and White 2004)

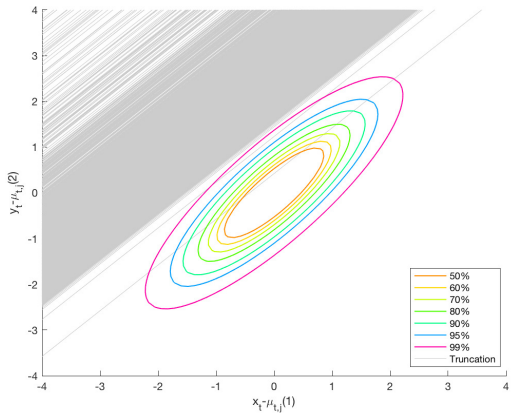
Component	α	C	B_1		B_2		Σ	
1	0.4184	0.3916	0.0681	−0.1033	−0.0276	0.0327	0.1838	0.1600
	(0.0428)	(0.0535)	(0.0331)	(0.0412)	(0.0368)	(0.0370)	(0.0230)	(0.0195)
		−0.2864	−0.0683	0.0801	−0.1285	0.1480	0.1600	0.1909
		(0.0688)	(0.0411)	(0.0501)	(0.0402)	(0.0397)	(0.0195)	(0.0204)
2	0.3635	0.3678	0.1758	−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)

Note: the numbers in parentheses are standard errors.

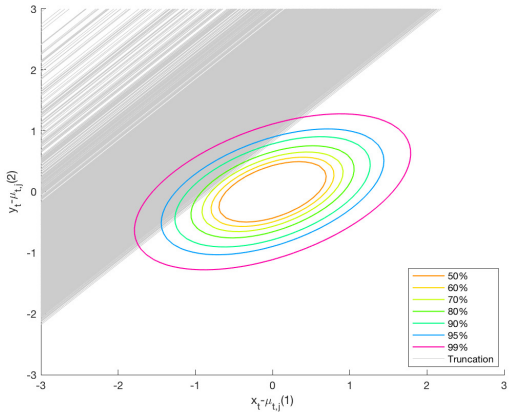
Table 7: Estimation results of $TMT(4, 2)$



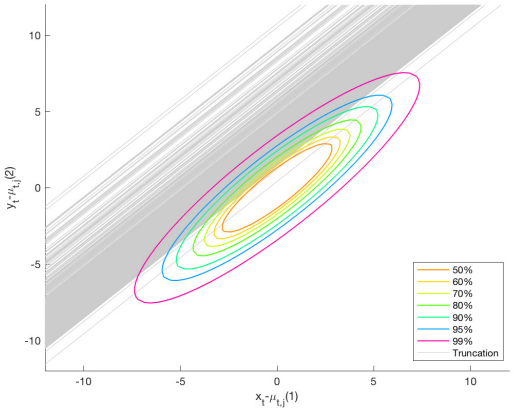
(a) First component



(b) Second component



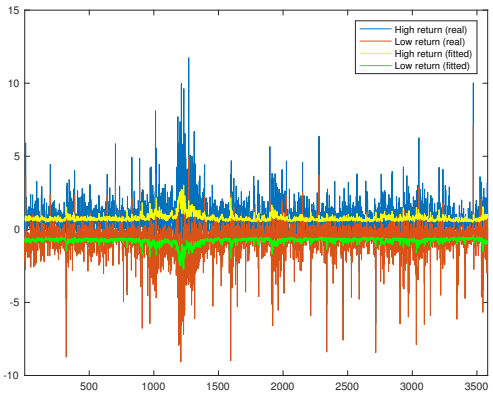
(c) Third component



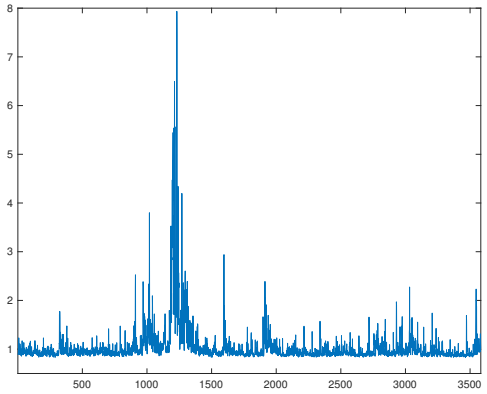
(d) Fourth component

Figure 5: Truncations for the fitted $TMT(4, 2)$

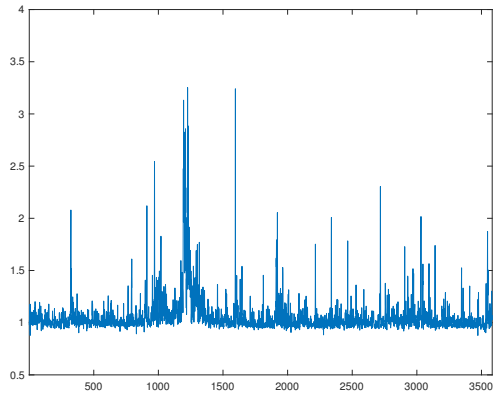
We plot in Figure 6 the fitted conditional means (2.3) together with the realized data. The persistency in the data seems to be well described. Figure 6 also shows the fitted conditional variances and correlation coefficients (2.5) of the high/low returns. The spikes in the fitted variances is aligned with the volatility clustering in the data. The contemporaneous conditional correlations stay at a relatively high level most of the time while drop toward zero during the volatile periods. It aligns with the observation that the ranges (gaps between upper and lower bounds) tend to be larger in these periods. In Figure 7, we plot some fitted conditional densities to illustrate the flexibility of the truncated normal mixture distribution.



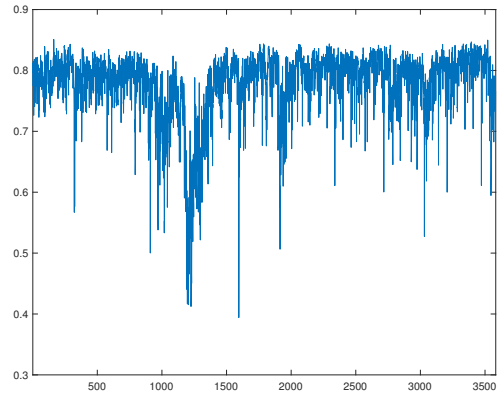
(a) Fitted Conditional Mean



(b) Fitted Variance (High Return)

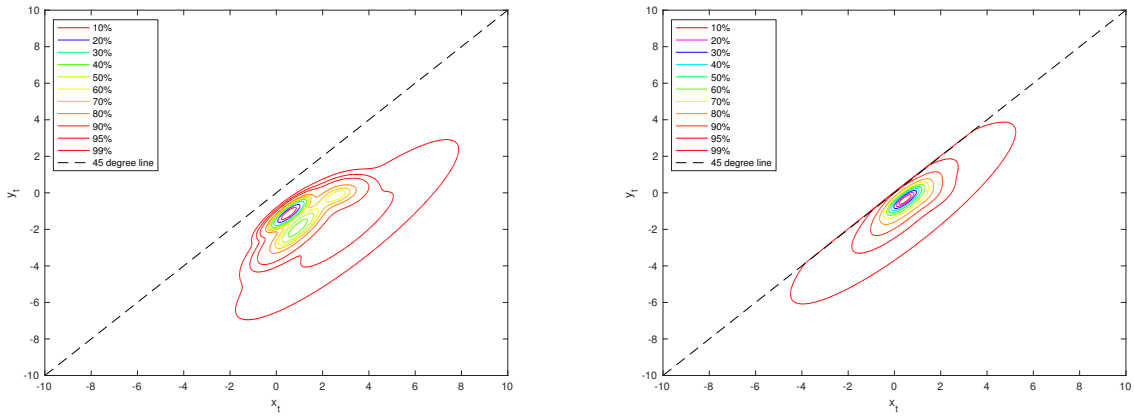


(c) Fitted Variance (Low Return)



(d) Fitted Correlation

Figure 6: Fitted Conditional Mean, Variance and Correlation of Daily IBM High/Low Stock Returns (2004/1/1 to 2018/4/1)



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

Figure 7: Fitted Conditional Density Contours

Given that not all the parameters in Table 7 are significant, and to account for the increase in parameters when the number of components grows larger, we also consider a restricted version of the model, $RTMT(P)$, where the restriction is imposed such that each component contains only one lag from the information set. For instance, $\mu_{t,j} = C_j + B_{j,j}Y_{t-j}$ and $B_{j,r} = 0$ for $r \neq j$. We consider up to seven components ($P = \{1, ..., 7\}$) for $RTMT$. Finally, we compare the TMT and $RTMT$ models with four other models. The number of lags for these models is selected using BIC. The linear vector autoregressive model serves as a benchmark. Two multivariate $GARCH$ models are considered to account for conditional heteroskedasticity in the data. See Bauwens et. al. (2006) for a review of multivariate $GARCH$ models. We also implement GL. Notice that, however, VAR and $VAR-MGARCH$ models cannot preserve the natural order of the ITS. A detailed comparison of the six models is summarized in Table 8.

Model for the mean	Model for the variance	Log-likelihood	Number of parameters	BIC
$VAR(7)$	_	-8604	30	-17,454
$VAR(7)$	MGARCH(1,1)-SBEKK	-8175	35	-16,064
$VAR(7)$	MGARCH(1,1)-DCC	-8155	39	-15,991
$GL(7)$	_	-8486	33	-17,243
$RTMT(5)$	_	-6975	49	-14,352
$TMT(4,2)$	_	-6833	55	-14,117

Table 8: Evaluation of models

TMT achieves the highest BIC and likelihood while using the most parameters. $RTMT$

trades the likelihood and BIC for a smaller number of parameters. VAR uses the smallest number of parameters and ends up having the smallest likelihood and BIC. After accounting for time-varying conditional variance, the $VAR - MGARCH$ models improve the performance over VAR significantly, implying that the data is conditional heteroskedastic. In terms of all criteria, GL lies in between VAR and $VAR - MGARCH$. This suggests that although GL preserves the natural oder of the interval data, it has a limited ability accommodating conditional heteroskedasticity.

7 Conclusions

We propose a truncated mixture transition model for the interval-valued time series. The natural order of the data (upper bound greater than lower bound) is guaranteed in our model using truncated normal distributions. The model enjoys great flexibility in terms of both parameter and density specifications. However, the standard EM algorithm to estimate mixture models fails since no closed-form solutions can be obtained in M step. Therefore, a new EM algorithm is proposed, which brings the pseudo data generating process to a higher level and encloses a closed-form solution in M step. We prove the consistency of the maximum likelihood estimator. Simulation results show that the new EM algorithm performs well. Last but not least, we illustrate the performance of the model with an application to the IBM daily high/low stock returns and it ourperforms other competing models.

References

[1] Amemiya, T. (1973), "Regression analysis when the dependent variable is truncated normal". *Econometrica*, Vol. 41, pp. 997–1016.

[2] Barndorff-Nielsen, O. (1965), "Identifiability of mixtures of exponential families," *Journal of Mathematical Analysis and Applications*, Vol. 12, pp. 115-121.

[3] Bauwens, et. al. (2006), "Multivariate GARCH models: a survey", *Journal of Applied Econometrics*, Vol. 21, pp. 79-109.

[4] Berchtold, André and Raftery, Adrian (2002), "The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series", *Statist. Sci.*, Vol. 17, pp. 328-356.

[5] Chang, S.H., et. al. (2011), "Chernoff-Type Bounds for the Gaussian Error Function", *IEEE Trans. Commun.*, Vol. 59, pp. 2939–2944.

[6] Dempster, A., et. al. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, pp. 1-38.

[7] González-Rivera, Gloria and Lin, Wei (2013), "Constrained Regression for Interval-valued Data," *Journal of Business and Economic Statistics*, Vol. 31, pp. 473–490.

[8] Hamilton, James D (1990), "Analysis of time series subject to changes in regime". *Journal of Econometrics*, Vol. 45, pp. 39-70.

[9] Hassan, M.Y. and Lii, Keh-Shin (2006), "Modeling marked point processes via bivariate mixture transition distribution models", *Journal of the American Statistical Association*, Vol. 101, pp. 1241-1252.

[10] Kalliovirta, Leena, et. al. (2016), "Gaussian mixture vector autoregression", *Journal of Econometrics,*Vol. 192, pp. 485-498.

[11] Krengel, U. (1985), *Ergodic Theorems.* de Gruyter, Berlin.

[12] Nhu D. Le, et. al. (1996), "Modeling flat stretches, bursts and outliers in time series using mixture transition distribution models", *Journal of the American Statistical Association*, Vol. 91, No. 436.

[13] Lee, Gyemin and Scott, Clayton (2012), "EM algorithms for multivariate Gaussian mixture models with truncated and censored data", *Computational Statistics & Data Analysis*, Vol. 56, pp. 2816-2829.

[14] Lima Neto, E., and de Carvalho, F. (2010), "Constrained linear regression models for symbolic interval-valued variables," *Computational Statistics and Data Analysis*, Vol. 54, pp. 333-347.

[15] McNeil and Frey (2000), "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach", *Journal of Empirical Finance*, Vol. 7, pp.271-300.

[16] Nath, G. Baikunth (1972), "Moments of a linearly truncated bivariate normal distribution", *Austral. J. Statist.*, Vol. 14, pp. 97-102.

[17] Potscher, B.M. and Prucha, I.R. (1991), "Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part i: consistency and approximation concepts", *Econometric Reviews*, Vol. 10, pp.125-216.

[18] Rao, R. Ranga (1962), "Relations between Weak and Uniform Convergence of Measures with Applications", *Ann. Math. Statist,* Vol. 33, pp.659-680.

[19] Straumann, D. and Mikosch, T. (2006), "Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach", *Ann. Statist*, Vol. 34, pp. 2449-2495.

[20] Teicher, Henry (1961), "Identifiability of Mixtures", *Ann. Math. Statist.*, Vol. 32, 244-248.

[21] Tu, Y. and Y. Wang (2016), "Center and log range models for interval-valued data with an application to forecast stock returns", *working paper*.

[22] Wong, C., & Li, W. (2000). "On a Mixture Autoregressive Model", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 62, pp. 95-115.

[23] Wu, C. F. Jeff. (1983), "On the Convergence Properties of the EM Algorithm", *Ann. Statist.*, Vol. 11, 95-103.

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

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

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

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

Hence, we found the upper bound of $\frac{1}{2}erfc(\frac{\lambda}{\sqrt{2}})\lambda$, and it suffices to show that

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

$$\iff 1 \geq \frac{1}{\frac{1}{2}+\sqrt{\frac{1}{4}+\frac{2}{\pi\lambda^2}}}$$

which obviously holds when $\lambda > 0$.

A2. The EM algorithm for normal mixture model (unconditional case)

This section reviews the EM algorithm when the component density $f_j(Y_t|\mu_j, \Sigma_j)$ in (3.2) is a bivariate normal distribution. EM algorithm transforms the problem into a missing data framework and constructs a pseudo complete data generating process. It starts by assuming that each observation comes from one of the P components, and there is a latent variable indicating which component the observation truly comes from. Let $z_{tj} \in \{0, 1\}$ be the indicator variable such that $z_{tj} = 1$ if Y_t is generated from component j and 0 otherwise. The objective is to maximize the pseudo complete likelihood of $\{Y, z\}$. Denote $z_t = \{z_{t1}, ..., z_{tP}\}$. To construct the complete likelihood, the latent variable z_{tj} is specified to follow a multinomial distribution:

$$g(z_t|\Theta) = \prod_{j=1}^{P} \alpha_j^{z_{tj}} \tag{7.1}$$

The conditional density of Y_t on z_t is

$$h(Y_t|z_t, \Theta) = \prod_{j=1}^{P} \Big[f_j(Y_t|\mu_j, \Sigma_j) \Big]^{z_{tj}} \tag{7.2}$$

The complete density function becomes

$$
\begin{aligned}
l(Y_t, z_t|\Theta) &= g(z_t|\Theta)h(Y_t|z_t,\Theta) \\
&= \prod_{j=1}^{P}\Big[\alpha f_j(Y_t|\mu_j,\Sigma_j)\Big]^{z_{tj}}
\end{aligned} \tag{7.3}
$$

Therefore, the complete log-likelihood function for Θ can be written as

$$
L^C(\Theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} z_{tj} log\alpha_j + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} z_{tj} log f_j(Y_t|\mu_j,\Sigma_j) \tag{7.4}
$$

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 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).

$$
Q(\Theta|\Theta^l) = E(L^C(\Theta)|Y,\Theta^l) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} \tilde{z}_{tj} log\alpha_j + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P} \tilde{z}_{tj} log f_j(Y_t|\mu_j,\Sigma_j) \tag{7.5}
$$

$$
\begin{aligned}
\tilde{z}_{tj} &\equiv E(z_{tj}|Y_t,\Theta^l) \\
&= P(z_{tj}|Y_t,\Theta^l) \\
&= \frac{P(z_{tj},Y_t,\Theta^l)}{P(Y_t,\Theta^l)} \\
&= \frac{\alpha_j^l f_j(Y_t|\mu_j^l,\Sigma_j^l)}{\sum_{k=1}^{P}\alpha_k^l f_k(Y_t|\mu_j^l,\Sigma_j^l)}
\end{aligned} \tag{7.6}
$$

M Step: The updated parameter set is obtained by $\Theta^{l+1} = \underset{\Theta}{argmax} Q(\Theta|\Theta^l)$:

$$\alpha_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj}}{T} \tag{7.7}$$

$$\mu_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj} Y_t}{\sum_{t=1}^{T} \tilde{z}_{tj}} \tag{7.8}$$

$$\Sigma_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj} (Y_t - \mu_j^{l+1})(Y_t - \mu_j^{l+1})'}{\sum_{t=1}^{T} \tilde{z}_{tj}} \tag{7.9}$$

Iterate E step and M step until convergence. Dempster et. al. (1977) pointed out that the likelihood (3.2) is closely related to the feasible pseudo complete likelihood (7.5): $L(\Theta^l) = Q(\Theta^l|\Theta^l) \leq Q(\Theta^{l+1}|\Theta^l) \leq L(\Theta^{l+1})$. Therefore, as $Q(\Theta|\Theta^l)$ is maximized in each iteration (which implies $Q(\Theta^l|\Theta^l) \leq Q(\Theta^{l+1}|\Theta^l)$), the likelihood (3.2) increases monotonically ($L(\Theta^{l+1}) \geq L(\Theta^l)$).

A.3 The EM algorithm for truncated normal mixture model (unconditional case)

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:

E Step: Following the same steps as Appendix A.1, the expression for $\tilde{z}_{tj}$ is the same as (7.6). However, $f_j(Y_t|\mu_j^l, \Sigma_j^l)$ is now a truncated bivariate normal distribution.

M Step:

$$\alpha_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj}}{T} \tag{7.10}$$

$$\mu_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj} Y_t}{\sum_{t=1}^{T} \tilde{z}_{tj}} - v_j(\mu_j^{l+1}, \Sigma_j^{l+1}) \tag{7.11}$$

$$\Sigma_j^{l+1} = \frac{\sum_{t=1}^{T} \tilde{z}_{tj} (Y_t - \mu_j^{l+1})(Y_t - \mu_j^{l+1})'}{\sum_{t=1}^{T} \tilde{z}_{tj}} + I_j(\mu_j^{l+1}, \Sigma_j^{l+1}) \tag{7.12}$$

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.3.1.

A.3.1 Derivation of the EM algorithm

Let Y follows a truncated bivariate normal distribution:

$$f(Y) = \frac{1}{2\pi|\Sigma|[1-\Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})]} exp[-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)] \tag{7.13}$$

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, we can write down the conditional expectation of the complete log-likelihood function:

$$\begin{aligned} Q(\Theta|\Theta^l) = E(L^C(\Theta)|Y,\Theta^l) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{P}\tilde{z}_{tj}\Bigg[log\alpha_j - log2\pi - \frac{1}{2}log|\Sigma_j| \\ - \frac{1}{2}(Y_t-\mu_j)'\Sigma_j^{-1}(Y_t-\mu_j) - log(1-\Phi(\frac{-w'\mu_j}{\sqrt{w'\Sigma_j w}}))\Bigg] \end{aligned}$$

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, we take 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

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

We can get:

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

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

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

First, we can get

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

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

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

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, we can get:

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

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.4 E step of the new EM algorithm

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

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

A.5 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 of the has the following form:

$$f(Y,\mu,\Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}[1-\Phi(\frac{w'\mu}{\sqrt{w'\Sigma w}})]} exp[-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)] \tag{7.14}$$

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

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

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

- Take derivative of (3.8) with respect to μ_j.

$$\frac{\partial Q(\Theta|\Theta^l)}{\partial \mu_j} = \frac{1}{T}\sum_{t=1}^{T} \tilde{z}_{tj}[\Sigma_j^{-1}Y_t - \Sigma_j^{-1}\mu_j + \tilde{n}_{t,j}\int(\Sigma_j^{-1}Y_{t,k} - \Sigma_j^{-1}\mu_j)(\frac{f_j^{N,l}(Y_{t,k})}{1-F_j^l})dY_{t,k}] = 0$$

$$\Rightarrow \sum_{t=1}^{T}\tilde{z}_{tj}Y_t - \mu_j\sum_{t=1}^{T}\tilde{z}_{tj} + \sum_{t=1}^{T}\tilde{z}_{tj}\tilde{n}_{t,j}(M_{d,j}^{1,l} + \mu_j^l) - \mu_j\sum_{t=1}^{T}\tilde{z}_{tj}\tilde{n}_{t,j} = 0$$

$$\Rightarrow \mu_j^{l+1} = \frac{\sum_{t=1}^{T}\tilde{z}_{tj}(Y_t + \tilde{n}_{t,j}(M_{d,j}^{1,l} + \mu_j^l))}{\sum_{t=1}^{T}\tilde{z}_{tj}(1+\tilde{n}_{t,j})}$$

where $M_{d,j}^{1,l}$ is M_d^1 with $\mu = \mu_j^l$, $\Sigma = \Sigma_j^l$.

- Take derivative of (3.8) with respect to Σ_j^{-1}.

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

where $M_{d',j}^2 = M_{d,j}^{2,l} + (\mu_j^l - \mu_j^{l+1})(M_{d,j}^{1,l})' + (M_{d,j}^{1,l})(\mu_j^l - \mu_j^{l+1})' + (\mu_j^l - \mu_j^{l+1})(\mu_j^l - \mu_j^{l+1})'$, and $M_{d,j}^{2,l}$ is M_d^2 with $\mu = \mu_j^l$, $\Sigma = \Sigma_j^l$.

A.6 M step of the new EM algorithm (conditional case)

The closed-form solution for α_j and Σ_j can be easily derived similar to the unconditional case. Here we focus on 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:

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

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 derivative of $L(A)$ with respect to $vec(A_j')$:

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

Then, we can write down $vec(A_j')$ as:

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

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.7 Proof of Theorem 1

First, we introduce a lemma that shows the mixture truncated normal distribution is identifiable.

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 distributions whose *density is given by*

$$f(Y,\nu) \;=\; \frac{1}{2\pi\sqrt{|\Sigma|}[1-\Phi(\frac{-w'\mu}{\sqrt{w'\Sigma w}})]}exp[-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)] \tag{7.15}$$

Then $\psi_\gamma(Y) = \sum_{j=1}^{P}\alpha_j F(Y,\nu_j)$*, the class of finite mixtures of* Λ*, is identifiable.* $\gamma = \{\alpha_j,\nu_j|\forall j\}$*,* $\alpha_j > 0$*, and* $\sum_{j=1}^{P}\alpha_j = 1$*. In other words,* $\psi_\gamma(Y) = \psi_{\gamma^*}(Y) \Rightarrow \gamma = \gamma^*$*.*

Proof of Lemma 1:

We first define the exponential family.

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

$$dF(Y,\tau) \;=\; a(\tau)b(Y)exp[\tau'h(Y)]d\mu(Y) \tag{7.16}$$

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 F is called an exponential family member.

Barndorff-Nielsen (1965) proves that the class ψ is identifiable if all of the following hold: (a) F belongs to the exponential family, (b) μ is n-dimensional Lebesgue measure, (c) functions h_j, $j = 1,2,\ldots,m$, are all continuous, and (d) the set $\{y : y = h(Y), b(Y) > 0, Y \in \mathbb{R}^n\}$ contains a nonempty open set.

First, we show that the distribution with density given by (7.15) belongs to exponential family as 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 can 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: $\sup_{\Psi \in \Xi} \mid L(\Psi) - E[L(\Psi)] \mid \to 0$ 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 $E[\sup_{\Psi \in \Xi} \mid l(\Psi) \mid] < \infty$ (see Rao (1962) or Straumann and Mikosch (2006 Theorem 2.7)). Kalliovirta et.al. (2016) prove the the above inequality holds for the likelihood in their model one side at a time. We are going to adapt similar similar procedures here. Specifically, we know that

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

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 $\gamma > 0$. We also know that $exp[-\frac{1}{2}(Y_t - A_j X_{t-1})' \Sigma_j^{-1} (Y_t - A_j X_{t-1})] \leq 1$. In addition, when $-w' A_j X_{t-1} / \sqrt{2w' \Sigma_j w} \leq 0$, $erfc(-w' A_j X_{t-1} / \sqrt{2w' \Sigma_j w}) \geq 1$, and thus we can see that $l(\Psi) \leq log(\pi^{-1} \delta^{-1/2})$. When $-w' A_j X_{t-1} / \sqrt{2w' \Sigma_j w} > 0$, we apply the inequality $erfc(x) \geq \frac{1}{2} exp(-2x^2)$ (see Chang et. al. (2011, Theorem 2)), thus

$$
\begin{aligned}
erfc(-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w}) &\geq \frac{1}{2}exp(-w'A_jX_{t-1}X'_{t-1}A'_jw/w'\Sigma_j w) \\
&\geq \frac{1}{2}exp[-\frac{1}{\gamma}tr(X_{t-1}X'_{t-1}A'_jww'A_j)] \\
&\geq \frac{1}{2}exp[-\frac{1}{\gamma}tr(X_{t-1}X'_{t-1})tr(A'_jww'A_j)] \\
&\geq \frac{1}{2}exp[-\frac{\kappa}{\gamma}X'_{t-1}X_{t-1}]
\end{aligned}
$$

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

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

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

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, $erfc(-w'A_jX_{t-1}/\sqrt{2w'\Sigma_j w}) \leq 2$, 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, we have $G_1-\frac{1}{2}\rho(1+Y_t'Y_t+X_{t-1}'X_{t-1}) \leq l(\Psi) \leq log(2\pi^{-1}\delta^{-1/2})+\frac{\kappa}{\gamma}X_{t-1}'X_{t-1}$, from which $E[\sup_{\Psi\in\Xi} \mid l(\Psi) \mid] < \infty$ holds because $X_{t-1}'X_{t-1} = 1+Y_{t-1}'Y_{t-1}+...+Y_{t-Q}'Y_{t-Q}$, and $E(Y_t'Y_t) < \infty$ 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} as , 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, 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.