Estimating SUR Tobit Model while Errors are Gaussian Scale Mixtures: with an Application to High Frequency Financial Data

Qian, Hang

Iowa State University

July 2009

Estimating SUR Tobit Model while Errors are Gaussian Scale Mixtures

— with an Application to High Frequency Financial Data

Hang Qian
Iowa State University

Abstract

This paper examines multivariate Tobit system with Scale mixture disturbances. Three estimation methods, namely Maximum Simulated Likelihood, Expectation Maximization Algorithm and Bayesian MCMC simulators, are proposed and compared via generated data experiments. The chief finding is that Bayesian approach outperforms others in terms of accuracy, speed and stability. The proposed model is also applied to a real data set and study the high frequency price and trading volume dynamics. The empirical results confirm the information contents of historical price, lending support to the usefulness of technical analysis. In addition, the scale mixture model is also extended to sample selection SUR Tobit and finite Gaussian regime mixtures.

1. Introduction

The seminal work of Amemiya (1974) inspired a variety of research on Seemingly Unrelated Tobit Regression (SUR Tobit) model. Since its advent, SUR Tobit model has been firmly integrated with economic theory where corner solution prevails due to non-negativity constraints, capacity limits or government regulations. Wales and Woodland (1983) derived the censored demand systems based on Kuhn-Tucker conditions. Lee and Pitt (1986) adopted dual approach to the demand system using virtual price in place of complementary slackness conditions. Those models *per se* are elegant; the remaining difficulties are largely computational – likelihood function entails evaluations of multiple integrals without analytical solution.

Recent years witness increasing popularity of SUR Tobit model in applied micro-econometric studies. The driving force is apparently the advance of numerical methods and computation resources. Broadly speaking, there are four common methods to estimate the SUR Tobit system, namely Maximum Simulated Likelihood (MSL), Heckman Two Steps, Expectation-Maximization (EM) algorithm and Bayesian Posterior Simulator.

Firstly, likelihood involving multivariate normal c.d.f could be evaluated via numerical methods. For two or three dimensions, Gaussian quadrature is a good choice. Cornick et.al (1994) examined fluid milk (whole, reduced and skim) consumption patterns. They neglected the 4% of consumers in the dataset who did not consume any of the three milk types. So they only dealt with at most bivariate normal probability, which can be evaluated easily by, for instance, adaptive quadrature (see Genz, 2004). For higher dimension problems, simulation technique is necessary. Arias and Cox (2001) adopted MSL to study a four-goods dairy sector (cheese, butter, dry milk and manufacturing milk). GHK simulator (Geweke, 1991;

Borsch-Supan and Hajivassiliou, 1993; Keane, 1994) was used to evaluate the integral. Kao et.al (2001) also adopted that simulator to study a seven-goods linear expenditure demand system.

Secondly, the two step procedure offers a less computational alternative to MSL. Nowadays, it becomes clear that the Heien and Wessells (1990)'s two step procedure suffers from internal inconsistency, but their ideas do set the seeds for Shonkwiler and Yen (1999) modification. Su and Yen (2000) applied that procedure to study censored system of cigarette and alcohol consumption.

Thirdly, E-M algorithm can be suitably applied to SUR Tobit model, offering a consistent and asymptotically efficient estimator. Huang et.al (1987) provides analytical formulae for the E-step and M-step in the two-dimension censoring system. Another author Huang (1999) extended it to the multiple-equation case by Monte Carlo E-step plus conditional M-step.

Fourthly, Bayesian posterior simulator competes with the traditional ML based point estimator. Chib and Greenberg (1995, 1998) examined limited dependent variable model in the Bayesian SUR framework. Essentially observable data are augmented by draws from their truncated conditional distribution whenever censoring occurs. To apply the Bayesian approach to the SUR Tobit, Huang (2001) outlined the MCMC steps to implement the simulation.

It should be noted that the main attraction of SUR Tobit system lies in its flexible correlation structure, induced by multivariate normally distributed disturbances. (On the other hand, this is the source of computational burden!) Admittedly, it is possible to circumvent the normal distribution by resorting to GEV specification (e.g. Pudney, 1989), error component structure (Butler and Moffitt,1982), Copula approach (e.g. Yen and Lin, 2008 used Frank's copula with log-Burr marginals), etc, we do not want to impose stringent restrictions on disturbance terms, but rather lessen it by introducing mixture models, which admit normality as a special case. It turns out that our methods are still computationally tractable.

In the main text, we will discuss the estimation of SUR Tobit model with Gaussian scale-mixture disturbances. In microeconomics, individual unobserved heterogeneity might justify the varied error variance. In finance, it is well documented that security prices always exhibit leptokurtosis and/or skewness (e.g. Fama,1965; Praetz, 1972; Peiro, 1994, to name a few). Investors usually feel the stock prices crawling upwards for months and plummet in a day. The magnitude of crash on Black Monday of 1987 is hard to reconcile with the relative thin tail of normal distribution. It is common for high-frequent financial data exhibiting asymmetry, leptokurtosis and extreme value. Gaussian scale-mixture has a marginal distribution as multivariate t, thus is able to accommodate leptokurtosis. Of course, there are other tractable mixtures to model departure from normality, such as finite Gaussian regime mixtures, skewed normal, which could capture skewness as well. We will briefly discuss some extensions at the end of the paper.

The focus of the paper is on the various estimation methods for the SUR Tobit model with mixture errors. Treating the unobserved heterogeneity as an additional latent variable, we develop three methods to estimate the model: MSL, EM and Bayesian posterior sampler. Comparisons of the three methods are investigated through artificial data set with known DGP. We also illustrate our approach by studying the price and volume interactions using high-frequency trading data of financial derivatives, where non-trading accounts for a non-negligible portion of the data.

The rest of the paper is organized as follows. Section 2 outlines three estimation schemes for the SUR Tobit model with scale-mixture disturbance. Section 3 compares the relative efficiency of the three methods through simulated data. Section 4 applies our methods to a real data set to study price and volume interactions. Section 5 extent the current model to sample selection case and other types of mixture error specifications.

2. Estimation Methods

2.1 The model

The linear regression form of SUR Tobit with Gaussian scale-mixtures can be written as:

$Y_{1t}^* = \mathbf{x_{1t}}\,\boldsymbol{\beta_1} + \varepsilon_{1t}$, $Y_{1t} = \max(0, Y_{1t}^*)$

$Y_{2t}^* = \mathbf{x_{2t}}\,\boldsymbol{\beta_2} + \varepsilon_{2t}$, $Y_{2t} = \max(0, Y_{2t}^*)$

……

$Y_{m\,t}^* = \mathbf{x_{m\,t}}\,\boldsymbol{\beta_m} + \varepsilon_{m\,t}$, $Y_{mt} = \max(0, Y_{mt}^*)$

$\boldsymbol{\varepsilon_t}\,|\tau_t \sim \mathrm{MVN}(\mathbf{0}, \tau_t\,\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is m*m positive definite matrix

$\tau_t \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$

For future reference, stack the variables in the obvious way,

$$\mathbf{Y_t^*} = \begin{pmatrix} Y_{1t}^* \\ \dots \\ Y_{mt}^* \end{pmatrix},\ \mathbf{Y_t} = \begin{pmatrix} Y_{1t} \\ \dots \\ Y_{mt} \end{pmatrix},\quad \mathbf{X_t} = \mathrm{diag}(\mathbf{x_{1t}}, \dots, \mathbf{x_{m\,t}}),\ \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \dots \\ \boldsymbol{\beta_m} \end{pmatrix},\ \boldsymbol{\varepsilon_t} = \begin{pmatrix} \varepsilon_{1t} \\ \dots \\ \varepsilon_{mt} \end{pmatrix}$$

$$\mathbf{Y^*} = \begin{pmatrix} \mathbf{Y_1^*} \\ \dots \\ \mathbf{Y_T^*} \end{pmatrix},\ \mathbf{Y} = \begin{pmatrix} \mathbf{Y_1} \\ \dots \\ \mathbf{Y_T} \end{pmatrix},\ \mathbf{X} = \begin{pmatrix} \mathbf{X_1} \\ \dots \\ \mathbf{X_T} \end{pmatrix},\ \boldsymbol{\tau} = \begin{pmatrix} \tau_1 \\ \dots \\ \tau_T \end{pmatrix}$$

The latent variables $\mathbf{Y^*}$ and $\boldsymbol{\tau}$ are not observable. The goal of statistical inference is to estimate the parameter $\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu$ from the data $\mathbf{Y}$ and $\mathbf{X}$. Notice that in the model, conditional on latent scale τ_t , $\boldsymbol{\varepsilon_t}$ is multivariate normal $\mathrm{MVN}(\mathbf{0}, \tau_t\,\boldsymbol{\Sigma})$. For tractability, the latent scale has an inverted Gamma distribution $\mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$ so that $\boldsymbol{\varepsilon_t}$'s marginal distribution is multivariate t $\mathrm{MVT}(\mathbf{0}, \boldsymbol{\Sigma}, \nu)$ (See Appendix A1).

It is true that in some applications, researchers might encounter models more complicated than the above setup. For instance, Wales and Woodland (1983)'s Kuhn-Tucker model does not have a regression form, $\mathbf{Y_t^*}$ could be intertwined with $\mathbf{X_t}$ and $\boldsymbol{\beta}$ so that the system is non-linear. In some other applications, say Su and Yen (2000), the determination of individual participation and activity level are separated. However, the estimation methods outlined below can be adapted and easily extended to non-linearity and sample selection models.

Compared with single-equation Tobit, the multivariate system might have various patterns of censoring. To be more exact, an m-variate system has 2^m patterns ranging from no censoring to all censorings. Without loss of generality, for some individual t, rearrange the order of variables so that the first m_u dependent variables are intact but the last $m_c \equiv m - m_u$ variates are censored to zero. We use the subscript "u" to denote the non-censored block, and subscript "c" to denote the censored block. (The special cases $m_c = 0$ and $m_u = 0$ correspond to no censoring and all censoring respectively. For notational simplicity, treat the former as an empty "c" block, and the latter as an empty "u" block.)

For some individual t, the variables are partitioned to be:

$$\mathbf{Y_t^*} = \begin{pmatrix} \mathbf{Y_{ut}^*} \\ \mathbf{Y_{ct}^*} \end{pmatrix},\ \mathbf{Y_t} = \begin{pmatrix} \mathbf{Y_{ut}} \\ \mathbf{Y_{ct}} \end{pmatrix},\ \mathbf{X_t} = \begin{pmatrix} \mathbf{X_{ut}} & \\ & \mathbf{X_{ct}} \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_u} \\ \boldsymbol{\beta_c} \end{pmatrix},\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{uu}} & \boldsymbol{\Sigma_{uc}} \\ \boldsymbol{\Sigma_{cu}} & \boldsymbol{\Sigma_{cc}} \end{pmatrix}$$

The partition is useful since in all the three methods discussed below, distribution of censored block, conditional on the knowledge of uncensored block, will be intensively explored. It should be noted that the conditional distribution of MVN variates (blocks) are still normal, while that of MVT do not have a familiar form. Therefore, the inclusion of latent scale τ_t bridges the flexible MVT and the tractable MVN.

2.2 Maximum Simulated Likelihood

Full information maximum likelihood is the most straightforward way (at least conceptually) to estimate the above system, provided that the dimension is not too high. It counts the contribution of censored probability plus that of uncensored density. Fishy at the first glance, the blending likelihood has been shown by Amemiya (1973) to be valid. Schnedler (2005) further established rules to construct likelihood for more general censoring problems. The maximizer is proven to be root-N consistent, asymptotically efficient and normally distributed under regularity conditions.

In our model, the log likelihood is $\log L = \sum_{t=1}^{T} \ln p(\mathbf{Y_{ut}}, \mathbf{Y_{ct}})$, where

$$p(\mathbf{Y_{ut}}, \mathbf{Y_{ct}}) = \int_0^{\infty} \int_{-\infty}^{\mathbf{0}} p(\mathbf{Y_{ut}^*}, \mathbf{Y_{ct}^*}, \tau_t)\, d\, \mathbf{Y_{ct}^*}\, d\tau_t$$

$$= p(\mathbf{Y_{ut}^*}) \cdot \int_0^{\infty} \int_{-\infty}^{\mathbf{0}} p(\tau_t|\mathbf{Y_{ut}^*})\, p(\mathbf{Y_{ct}^*}|\, \mathbf{Y_{ut}^*}, \tau_t)\, d\, \mathbf{Y_{ct}^*}\, d\tau_t$$

To avoid confusion, the variables and the intervals have the order like $\int_{x_1 \in A_1} \int_{x_2 \in A_2} f(x_1, x_2)\, dx_2\, dx_1$

It can be shown that (see Appendix A2)

$\mathbf{Y_{ut}^*} \sim \mathrm{MVT}(\mathbf{X_{ut}}\, \boldsymbol{\beta_u}\, ;\, \boldsymbol{\Sigma_{uu}}, \nu)$

$\tau_t|\mathbf{Y_{ut}^*} \sim \mathrm{IG}\left\{ \frac{\nu+m_u}{2}\, , \left[\frac{\nu}{2} + \frac{1}{2} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\, \boldsymbol{\beta_u})' \boldsymbol{\Sigma_{uu}^{-1}} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\, \boldsymbol{\beta_u}) \right]^{-1} \right\}$

$\mathbf{Y_{ct}^*}|\, \mathbf{Y_{ut}^*}, \tau_t \sim \mathrm{MVN}\left[\mathbf{X_{ct}}\, \boldsymbol{\beta_c} + \boldsymbol{\Sigma_{cu}} \boldsymbol{\Sigma_{uu}^{-1}} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\, \boldsymbol{\beta_u})\, , \tau_t \left(\boldsymbol{\Sigma_{cc}} - \boldsymbol{\Sigma_{cu}} \boldsymbol{\Sigma_{uu}^{-1}} \boldsymbol{\Sigma_{uc}} \right) \right]$

The crux of ML estimation is to calculate $\int_0^{\infty} \int_{-\infty}^{\mathbf{0}} p(\tau_t|\mathbf{Y_{ut}^*})\, p(\mathbf{Y_{ct}^*}|\, \mathbf{Y_{ut}^*}, \tau_t)\, d\, \mathbf{Y_{ct}^*}\, d\tau_t$, which belongs to the c.d.f of the "multivariate generalized t-distributions family" (Arslan, 2004), with probability evaluated at rectangular region $(-\infty, 0)$ for each element. The integral does not have a close-form solution. For lower dimension case, Gaussian quadrature might provide a faster and more accurate solution, but simulation techniques must be employed for higher dimensional problems.

Nowadays, GHK simulator (Geweke, 1991; Borsch-Supan and Hajivassiliou, 1993; Keane, 1994) is a popular choice for evaluating probabilities of multivariate normal distribution. Hajivassiliou et al. (1996) reviewed many available simulators and concluded that GHK simulator outperforms others.

In this paper, we extend GHK simulator to evaluate multivariate t family c.d.f.. It turns out that only one more layer need to be added to the recursive conditionings. That is, we first obtain a draw from IG distribution and then start the usual GHK recursive truncated normal draws, and finally calculate a series of univarite normal c.d.f.s.

The details are as follows.

First consider a general problem: $\sim \mathrm{MVT}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, i.e. $\mathbf{X}|\tau \sim \mathrm{MVN}(\boldsymbol{\mu}, \tau\boldsymbol{\Sigma})$, $\tau \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$

We want to calculate $\Pr(\mathbf{X} < \boldsymbol{a})$

Cholesky decompose the covariance matrix $\boldsymbol{\Sigma} = \mathbf{P} \cdot \mathbf{P}'$, denote $\mathbf{P} = \begin{pmatrix} p_{11} & & & \\ p_{21} & p_{22} & & \\ \cdots & & & \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}$

Define $\mathbf{Y} = \tau^{-\frac{1}{2}} \mathbf{P}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ so that $\mathbf{Y}|\tau \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I})$

By change of variables, it is easy to show

$$\Pr(\mathbf{X} < a) = \int_0^{\infty} \int_{-\infty}^{c_1} \dots \int_{-\infty}^{c_k} \phi(Y_1) \cdot \dots \cdot \phi(Y_k)\, dY_k \dots dY_1 d\tau$$

$$= E_\tau\, E_{Y_1} \dots E_{Y_{k-1}} [\Phi(c_1) \cdot \dots \cdot \Phi(c_k)]$$

where $c_1 = \frac{a_1 - \mu_1}{\sqrt{\tau}\, P_{11}}$, $c_2 = \frac{a_2 - \mu_2 - \sqrt{\tau}\, P_{21} Y_1}{\sqrt{\tau}\, P_{22}}$, …, $c_k = \frac{a_k - \mu_k - \sqrt{\tau}\,(P_{k1} Y_1 + \dots + P_{k-1,k-1} Y_{k-1})}{\sqrt{\tau}\, P_{kk}}$

a_i , μ_i are i th element of $\mathbf{a}, \boldsymbol{\mu}$

$E_\tau(\cdot)$ is expectation w.r.t. $\tau \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$

$E_{Y_1}(\cdot)$ is expectation w.r.t. $Y_1 \sim \mathrm{TN}_{(-\infty, c_1)}(0,1)$

… …

$E_{Y_{k-1}}(\cdot)$ is expectation w.r.t. $Y_{k-1} \sim \mathrm{TN}_{(-\infty, c_{k-1})}(0,1)$

So the GHK simulator adapted to MVT c.d.f has the following algorithm

i) Generate a draw $\tau \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$

ii) Conditional on τ , recursively generate draws Y_1 , … , Y_{k-1}

iii) Calculate $\Phi(c_1) \cdot \dots \cdot \Phi(c_k)$

iv) Repeat i) - iii), and take the average of results of iii), which approximates $\Pr(\mathbf{X} < a)$

To apply the above simulator to our problem, we simply need to modify step i) by replacing the standard $\mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$ with the posterior $\mathrm{IG}\left\{\frac{\nu + m_u}{2}, \left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\, \boldsymbol{\beta_u})' \boldsymbol{\Sigma_{uu}^{-1}} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\, \boldsymbol{\beta_u})\right]^{-1}\right\}$

2.3 Expectation-Maximization Algorithm

In their influential paper, Dempster et.al (1977) proposed an intuitive and insightful method to estimate the latent variables model. The basic idea of E-M algorithm is that if the latent variables were known, the complete data likelihood would be much easier to maximize, usually with close-form solution. To address the unknown, we take the expectation of complete data likelihood w.r.t. the posterior distribution of latent variables (conditional on our latest estimated parameters). E-M algorithm can be viewed as lower bound optimization, and each round of E-step and M-step will lead to an updated estimator. The beauty of the approach is that the likelihood will never decrease in iteration and will converge to the maxima under fairly weak regularity conditions.

Huang et.al (1987) applied E-M algorithm to SUR Tobit model. They elucidated E-steps and M-steps analytically. However, the first and second moments of truncated MVN are required in E-step. Birnbaum

and Meyer(1953) , Tallis(1960) did provide the formulae, which, unfortunately, involve a series of cumbersome MVN c.d.f. As is known, the robust E-M algorithm comes with the price of slow convergence. It usually takes hundreds of iterations before convergence is achieved. If dozens of c.d.f. are calculated in each iteration, it would take very long time to implement the algorithm. In this paper, we extend Huang(1999)'s algorithm featuring Monte Carlo E-step plus conditional M-step. In our model, there are two latent variables— $\mathbf{Y_t^*}$ and τ_t. Direct Monte Carlo E-step is not applicable, so we propose the MCMC to carry out the E-step.

The details of the algorithm are as follows:

The complete data log likelihood can be written as

$$\begin{aligned}
&\ln p(\mathbf{Y}, \mathbf{Y}^*, \boldsymbol{\tau} | \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu) \\
&= \sum_{t=1}^{T} [\ln p(\tau_t|\nu) + \ln p(\mathbf{Y_t^*}|\tau_t, \boldsymbol{\beta}, \boldsymbol{\Sigma}) + \ln p(\mathbf{Y_t})] \\
&= \sum_{t=1}^{T} \ln p(\tau_t|\nu) + \ln p(\mathbf{Y_t^*}|\tau_t, \boldsymbol{\beta}, \boldsymbol{\Sigma}) + \text{constant} \\
&= \sum_{t=1}^{T} \begin{bmatrix} \ln \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) - \left(\frac{\nu}{2}+1\right)\ln\tau_t - \frac{\nu}{2}\tau_t^{-1} \\ -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\tau_t^{-1}(\mathbf{Y_t^*} - \mathbf{x_t}\,\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y_t^*} - \mathbf{x_t}\,\boldsymbol{\beta}) \end{bmatrix} + \text{constant}
\end{aligned}$$

We use the informal notation "constant" to represent those additive terms which do not interact with model parameters $\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu$, The rational is that our goal is to maximize the expected likelihood w.r.t. $\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu$ where terms labeled as "constant" play no role. Note that some stand-alone functions of latent variables, such as $-\frac{m}{2}\ln\tau_t$, can also be dumped.

To calculate the expectation of above likelihood, we first need to find out the posterior distribution of latent variables, i.e. $\mathbf{Y}^*, \boldsymbol{\tau} \mid \mathbf{Y}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \nu^{(0)}$, where $\left(\boldsymbol{\beta}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \nu^{(0)}\right)$ stands for estimators produced by latest round of iteration. (E-step must be implemented in that fashion, or the likelihood cannot be guaranteed to increase).

Unfortunately, the (joint) posterior distribution of $\mathbf{Y}^*, \boldsymbol{\tau}$ do not belong to any recognizable family. That is to say, the posterior mean $E(\ln\tau_t)$, $E(\tau_t^{-1})$, $E\left[\tau_t^{-1}(\mathbf{Y_t^*} - \mathbf{x_t}\,\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y_t^*} - \mathbf{x_t}\,\boldsymbol{\beta})\right]$ do not have a close form expression, let alone analytical expected likelihood. Simulations are needed to perform E-step.

Fortunately, the posterior conditional distributions do have a recognized form, so that Gibbs sampler can be employed to obtain posterior draws, which could be used to simulate expected likelihood.

For the time being (only in Section 2.3), define $\boldsymbol{\beta}^{(0)} = \begin{pmatrix} \boldsymbol{\beta_u} \\ \boldsymbol{\beta_c} \end{pmatrix}$, $\boldsymbol{\Sigma}^{(0)} = \begin{pmatrix} \boldsymbol{\Sigma_{uu}} & \boldsymbol{\Sigma_{uc}} \\ \boldsymbol{\Sigma_{cu}} & \boldsymbol{\Sigma_{cc}} \end{pmatrix}$, other notations are same as Section 2.1

$$\begin{aligned}
&\mathbf{Y_{ut}^*} \equiv \mathbf{Y_{ut}} \\
&\mathbf{Y_{ct}^*}\left|\tau_t, \mathbf{Y_{ut}^*}, \mathbf{Y_t}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \nu^{(0)} \right. \sim \mathrm{MVN}_{(-\infty,\mathbf{0})}\left[\mathbf{X_{ct}}\boldsymbol{\beta_c} + \boldsymbol{\Sigma_{cu}}\boldsymbol{\Sigma_{uu}^{-1}}(\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\boldsymbol{\beta_u}) \ , \ \tau_t\left(\boldsymbol{\Sigma_{cc}} - \boldsymbol{\Sigma_{cu}}\boldsymbol{\Sigma_{uu}^{-1}}\boldsymbol{\Sigma_{uc}}\right)\right] \\
&\tau_t\left|\mathbf{Y_t^*}, \mathbf{Y_t}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \nu^{(0)} \right. \sim \mathrm{IG}\left\{\frac{\nu^{(0)}+m}{2}, \left[\frac{\nu^{(0)}}{2} + \frac{1}{2}\left(\mathbf{Y_t^*} - \mathbf{x_t}\,\boldsymbol{\beta}^{(0)}\right)'\left(\boldsymbol{\Sigma}^{(0)}\right)^{-1}\left(\mathbf{Y_t^*} - \mathbf{x_t}\,\boldsymbol{\beta}^{(0)}\right)\right]^{-1}\right\}
\end{aligned}$$

Note that the posterior conditionals of $\mathbf{Y_{ct}^*}$ is truncated MVN. Geweke(1991) discussed sampling from univariate as well as multivariate truncated normal distribution. For truncated MVN, the distribution of one element conditional on other elements is univariate truncated normal, so that draws can be obtained via Gibbs sampler. In our case, an additional step, namely $\tau_t|\cdot$, will be added to the Gibbs cycle to obtain posterior draws $\left\{\mathbf{Y_{t\,(r)}^*}, \tau_{t\,(r)}\right\}_{r=1}^{R}$, $t = 1, \ldots, T$.

The expected complete data likelihood then can be approximated by the ergodic mean:

$E[\ln p(\mathbf{Y}, \mathbf{Y}^*, \boldsymbol{\tau}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)]$

$$\approx \frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T}\begin{bmatrix} -\ln\Gamma\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\ln\left(\frac{\nu}{2}\right)-\left(\frac{\nu}{2}+1\right)\ln\tau_{t\,(r)}-\frac{\nu}{2}\tau_{t\,(r)}^{-1} \\ -\frac{m}{2}\ln\tau_{t\,(r)}-\frac{1}{2}\ln|\boldsymbol{\Sigma}|-\frac{1}{2}\tau_{t\,(r)}^{-1}\left(\mathbf{Y}_{\mathbf{t}\,(\mathbf{r})}^{*}-\mathbf{x_t}\,\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_{\mathbf{t}\,(\mathbf{r})}^{*}-\mathbf{x_t}\,\boldsymbol{\beta}\right) \end{bmatrix}+\text{constant}$$

Now that E-step has been done, M-step is on the way.

Note that even in SUR model, the likelihood function cannot be optimized analytically w.r.t. $\boldsymbol{\beta},\boldsymbol{\Sigma}$ all at once. Here we are in the same position. Only if we maximize $\boldsymbol{\beta}$ conditional on $\boldsymbol{\Sigma}$, ν or optimize $\boldsymbol{\Sigma}$ conditional on $\boldsymbol{\beta},\boldsymbol{\nu}$ can we obtain close-form solutions. The conditional M-step still ensures the non-decreasing of likelihood as well as convergence. It is easy to show the updated estimators are: (See Appendix A3)

$$\boldsymbol{\beta}^{(1)}=\left[\sum_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\right)\mathbf{x}_t'\cdot\boldsymbol{\Sigma}^{-1}\cdot\mathbf{x_t}\right]^{-1}\cdot\sum_{t=1}^{T}\left[\mathbf{x}_t'\cdot\boldsymbol{\Sigma}^{-1}\cdot\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\,\mathbf{Y}_{\mathbf{t}\,(\mathbf{r})}^{*}\right)\right]$$

$$\boldsymbol{\Sigma}^{(1)}=\frac{1}{T\,R}\sum_{r=1}^{R}\sum_{t=1}^{T}\left[\tau_{t\,(r)}^{-1}\left(\mathbf{Y}_{\mathbf{t}\,(\mathbf{r})}^{*}-\mathbf{x_t}\,\boldsymbol{\beta}\right)\left(\mathbf{Y}_{\mathbf{t}\,(\mathbf{r})}^{*}-\mathbf{x_t}\,\boldsymbol{\beta}\right)'\right]$$

$$\nu^{(1)}=\arg\max\sum_{t=1}^{T}\left[-\ln\Gamma\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\ln\left(\frac{\nu}{2}\right)-\left(\frac{\nu}{2}+1\right)\left(\frac{1}{R}\sum_{r=1}^{R}\ln\tau_{t\,(r)}\right)-\frac{\nu}{2}\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\right)\right]$$

Intuitively, $\boldsymbol{\beta}^{(1)}$ is a weighted OLS estimator by disturbance precisions, $\boldsymbol{\Sigma}^{(1)}$ is the average of residual covariance matrix across equations. $\nu^{(1)}$ does not have a close-form, but the scalar can be easily optimize by Newton type numerical procedure, or even brute-force grid search!

2.4 Bayesian Posterior Simulator

E-M algorithm lies somewhere between the ML and Bayesian approach. On the one hand, E-M focus on the maximization of likelihood function and indeed obtain point estimators of $\boldsymbol{\beta},\boldsymbol{\Sigma},\nu$. On the other hand, E-M also yields, as a by-product, the posterior distribution of the latent variables $\mathbf{Y}^{*},\boldsymbol{\tau}$.

Bayesian approach treats the latent variable in the same manner as the parameters, and aims at joint posteriors of $\boldsymbol{\beta},\boldsymbol{\Sigma},\nu,\mathbf{Y}^{*},\boldsymbol{\tau}$. Bayesian posterior simulator addresses the latent variables in the most intuitive way—simply take a random draw from its posterior distribution so as to augment the data.

Chib and Greenberg (1995, 1998) examined limited dependent variable model in the Bayesian SUR framework. Huang (2001) outlines the full set of conditionals for Gibbs sampler of SUR Tobit model. In our model with Gaussian scale-mixtures, only two additional steps will be added to the sampler—one is to take posterior draws for latent scale τ_t, the other is to sample degree of freedom parameter ν. The latter do not have a recognizable distribution, and therefore Metropolis-within-Gibbs procedure is employed to carry out MCMC. Compared with E-M algorithm outlined in the previous section, the posterior conditionals of latent variables $\mathbf{Y}^{*},\boldsymbol{\tau}$ are almost identical, but we need to specify the suitable priors to derive the posteriors of $\boldsymbol{\beta},\boldsymbol{\Sigma},\nu$ as well.

The proper priors are assumed to be:

$\boldsymbol{\beta}\sim\text{MVN}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}},\mathbf{V}_{\boldsymbol{\beta}}\right)$

$\boldsymbol{\Sigma}^{-1}\sim\text{Wishart}\left(\boldsymbol{\Omega},k\right)$

$\nu\sim\text{Uniform}\left(\underline{\nu},\overline{\nu}\right)$

The full set of posterior conditionals are: (See Appendix A4)

$\boldsymbol{\beta}|\,\mathbf{Y},\boldsymbol{\Sigma},\nu,\mathbf{Y}^{*},\boldsymbol{\tau}\sim\text{MVN}(\mathbf{Dd},\mathbf{D})$

$$\text{where } \mathbf{D} = \left[\sum_{t=1}^{T} \tau_t^{-1} \cdot \mathbf{x}_t' \, \boldsymbol{\Sigma}^{-1} \, \mathbf{x}_t + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\right]^{-1}$$

$$\mathbf{d} = \left[\sum_{t=1}^{T} \tau_t^{-1} \cdot \mathbf{x}_t' \, \boldsymbol{\Sigma}^{-1} \, \mathbf{Y}_t^* + \mathbf{V}_{\boldsymbol{\beta}}^{-1} \, \boldsymbol{\mu}_{\boldsymbol{\beta}}\right]$$

$$\boldsymbol{\Sigma}^{-1} | \, \mathbf{Y}, \boldsymbol{\beta}, \nu, \mathbf{Y}^*, \boldsymbol{\tau} \sim \text{Wishart} \left\{ \left[\boldsymbol{\Omega}^{-1} + \sum_{t=1}^{T} \tau_t^{-1} \cdot (\mathbf{Y}_t^* - \mathbf{x}_t \boldsymbol{\beta})(\mathbf{Y}_t^* - \mathbf{x}_t \boldsymbol{\beta})'\right]^{-1}, \; T + k \right\}$$

$$\tau_t | \, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu, \mathbf{Y}^* \sim \text{IG} \left\{ \frac{\nu + m}{2}, \left[\frac{\nu}{2} + \frac{1}{2} (\mathbf{Y}_t^* - \mathbf{x}_t \boldsymbol{\beta})' \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{Y}_t^* - \mathbf{x}_t \boldsymbol{\beta})\right]^{-1} \right\}$$

$$\mathbf{Y}_{ct}^* | \, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu, \boldsymbol{\tau} \sim \text{MVN}_{(-\infty, \mathbf{0})} \left[\mathbf{X}_{ct} \boldsymbol{\beta}_c + \boldsymbol{\Sigma}_{cu} \boldsymbol{\Sigma}_{uu}^{-1} (\mathbf{Y}_{ut}^* - \mathbf{X}_{ut} \boldsymbol{\beta}_u) \; , \; \tau_t \left(\boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{cu} \boldsymbol{\Sigma}_{uu}^{-1} \boldsymbol{\Sigma}_{uc}\right)\right]$$

$$p(\nu | \, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu, \mathbf{Y}^*, \boldsymbol{\tau}) \propto \prod_{t=1}^{T} \left[\Gamma\left(\frac{\nu}{2}\right) \cdot \left(\frac{2}{\nu}\right)^{\nu/2}\right]^{-1} (\tau_t)^{-\left(\frac{\nu}{2}+1\right)} e^{-\frac{\nu}{2}\tau_t^{-1}} \cdot I\left(\underline{\nu} < \nu < \overline{\nu}\right)$$

All posterior conditionals come from easy-to-draw distributions except for $\nu | \cdot$, which does not have a recognizable form. Generally speaking, there are three ways to handle the case: one is to discretize the degree of freedom parameter (Albert and Chib, 1993). However, experience tells us that Gibbs sampler might perform poorly in the presence of discrete variates, since the chain usually move slowly across the parameter space. It is difficult to take a sudden leap to another discrete regime. Another approach is direct rejection sampling, but the supremum of the source and target density might be difficult to calculate in this case. Finally, the preferred approach is Metropolis-Hasting sampler within Gibbs. (In fact, Gibbs is a special case of M-H with acceptance probability of one). Here we use the normal random walk proposal, which performs satisfactorily.

3. Comparison of Estimators

In this section, we will apply the three estimation methods, namely Maximum simulated likelihood (ML hereafter), Expectation Maximization (EM hereafter), and Bayesian posterior simulators (BAY hereafter), to simulated data. The purpose of generated data experiment is twofold. First and foremost, despite the fact that they all yield consistent estimators, their finite sample properties are unknown (not applicable to Bayesian estimators, they are all finite sample in nature). Through generated data Monte Carlo experiments, we can compare their relative performance for a given sample size. Secondly, SUR Tobit model is easier said than done — practically computation poses the greatest challenge. Through the experiment, we attempt to find out the relatively economical way to estimate the model.

We will conduct two types of experiments. One is repeated experiments with fixed sample size of generated data, from which we can study the finite sample relative efficiency of varied estimators without resorting to asymptotics. The other is experiments with increasing sample size, so that we can learn the sensitivity of sample size to the performance of differed estimation methods.

Codes are run on an ordinary desktop computer (Pentium Dual 2.5GHZ / 3GB RAM / 64 bit Vista / Matlab 2009a). Computation time is recorded for reference purpose only, which obviously have much to do with the quality of the codes and the ability of the programmer.

As for the criteria of comparison, basically we compare the MSE of varied estimators. Since the researcher sets the DGP, the true parameter vector $\boldsymbol{\theta}$ is known. Using one set of generated data, we obtain estimators $\widehat{\boldsymbol{\theta}}_{\text{ML}}$, $\widehat{\boldsymbol{\theta}}_{\text{EM}}$, and $\widehat{\boldsymbol{\theta}}_{\text{Bay}}$ (posterior mean, by default). Repeat the process R times, we obtain $\left\{\widehat{\boldsymbol{\theta}}_{\text{ML (r)}}\right\}_{r=1}^{R}$, $\left\{\widehat{\boldsymbol{\theta}}_{\text{EM (r)}}\right\}_{r=1}^{R}$, $\left\{\widehat{\boldsymbol{\theta}}_{\text{Bay (r)}}\right\}_{r=1}^{R}$.

We first compare and report the regression coefficients one by one. Let θ , $\hat{\theta}_{\mathrm{ML}}$, $\hat{\theta}_{\mathrm{EM}}$, and $\hat{\theta}_{\mathrm{Bay}}$ be a corresponding element of the parameter vector $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{\mathbf{ML}}$, $\widehat{\boldsymbol{\theta}}_{\mathbf{EM}}$, $\widehat{\boldsymbol{\theta}}_{\mathbf{Bay}}$. We report the following:

Sample mean of estimators: $\widehat{\mathrm{E}}\left(\hat{\theta}_{\mathrm{ML}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}} \hat{\theta}_{\mathrm{ML\,(r)}}$

Standard deviation of estimators: $\hat{\mathrm{s}}\left(\hat{\theta}_{\mathrm{ML}}\right) = \sqrt{\frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left[\hat{\theta}_{\mathrm{ML\,(r)}} - \widehat{\mathrm{E}}\left(\hat{\theta}_{\mathrm{ML}}\right)\right]^2}$

MSE of estimators: $\mathrm{MSE}\left(\hat{\theta}_{\mathrm{ML}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\hat{\theta}_{\mathrm{ML\,(r)}} - \theta\right)^2$

E-M and Bayesian counterparts are defined in the same manner.

In addition, we also report some "composite indicators". First of all, MSE matrixes are calculated:

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{ML\,(r)}} - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}}_{\mathbf{ML\,(r)}} - \boldsymbol{\theta}\right)'$$

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{EM}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{EM\,(r)}} - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}}_{\mathbf{EM\,(r)}} - \boldsymbol{\theta}\right)'$$

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{Bay}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{Bay\,(r)}} - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}}_{\mathbf{Bay\,(r)}} - \boldsymbol{\theta}\right)'$$

If, for example, $\mathrm{MSE}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}}\right) - \mathrm{MSE}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{EM}}\right)$ is positive definite, then we conclude $\widehat{\boldsymbol{\theta}}_{\mathbf{EM}}$ has a smaller MSE, thus is preferred to $\widehat{\boldsymbol{\theta}}_{\mathbf{ML}}$

Since it is possible that the difference of MSE matrix is neither positive definite nor negative definite, (i.e. some of the eigenvalues are positive and others are negative, hence no conclusion), we report a scalar indicator "quadratic risk" as well, which always enables a ranking.

Quadratic risk:

$$\rho\left(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{ML\,(r)}} - \boldsymbol{\theta}\right)'\left(\widehat{\boldsymbol{\theta}}_{\mathbf{ML\,(r)}} - \boldsymbol{\theta}\right)$$

$$\rho\left(\widehat{\boldsymbol{\theta}}_{\mathbf{EM}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{EM\,(r)}} - \boldsymbol{\theta}\right)'\left(\widehat{\boldsymbol{\theta}}_{\mathbf{EM\,(r)}} - \boldsymbol{\theta}\right)$$

$$\rho\left(\widehat{\boldsymbol{\theta}}_{\mathbf{Bay}}\right) = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\widehat{\boldsymbol{\theta}}_{\mathbf{Bay\,(r)}} - \boldsymbol{\theta}\right)'\left(\widehat{\boldsymbol{\theta}}_{\mathbf{Bay\,(r)}} - \boldsymbol{\theta}\right)$$

Estimator with smaller quadratic risk is preferable.

3.1 Repeated experiments with fixed sample size

The data generating process are as follows:

$\mathrm{Y}_{1\mathrm{t}}^{*} = \beta_{10} + \mathrm{x}_{11\mathrm{t}}\,\beta_{11} + \mathrm{x}_{12\mathrm{t}}\,\beta_{12} + \varepsilon_{1\mathrm{t}}$, $\mathrm{Y}_{1\mathrm{t}} = \max(0, \mathrm{Y}_{1\mathrm{t}}^{*})$

$\mathrm{Y}_{2\mathrm{t}}^{*} = \beta_{20} + \mathrm{x}_{21\mathrm{t}}\,\beta_{21} + \mathrm{x}_{22\mathrm{t}}\,\beta_{22} + \varepsilon_{2\mathrm{t}}$, $\mathrm{Y}_{2\mathrm{t}} = \max(0, \mathrm{Y}_{2\mathrm{t}}^{*})$

$\mathrm{Y}_{3\mathrm{t}}^{*} = \beta_{30} + \mathrm{x}_{31\mathrm{t}}\,\beta_{31} + \mathrm{x}_{32\mathrm{t}}\,\beta_{32} + \varepsilon_{3\mathrm{t}}$, $\mathrm{Y}_{3\mathrm{t}} = \max(0, \mathrm{Y}_{3\mathrm{t}}^{*})$

$\mathrm{Y}_{4\mathrm{t}}^{*} = \beta_{40} + \mathrm{x}_{41\mathrm{t}}\,\beta_{41} + \mathrm{x}_{42\mathrm{t}}\,\beta_{42} + \varepsilon_{4\mathrm{t}}$, $\mathrm{Y}_{4\mathrm{t}} = \max(0, \mathrm{Y}_{4\mathrm{t}}^{*})$

The true parameter values are set to be

$\beta_{10} = \beta_{11} = \beta_{12} = 1$, $\beta_{20} = \beta_{21} = \beta_{22} = 2$, $\beta_{30} = \beta_{31} = \beta_{32} = 3$, $\beta_{40} = \beta_{41} = \beta_{42} = 4$

$$\begin{pmatrix}\varepsilon_{1t}\\ \varepsilon_{2t}\\ \varepsilon_{3t}\\ \varepsilon_{4t}\end{pmatrix}|\tau_t \sim \mathrm{MVN}\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \tau_t\begin{pmatrix}1 & 0.3 & 0.3 & 0.3\\ 0.3 & 1 & 0.3 & 0.3\\ 0.3 & 0.3 & 1 & 0.3\\ 0.3 & 0.3 & 0.3 & 1\end{pmatrix}\right],\ \tau_t \sim \mathrm{IG}\left(\frac{6}{2}, \frac{2}{6}\right),$$

Regressors are drawn from standard normal.

The sample size is fixed at $T = 2000$. For a given sample, three estimation methods are applied to the same data. Then generated data experiments are repeated for $R = 100$ times (Due to limits of computation resources, repetition times are not very large. However, statistics obtained from each experiment are strictly iid, so it is still quite informative).

As is seen from Table 3.1.1, more than 68% of the data are subject to censoring, to some varied degree from one variable (40% of the data) to four variables (1% of the data). In the presence of non-negligible portion of censored data, use OLS or GLS to estimate the SUR system will cause inconsistency. Tobit-type approach must be used to explicitly account for censoring. This is confirmed by Table 3.3.2, which shows the native regression on the basis of equation-wise OLS and GLS. The results suggest that the estimators are severely biased downwards and the MSEs are huge.

Table 3.1.3 to 2.2.5 compares the three consistent and asymptotically efficient estimators, namely ML, EM and Bayesian approach, in the context of given sample size T=2000. The major finding of the experiments is that Bayesian posterior simulators outperform any other estimators, unparalleled in terms of accuracy, stability and speed.

The accuracy of estimators can be compared by MSE of individual coefficients and the composite indictor such as quadratic risk. Though the mean of ML, EM and Bayesian estimators are all reasonably close to the true parameters, they differ substantially in MSE. The EM algorithm has the largest MSE, followed by ML estimators. Bayesian simulators has the least MSE in almost every individual coefficient. It is not because the EM algorithm itself is less accurate; the lack of preciseness is most likely induced by MCMC in the E-step. We only use 5000 draws to simulate the expected complete data likelihood, which could be less precise since MCMC draws are not independent. However, to choose 5000 draws is largely a practical concern—the fact is that every iteration entails such a simulation and EM algorithm usually takes hundreds of iterations to achieve convergence. It would be too costly to take larger number of draws.

A hidden factor affecting accuracy of estimators is starting values. For ML and EM, we use single equation Tobit with normal disturbances as the starting values, which are already reasonably close to the true parameter (but not consistent due to MVT errors). If we had picked the starting values at random, it could have been a disaster for ML and EM estimators. However, for Bayesian approach, we indeed picked the starting value at random, simply drawing from the prior, which can be far away from the true parameters. For example, degree of freedom is initially drawn from $\mathrm{U}(1, 20)$. The beauty of Gibbs sampler is that convergence is guaranteed despite the starting value under regularity conditions of Markov Chain. Indeed people use dispersed starting value to test the convergence of the chain. From this perspective, the Bayesian simulator is superior. In our experiments, the number of draws is 30000 with the first 10% draws burned in. The posterior mean of simulators are quite close in every experiments—an informal evidence suggesting convergence of the Markov Chain.

Another aspect of comparison is stability of estimators, which can be judged by the standard deviation of estimators. Though in each sample, asymptotical standard errors can be computed by inverted Hessian or BHHH methods, we did not report that since they are asymptotics which could be imprecise or even misleading in finite sample. As a matter of fact, for repeated experiments, we do not need to resort to

asymptotic distribution simply because we set the DGP by ourselves. By repeating the experiments independently for many times, we simply calculate the sample standard deviation and thus know the variability of estimators. As is seen in Table 3.1.3, the standard deviation of Bayesian simulators is smallest for every individual parameter. Again, the EM estimators have the largest standard errors. But the silver lining is that the difference between EM estimators and the truth are always within two standard errors (divided by $\sqrt{R}$), roughly indicating that the 95% credible intervals will cover the true parameters. In this sense, EM estimators are still a fairly good one.

The last but not the least aspect of comparison is speed. Before the discussion, we reiterate that the speed have much to do with the quality of the codes, the software platform and the ability of the programmer. Best efforts have been made by the author to make the code run faster. As is seen from Table 3.1.5, the average computation time for each sample (of size 2000) is 187 seconds for Bayesian simulators, 817 seconds for ML and 1512 seconds for EM. The speed advantage of Bayesian approach is apparent for high dimensional SUR systems. Also note that the speed of Bayesian simulator is predictable and controllable since it is perfectly proportional to the number of draws. That is not the case in ML, because we have no exact idea how many iterations are needed to achieve convergence, nor how many function evaluations in the line search during the iteration.

We intentionally test the codes in Matlab since she has exceedingly good debugging tools and speed profiler. Latest version of Matlab has remarkable improvement in executing codes in the loop structure, and our coded in Fortran ".mex" files do not substantially increase the speed. Furthermore, we circumvent the slow Matlab optimization toolbox and use our own BFGS codes in ML estimation. Our test shows that the bottleneck of the ML is millions of the normal c.d.f. evaluation in the adapted GHK simulator, which accounts for about 80% of total CPU time. The bottle neck of EM algorithm turns out to be the truncated normal draws in the Monte-Carlo E step (50% of the CPU time). We used Geweke (1991)'s mixed source reject sampling to obtain the truncated normal and multivariate normal draws. The algorithm is indeed much faster than the inversion of normal c.d.f. But clearly millions or even billions of draws still requires substantial computation. Computation of Bayesian approach is light and CPU time is roughly evenly distributed in draws of truncated normal, Wishart matrix as well as inverted gamma draws. The only unfavorable aspect of Bayesian Gibbs sampler is that the memory usage is large. But nowadays RAMs are measured in GB and their I/O speed is several GBs per seconds. Therefore, high memory consumption becomes less relevant and matrix-oriented computation may accelerate the codes, especially ideal for the multi-core CPU.

Table 3.1.1 Censoring regime of generated data

number censored	censoring regime	censoring ratio
0	(+,+,+,+)	31.25%
1	(0,+,+,+)	11.45%
1	(+,0,+,+)	9.84%
1	(+,+,0,+)	9.49%
1	(+,+,+,0)	9.60%
2	(0,0,+,+)	4.37%

2	(0,+,0,+)	4.07%
2	(0,+,+,0)	3.86%
2	(+,0,0,+)	3.24%
2	(+,0,+,0)	3.26%
2	(+,+,0,0)	3.08%
3	(0,0,0,+)	1.71%
3	(0,0,+,0)	1.61%
3	(0,+,0,0)	1.42%
3	(+,0,0,0)	1.12%
4	(0,0,0,0)	0.66%

Note:
"+" means non-censor,
"0" means censor to zero

Table 3.1.2 Inconsistency of OLS and GLS

	TRUE	OLS	OLS_STD	OLS_MSE	GLS	GLS_STD	GLS_MSE
β_{11}	1	1.345	0.022	0.120	1.345	0.022	0.120
β_{12}	1	0.714	0.024	0.082	0.714	0.023	0.082
β_{13}	1	0.707	0.022	0.086	0.708	0.022	0.086
β_{21}	2	2.478	0.032	0.230	2.478	0.032	0.230
β_{22}	2	1.488	0.029	0.263	1.487	0.028	0.264
β_{23}	2	1.481	0.034	0.271	1.481	0.034	0.270
β_{31}	3	3.652	0.033	0.426	3.652	0.033	0.426
β_{32}	3	2.254	0.039	0.558	2.255	0.040	0.557
β_{33}	3	2.250	0.039	0.565	2.250	0.039	0.565
β_{41}	4	4.834	0.039	0.697	4.834	0.039	0.697
β_{42}	4	3.017	0.052	0.969	3.018	0.052	0.968
β_{43}	4	3.017	0.054	0.970	3.018	0.053	0.967
Σ_{11}	1	0.980	0.047	0.003	0.980	0.047	0.003
Σ_{21}	0.3	0.235	0.038	0.006	0.236	0.038	0.006
Σ_{31}	0.3	0.237	0.038	0.005	0.237	0.038	0.005
Σ_{41}	0.3	0.232	0.042	0.006	0.233	0.042	0.006
Σ_{22}	1	1.441	0.062	0.198	1.441	0.062	0.198
Σ_{32}	0.3	0.249	0.048	0.005	0.249	0.048	0.005
Σ_{42}	0.3	0.248	0.052	0.005	0.248	0.052	0.005
Σ_{33}	1	2.087	0.085	1.189	2.087	0.085	1.190
Σ_{43}	0.3	0.255	0.070	0.007	0.256	0.070	0.007
Σ_{44}	1	2.975	0.136	3.921	2.975	0.136	3.921

Note:
Estimators (OLS, GLS) are the sample mean of estimators of 100 experiments.
Standard errors are the sample standard deviations, MSEs are also sample analog.

Table 3.1.3 comparison of ML, EM and Bayesian, individual estimators

		ML			EM			BAYESIAN		
	TRUE	MEAN	STD	MSE	MEAN	STD	MSE	MEAN	STD	MSE
β 11	1	1.000	0.034	0.001	1.006	0.047	0.002	1.001	0.030	0.001
β 12	1	0.992	0.033	0.001	1.003	0.059	0.003	1.004	0.029	0.001
β 13	1	0.987	0.029	0.001	0.999	0.055	0.003	0.998	0.026	0.001
β 21	2	1.993	0.041	0.002	2.012	0.094	0.009	2.000	0.031	0.001
β 22	2	1.997	0.030	0.001	1.993	0.107	0.011	2.002	0.027	0.001
β 23	2	1.994	0.031	0.001	1.984	0.095	0.009	1.997	0.030	0.001
β 31	3	2.993	0.032	0.001	2.985	0.145	0.021	2.999	0.029	0.001
β 32	3	3.002	0.035	0.001	3.011	0.167	0.028	2.998	0.028	0.001
β 33	3	3.000	0.033	0.001	3.005	0.191	0.036	2.996	0.027	0.001
β 41	4	3.998	0.038	0.001	3.984	0.159	0.025	3.992	0.032	0.001
β 42	4	4.018	0.041	0.002	3.993	0.267	0.071	4.003	0.032	0.001
β 43	4	4.019	0.033	0.001	3.966	0.257	0.066	4.005	0.029	0.001
Σ 11	1	0.949	0.051	0.005	1.016	0.048	0.003	0.989	0.041	0.002
Σ 21	0.3	0.292	0.039	0.002	0.300	0.030	0.001	0.294	0.029	0.001
Σ 31	0.3	0.301	0.031	0.001	0.303	0.028	0.001	0.295	0.026	0.001
Σ 41	0.3	0.280	0.041	0.002	0.300	0.033	0.001	0.293	0.031	0.001
Σ 22	1	0.959	0.056	0.005	1.031	0.046	0.003	0.994	0.043	0.002
Σ 32	0.3	0.301	0.041	0.002	0.304	0.032	0.001	0.298	0.030	0.001
Σ 42	0.3	0.273	0.040	0.002	0.301	0.038	0.001	0.296	0.029	0.001
Σ 33	1	0.971	0.058	0.004	1.068	0.064	0.009	0.992	0.039	0.002
Σ 43	0.3	0.232	0.064	0.009	0.303	0.043	0.002	0.297	0.032	0.001
Σ 44	1	0.974	0.064	0.005	1.113	0.130	0.030	0.990	0.041	0.002
DF	6	5.552	0.516	0.464	6.428	0.537	0.469	6.016	0.491	0.239

Note:

Estimators (ML,EM,BAY) are the sample mean of estimators of 100 experiments.

Standard errors are the sample standard deviations, MSEs are also sample analog.

Table 3.1.4 comparison of ML, EM and Bayesian, composite estimators

	ML	EM	BAY
Quadric risk of β	0.015	0.285	0.010
Quadric risk of Σ	0.036	0.051	0.012

MSE definiteness	MSE(ML β) – MSE(EM β) < 0 , others not comparable

Table 3.1.5 computation time of ML, EM and Bayesian approach (Seconds)

	ML	EM	BAY
Mean	817	1512	187
Max	1348	1554	190
Min	555	1477	185

Note: computation time is ONE experiment with 2000 observations

Intel Pentium Dual 2.5GHZ / 3GB RAM / 64 bit Vista / Matlab 2009a

3.2 Effects of sample size

As we know, if the data generating process is exactly the same as what is described in Section 3.1, then the ML, EM and BAY estimators should get closer and closer to the true parameters as the sample size goes to infinity. In practice, however, every data set is of finite size. Now we conduct a series of experiments with varied sample size, and our goal is to answer the question "How large is large?" for each estimation method.

The sample sizes are set to be $\mathrm{T} = 500, 1000, 2000, 5000, 10000, 20000$ respectively.

One limitation of this study is that it takes hours to estimate the model with 20000 observations. Therefore, it is infeasible to conduct repeated experiments for hundreds of times (on my computer). Since we only do it once, the estimators might be less stable due to sampling randomness. The statistics should be read with reference to asymptotical standard errors.

The major findings of those experiments are Bayesian approach performs satisfactorily even under relatively small sample and computation speed increases moderately as sample size grows, compared with ML and EM estimators.

As is seen in Table 3.2.1 – 3.2.4, in the case of $\mathrm{T} = 500$, ML and EM estimators exhibit apparent deviation from the true coefficients. Particularly, the degree of freedom parameter wanders far from $\nu = 6$. However, Bayesian estimator works much better—regression coefficients, covariance matrix and degree of freedom estimators are all reasonably close to the true parameters. The quadratic risks (the squared sum of estimator deviations) are 0.026 for ML, 0.131 for EM and merely 0.009 for Bayesian estimators.

As the sample size grows, the quality of ML and EM estimators improves steadily. When the sample size increases to 10000 and up, the two estimators become good enough to square with the Bayesian estimators. In the case of $\mathrm{T} = 20000$, the quadratic risks for the three estimators are 0.002, 0.005, 0.004 respectively. Comparing individual coefficients, our eyeball can barely detect any difference between the ML and Bayesian estimators, and the differences for EM estimator are somewhat visible but not significant at all.

To attain the same quality, the costs are of great disparity. To estimate the data set of 20000, it takes roughly 5.3 hours via ML, 4.5 hours via EM, but merely 45 minutes by Bayesian approach. The savings of Bayesian estimators are awesome. Furthermore, the speed of Gibbs sampler is almost perfectly predictable, so that the researcher can add a status bar when the codes are running, watching and planning ahead. However, for ML estimator, the researcher never knows how many iterations it will take, so he waits and wishes it could end sooner.

Table 3.2.1 ML estimators with varied sample size

	TRUE	500	1000	2000	5000	10000	20000
β11	1	1.030	0.952	0.983	1.021	0.978	1.013
		[0.050]	[0.039]	[0.027]	[0.017]	[0.012]	[0.009]
β12	1	1.085	1.062	0.983	0.981	1.004	1.003
		[0.053]	[0.036]	[0.026]	[0.016]	[0.012]	[0.009]
β13	1	0.963	0.990	1.001	0.969	0.999	1.006
		[0.050]	[0.035]	[0.027]	[0.016]	[0.012]	[0.009]
β21	2	2.102	2.082	2.054	1.984	2.021	2.009
		[0.057]	[0.038]	[0.028]	[0.019]	[0.013]	[0.009]
β22	2	1.949	1.962	2.039	2.033	1.992	1.997
		[0.058]	[0.039]	[0.026]	[0.019]	[0.013]	[0.009]
β23	2	2.033	2.008	1.973	2.004	1.986	2.004
		[0.058]	[0.035]	[0.026]	[0.019]	[0.012]	[0.009]
β31	3	3.135	2.926	3.042	3.009	2.989	3.019
		[0.053]	[0.046]	[0.029]	[0.018]	[0.013]	[0.009]
β32	3	3.002	3.001	2.981	2.969	2.996	2.976
		[0.054]	[0.046]	[0.030]	[0.018]	[0.013]	[0.009]
β33	3	3.010	3.103	3.021	2.990	2.987	2.991
		[0.050]	[0.047]	[0.029]	[0.018]	[0.013]	[0.009]
β41	4	3.969	3.968	4.008	4.015	4.013	4.013
		[0.059]	[0.047]	[0.028]	[0.018]	[0.013]	[0.009]
β42	4	4.154	3.922	4.131	4.004	4.003	3.983
		[0.062]	[0.050]	[0.030]	[0.018]	[0.013]	[0.009]
β43	4	4.045	4.007	4.046	3.995	3.980	3.993
		[0.054]	[0.048]	[0.029]	[0.018]	[0.013]	[0.009]
Σ11	1	0.809	0.987	0.984	0.998	1.041	1.018
		[0.074]	[0.066]	[0.047]	[0.030]	[0.022]	[0.015]
Σ21	0.3	0.276	0.375	0.307	0.310	0.298	0.298
		[0.053]	[0.042]	[0.030]	[0.020]	[0.014]	[0.010]
Σ31	0.3	0.231	0.304	0.321	0.339	0.300	0.299
		[0.051]	[0.047]	[0.030]	[0.019]	[0.014]	[0.010]
Σ41	0.3	0.267	0.336	0.195	0.317	0.304	0.293
		[0.054]	[0.050]	[0.029]	[0.019]	[0.014]	[0.010]
Σ22	1	0.869	0.962	0.973	1.050	1.007	1.003
		[0.078]	[0.064]	[0.045]	[0.031]	[0.020]	[0.014]
Σ32	0.3	0.273	0.321	0.329	0.374	0.299	0.297
		[0.054]	[0.047]	[0.031]	[0.021]	[0.014]	[0.010]

Σ42	0.3	0.171	0.269	0.296	0.308	0.286	0.286
		[0.055]	[0.048]	[0.028]	[0.020]	[0.014]	[0.010]
Σ33	1	0.849	1.113	0.989	1.018	1.004	1.001
		[0.080]	[0.082]	[0.045]	[0.030]	[0.020]	[0.014]
Σ43	0.3	0.280	0.147	0.227	0.288	0.296	0.311
		[0.057]	[0.055]	[0.029]	[0.019]	[0.014]	[0.010]
Σ44	1	0.980	1.172	0.874	0.912	0.998	0.993
		[0.094]	[0.084]	[0.040]	[0.027]	[0.020]	[0.014]
DF	6	5.346	5.323	5.265	5.254	6.194	6.215
		[0.807]	[0.638]	[0.382]	[0.243]	[0.210]	[0.148]

Note: (Asymptotical) Standard errors are reported in brackets.

Table 3.2.2 EM estimators with varied sample size

	TRUE	500	1000	2000	5000	10000	20000
β 11	1	0.938	0.986	0.883	1.008	0.980	1.027
		[0.057]	[0.041]	[0.030]	[0.017]	[0.012]	[0.009]
β 12	1	1.249	1.040	1.078	1.019	1.001	1.021
		[0.059]	[0.040]	[0.029]	[0.017]	[0.012]	[0.009]
β 13	1	1.041	1.027	1.072	0.950	0.965	0.992
		[0.057]	[0.040]	[0.029]	[0.016]	[0.012]	[0.009]
β 21	2	2.083	2.029	2.262	1.916	2.116	2.060
		[0.060]	[0.040]	[0.029]	[0.019]	[0.012]	[0.009]
β 22	2	1.880	1.967	2.010	2.154	1.917	1.956
		[0.060]	[0.041]	[0.027]	[0.019]	[0.012]	[0.009]
β 23	2	2.198	1.964	1.858	2.195	1.873	1.959
		[0.057]	[0.038]	[0.027]	[0.020]	[0.012]	[0.009]
β 31	3	2.932	3.134	2.901	3.132	3.026	2.979
		[0.062]	[0.041]	[0.033]	[0.018]	[0.013]	[0.009]
β 32	3	3.405	2.857	3.152	2.762	3.034	2.960
		[0.066]	[0.041]	[0.033]	[0.018]	[0.013]	[0.009]
β 33	3	3.179	2.880	3.419	2.921	2.883	3.037
		[0.063]	[0.042]	[0.032]	[0.018]	[0.013]	[0.009]
β 41	4	3.648	3.918	4.291	4.114	4.008	4.064
		[0.068]	[0.047]	[0.030]	[0.018]	[0.013]	[0.009]
β 42	4	4.578	3.840	3.952	3.789	4.083	3.952
		[0.072]	[0.048]	[0.030]	[0.017]	[0.013]	[0.009]
β 43	4	4.401	4.353	3.787	3.798	3.959	4.009
		[0.062]	[0.044]	[0.029]	[0.018]	[0.013]	[0.009]
Σ 11	1	1.002	1.072	1.113	1.026	1.046	1.022
		[0.092]	[0.070]	[0.053]	[0.030]	[0.022]	[0.015]
Σ 21	0.3	0.325	0.320	0.295	0.335	0.293	0.297

		[0.062]	[0.043]	[0.033]	[0.020]	[0.013]	[0.009]
Σ 31	0.3	0.257	0.259	0.340	0.305	0.299	0.301
		[0.068]	[0.043]	[0.035]	[0.019]	[0.014]	[0.010]
Σ 41	0.3	0.304	0.335	0.273	0.281	0.305	0.293
		[0.066]	[0.050]	[0.034]	[0.019]	[0.014]	[0.010]
Σ 22	1	0.987	1.016	1.047	1.117	0.994	0.995
		[0.083]	[0.065]	[0.047]	[0.033]	[0.020]	[0.014]
Σ 32	0.3	0.275	0.258	0.325	0.324	0.293	0.295
		[0.064]	[0.042]	[0.035]	[0.021]	[0.014]	[0.010]
Σ 42	0.3	0.256	0.318	0.361	0.330	0.282	0.286
		[0.067]	[0.047]	[0.031]	[0.020]	[0.014]	[0.010]
Σ 33	1	1.142	0.991	1.210	1.037	1.014	1.016
		[0.113]	[0.065]	[0.056]	[0.030]	[0.020]	[0.015]
Σ 43	0.3	0.387	0.296	0.333	0.313	0.297	0.312
		[0.078]	[0.049]	[0.035]	[0.019]	[0.014]	[0.010]
Σ 44	1	1.292	1.209	0.999	0.961	1.015	0.994
		[0.129]	[0.082]	[0.045]	[0.028]	[0.020]	[0.014]
DF	6	7.396	6.898	6.188	6.050	6.327	6.297
		[1.283]	[0.795]	[0.477]	[0.281]	[0.216]	[0.150]

Note: (Asymptotical) Standard errors are reported in brackets.

Table 3.2.3 Bayesian estimators with varied sample size

	TRUE	500	1000	2000	5000	10000	20000
β 11	1	1.015	0.957	1.001	1.015	0.977	1.012
		[0.052]	[0.040]	[0.028]	[0.017]	[0.012]	[0.009]
β 12	1	1.098	1.053	0.982	0.998	1.004	1.004
		[0.053]	[0.039]	[0.026]	[0.017]	[0.012]	[0.009]
β 13	1	0.979	1.018	1.008	0.989	0.999	1.006
		[0.052]	[0.039]	[0.027]	[0.016]	[0.012]	[0.009]
β 21	2	2.129	2.085	2.048	1.992	2.021	2.008
		[0.052]	[0.039]	[0.029]	[0.018]	[0.013]	[0.009]
β 22	2	1.926	1.960	2.023	2.040	1.992	1.997
		[0.055]	[0.038]	[0.028]	[0.018]	[0.013]	[0.009]
β 23	2	2.013	1.969	1.969	2.008	1.986	2.004
		[0.052]	[0.038]	[0.028]	[0.018]	[0.013]	[0.009]
β 31	3	3.113	2.923	3.057	3.020	2.989	3.018
		[0.056]	[0.042]	[0.029]	[0.018]	[0.013]	[0.009]
β 32	3	3.019	2.965	3.003	2.961	2.996	2.976
		[0.057]	[0.040]	[0.030]	[0.018]	[0.013]	[0.009]
β 33	3	2.997	3.050	3.028	2.986	2.987	2.991
		[0.057]	[0.042]	[0.029]	[0.018]	[0.013]	[0.009]

β_{41}	4	3.974	4.007	4.026	4.010	4.012	4.013
		[0.059]	[0.043]	[0.029]	[0.018]	[0.013]	[0.009]
β_{42}	4	4.121	3.939	4.068	3.989	4.003	3.983
		[0.061]	[0.044]	[0.029]	[0.018]	[0.013]	[0.009]
β_{43}	4	4.097	3.973	4.016	3.980	3.980	3.993
		[0.059]	[0.042]	[0.028]	[0.018]	[0.013]	[0.009]
Σ_{11}	1	0.896	1.058	0.996	1.011	1.043	1.021
		[0.084]	[0.071]	[0.047]	[0.030]	[0.021]	[0.015]
Σ_{21}	0.3	0.279	0.313	0.294	0.322	0.298	0.299
		[0.055]	[0.043]	[0.030]	[0.019]	[0.014]	[0.010]
Σ_{31}	0.3	0.213	0.261	0.315	0.305	0.300	0.299
		[0.054]	[0.043]	[0.031]	[0.020]	[0.014]	[0.010]
Σ_{41}	0.3	0.259	0.321	0.278	0.283	0.304	0.293
		[0.057]	[0.046]	[0.030]	[0.019]	[0.014]	[0.010]
Σ_{22}	1	0.871	0.981	1.018	1.038	1.008	1.005
		[0.083]	[0.061]	[0.046]	[0.030]	[0.020]	[0.015]
Σ_{32}	0.3	0.263	0.269	0.326	0.322	0.299	0.297
		[0.054]	[0.041]	[0.032]	[0.020]	[0.014]	[0.010]
Σ_{42}	0.3	0.235	0.287	0.340	0.323	0.286	0.287
		[0.056]	[0.043]	[0.031]	[0.019]	[0.014]	[0.010]
Σ_{33}	1	0.909	0.977	1.023	1.029	1.005	1.003
		[0.083]	[0.062]	[0.047]	[0.030]	[0.020]	[0.014]
Σ_{43}	0.3	0.290	0.275	0.322	0.320	0.297	0.312
		[0.060]	[0.044]	[0.031]	[0.019]	[0.014]	[0.010]
Σ_{44}	1	0.995	1.028	0.946	0.941	0.999	0.995
		[0.090]	[0.067]	[0.044]	[0.027]	[0.020]	[0.014]
DF	6	6.290	6.613	5.669	5.870	6.248	6.292
		[0.947]	[0.737]	[0.401]	[0.271]	[0.205]	[0.153]

Note: (Asymptotical) Standard errors are reported in brackets.

Table 3.2.4 Comparison of quadratic risk

	500	1000	2000	5000	10000	20000
ML	0.026	0.025	0.026	0.025	0.002	0.002
EM	0.131	0.047	0.024	0.012	0.007	0.005
BAY	0.009	0.018	0.006	0.001	0.003	0.004

Table 3.2.5 Comparison of computation time (Seconds)

	500	1000	2000	5000	10000	20000
ML	173	556	873	2243	9412	19176
EM	635	967	1600	3811	7905	16153
BAY	66	105	192	540	1122	2660

4. Application

4.1 The topic

In this section, we apply the SUR Tobit scale-mixture model to high frequency financial data, and will study the interaction between security price and trading volume.

In finance, there is dichotomy between the academia and the practitioners on technical analysis. Chartists forecast future stock price using various figures patterns, formulas and indicators, the inputs of which are essentially two: historical prices and trading volumes. One of the popular indicators is moving average (MA) crossings (i.e. filtering rules). When, for example, the short-run MA line cuts the long-run MA line from the bottom, it is a buying signal. In a highly simplified model, we will show the above filtering-rule trading strategy could admit a Vector Auto-Regression (VAR) representation consisting of price and volume.

However, the informational value of past prices and trading volumes is underappreciated in academia in general. Many of us are happy with the Efficient Market Hypothesis (Fama, 1965). The argument is that there is no free lunch in Wall Street since the current price has already contains all the historical information; technical analysis, therefore, is in vain. The implication for the econometric model is that the lagged regressors will have no explanatory power on the current price.

Time series analysis to the technical analysis, is something like astronomy to the astrology. Campbell et al. (1997) insightfully pointed out that despite the linguistic barriers, time series and technical analysis do study the same thing—the informational content of historical prices. Studies by Brock et al. (1992), Taylor and Allen (1992), Lo et.al (2000) provided some foundations of technical analysis from the perspective of statistics.

Basically what we do in this section is to run a VAR model, study the associated impulse response functions and test the significance of lagged explanatory variables. As we are taught in the textbook, “Vector auto-regressions are invariably estimated on the basis of conditional likelihood function rather than the full sample unconditional likelihood.” (Hamilton, 1994). The conditional ML turns out to be OLS equation by equation. However, when the model is applied to high frequency data recorded in seconds intervals, there are at least two complications:

1. There might be no transactions in a certain time interval, hence zero trading volume.

2. High frequency data are well-known to exhibit non-normality, typically with a fatter tail.

The first issue can be viewed as censored observation around zero. Since the censored data could constitute a non-negligible portion, the OLS will be apparently inconsistent.

Even if in the absence of the first issue, the presence of the second issue may render the OLS less efficient. The fat tail can be largely addressed by models with scale-mixture errors.

Once we are working on VAR model using the time series data, it is necessary to investigate the stationarity of the variables, which determine the specification of VAR (levels, 1st difference, or error

correction, etc). Since we are dealing with very high frequency data, it stands to reason to believe shocks within a minute come from idiosyncratic sources such as individuals or institutional investors. The micro-sourced shocks should not have impacts last forever. Therefore there is no good reason to worry about the unit root, let alone the cointegrating constraints. With this assumption, we can simply treat the reduced form VAR as a SUR model, and denote all the predetermined lagged variables $\mathbf{X}$, and all the coefficients $\boldsymbol{\beta}$, so that we can use all the micro-econometric methods to estimate the model.

All in all, the model we use is essentially the SUR Tobit system with scale-mixture errors, exactly as is described in Section 2.1

4.2 The simple model

To justify the use of VAR, we provide a highly simplified (maybe trivial) model of investors' behavior and determination of price and volumes.

The investors are assumed to be chartists who use filtering rule on the basis of MA lines to determine the quote and the amount to buy/sell. Since actual transactions do occur, agents must be heterogeneous (If agents' preference are identical, price could be pinned down when the market clears, but no transaction indeed. Another possibility is to introduce other type of investors such as risk-averse hedgers, fundamentalists, noise traders, etc. That leaves for future research.)

Let the historical price be $\{P_{t-j}\}_{j=1}^{\infty}$, the moving average chart of order q can be written as:

$$MA_q = \frac{1}{q}\sum_{j=1}^{q} P_{t-j}$$

In a call auction mechanism, an investor places her limit orders with desired price and quantity, which, by assumption, is determined by the differences of various MA lines. In its reduced form, this is equivalent to be determined by weighted average of MA lines.

If she is a potential buyer,

Desired buying price: $d_{B,i} = \sum_q \omega_{i,q} \cdot MA_q$

Desired amount: $Q_{B,i} = \sum_q \tilde{\omega}_{i,q} \cdot MA_q$

If she is a potential seller,

Desired selling price: $d_{S,i} = \sum_q \omega_{i,q} \cdot MA_q$

Desired amount: $Q_{S,i} = \sum_q \tilde{\omega}_{i,q} \cdot MA_q$

Where the weights $\omega_{i,q}$, $\tilde{\omega}_{i,q}$ reflects the individual's preference and judgment.

The price P_t that enables the largest number of orders to be executed is then chosen. The market clearing requires:

$\sum_i Q_{B,i} \cdot I(d_{B,i} > P_t) + \sum_i Q_{S,i} \cdot I(d_{S,i} < P_t) = 0$

The trading volume $Q_t = \sum_i Q_{B,i} \cdot I(d_{B,i} > P_t)$

Where $I(\cdot)$ is the indicator function that takes the value of 1 if the expression in bracket are true, and 0 otherwise.

Apparently, both the P_t and Q_t are linear functions of moving average lines, hence linear with respect to historical prices and trading volumes.

In econometrics, this implies a VAR representation of price and volumes

$$\mathbf{Y_t} = \boldsymbol{\Phi_1}\mathbf{Y_{t-1}} + \cdots + \boldsymbol{\Phi_p}\mathbf{Y_{t-p}} + \boldsymbol{\varepsilon_t}$$

Where $\mathbf{Y_t} = \begin{pmatrix} Q_t \\ P_t \end{pmatrix}$

The disturbance $\boldsymbol{\varepsilon_t}$ could be interpreted as the change of individual's taste, measure errors, etc.

4.3 Data and estimation

Issued in August 22, 2005, Baosteel warrants (a type of call options) was the first financial derivatives traded in Shanghai Stock Exchange, China. The options were spectacular in that its price featured extreme volatility and 5-10 times higher than what Black-Scholes formula implied. Academia, in general, could not provide convincing rationales why the options price wandered far away from the fundamentals—bubbles, they guess. It is believed that the participants are all speculative individuals who may sell the options minutes after purchase. At that time, maybe only the chartists took delight in talking about its price movement and trading strategies.

We use price and volume data of 30 consecutive trading days starting from March 23, 2006. Each day has four trading hours. The frequency of the data is 3 seconds. So we have the sample size of 144,000. In this high frequency data set, 21.94% of the volume data are recorded as zero, implying no transaction in that time interval.

Figure 4.3.1 Price sequence (candlesticks) and trading volume

Note: As the convention, candlestick chart are based on daily trading data.

To run a VAR regression, we need to choose a proper lag length, which can be determined by the Schwarz information criterion (SIC):

$$\mathrm{SIC} = -2\frac{\mathrm{logL}}{\mathrm{T}} + \frac{\mathrm{k\,ln(T)}}{\mathrm{T}}$$

Where $\log L$ is the log-likelihood function, k is number of regressors and T is number of observations.

Apparently, to perform SIC, one must run the regression for every lag specification, say, from 1 lag model to 100 lags model. That is too much computation for our model with censoring and scale-mixtures. So when we choose the lag structure, we neglect the two complications and use the equation-wise-OLS based conditional likelihood. This is not very correct, strictly speaking, but merely some expediency. The SIC suggests lag length of 27.

When we estimate the VAR with 27 lags, we do use the full information maximum likelihood and Bayesian approach, as is described in Section 2.1. Since the Monte Carlo experiments in Section 3 suggest poor performance of the slow EM algorithm, we do not use that approach here.

In this application, the model is a bivariate SUR Tobit. It is important to note that numerical algorithms are different for lower and higher dimensional problems. For instance, in lower dimension, Gaussian quadratures usually provides a much faster and more accurate approximation, while Monte Carlo methods are more suitable for higher dimensional problem because it suffers less from *the curse of dimension*.

For speed and accuracy, in this bivariate system, Gaussian quadratures are employed instead of direct Monte Carlo simulation whenever possible. In maximum likelihood, the probability of censored observation is essentially an integrand consisting of a univarite normal c.d.f. and an IG p.d.f.. If we took thousands of IG draws to directly simulate the probability, it would be too slow. We instead divide the integration range into subintervals and use Gauss-Laguerre and Gauss-Legendre quadratures accordingly. In that fashion, hundreds of weighed sum of univarite normal c.d.f. will provide a much faster and more accurate censoring probability.

Another feature of VAR model is the presence of a large amount of regressors. With 27 lags, there are 108 lagged explanatory variables (and 2 constants), plus covariance matrix and degree of freedom parameters. To accelerate the codes, it is also important to provide the analytic gradient in ML estimation. If we let the computer to calculate the numerical derivatives, it would require excessive calls to the objective function. Analytic gradients, however, requires little additional computer's efforts (but substantial human debugging efforts!), so that the codes could be ten times faster.

Estimation results are provided in Table 4.3.1. The left column is ML estimators and the right column is Bayesian simulators (posterior mean). Since large amount of regressors consume so much paper space, the standard errors are not reported for ecological reasons. (They are available upon request.) In fact, the covariance matrix of estimators is useful in that later we need it to calculate 95% intervals of impulse response functions. As we can see from the table, the ML and Bayesian approach produce very similar results. The coefficients corresponding to the first lag are relatively large, and decay with the lags. The estimated degree of freedom is very small, implying that the disturbance terms do deviate from normality to a large extend.

On interesting feature of VAR is that it can generate impulse response functions—the dynamic response of endogenous variables to shocks. However, impulse response analysis requires additional identification constraints, which must come from economic theory and thus difficult to provide in general. Fortunately, in our bivariate case, only one additional identification condition is needed. We impose a contemporaneous constraint, namely "volume comes first"—it takes one period for volume responding to price shocks. That is the case for chartists who only use historical price and volume information to determine how much they will buy today. Figure 4.3.2-4.3.5 depict the impulse response functions. ML and Bayesian approaches produce almost identical dynamic response patterns. The two red dashed lines are two boundaries of 95% confidence interval. For Bayesian simulators, it is straight forward to provide

the interval since we already have posterior draws. For each draw, we invert the VAR into $\mathrm{MA}(\infty)$ representation and the coefficients are response magnitudes. Repeat for all draws, the 2.5% and 97.5% percentiles are the two boundaries of the 95% intervals. As for the ML approach, we assume the normality of estimators (though the disturbances are scale mixtures, the ML estimators should be asymptotically normal anyway), so that we can use parametric bootstrap to obtain simulated draws of estimators, get $\mathrm{MA}(\infty)$ representations and depict the 95% intervals.

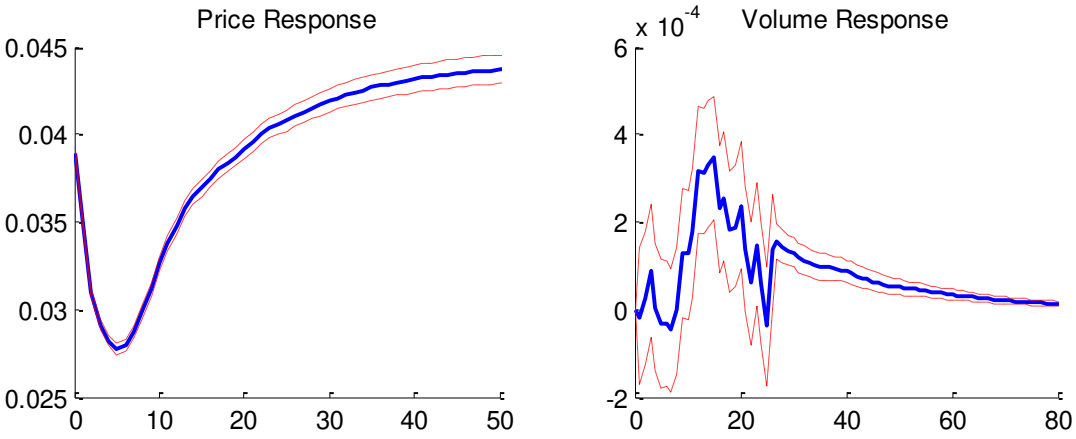
Finally, we test the information contents of historical price and trading volumes. The easiest way to test the null (namely efficient market hypothesis) is LR test which has χ^2 distribution asymptotically. Simply remove all the lagged regressors, as we expect, the log likelihood decreases drastically, thus the efficient market hypothesis is rejected in our model.

Table 4.3.1 Reduced form VAR estimation

	ML		BAY	
	Eq(1)	Eq(2)	Eq(1)	Eq(2)
Const	0.003	-0.007	0.003	-0.006
Φ (1)	0.894	0.003	0.887	0.003
	0.000	0.044	-0.001	0.054
Φ (2)	-0.005	0.005	-0.002	0.005
	0.001	0.034	0.000	0.035
Φ (3)	0.041	0.001	0.042	0.000
	0.002	0.028	0.001	0.028
Φ (4)	0.024	0.003	0.025	0.003
	-0.002	0.028	-0.002	0.028
Φ (5)	0.013	-0.001	0.014	-0.001
	-0.001	0.023	-0.001	0.023
Φ (6)	0.023	0.000	0.024	-0.001
	0.000	0.022	0.000	0.022
Φ (7)	0.020	0.001	0.020	0.001
	-0.001	0.016	-0.001	0.016
Φ (8)	0.017	0.003	0.017	0.003
	0.001	0.019	0.001	0.019
Φ (9)	0.015	0.000	0.016	0.000
	0.003	0.017	0.004	0.017
Φ (10)	0.010	-0.001	0.010	-0.001
	0.000	0.019	0.000	0.020
Φ (11)	0.000	-0.001	0.000	-0.001
	0.002	0.019	0.002	0.019
Φ (12)	-0.001	-0.003	-0.002	-0.003
	0.004	0.015	0.003	0.015
Φ (13)	0.000	0.000	0.000	0.000
	0.000	0.017	0.001	0.017
Φ (14)	-0.007	0.000	-0.007	0.000

	0.001	0.017	0.001	0.018
Φ (15)	-0.008	-0.001	-0.009	-0.001
	0.001	0.017	0.001	0.017
Φ (16)	-0.005	-0.001	-0.005	-0.001
	-0.003	0.012	-0.004	0.012
Φ (17)	-0.004	-0.001	-0.004	-0.001
	0.000	0.015	0.001	0.015
Φ (18)	-0.007	-0.002	-0.007	-0.002
	-0.002	0.016	-0.003	0.016
Φ (19)	-0.005	-0.001	-0.005	-0.001
	0.000	0.014	0.000	0.014
Φ (20)	-0.002	-0.001	-0.002	-0.001
	0.001	0.015	0.000	0.015
Φ (21)	-0.001	-0.001	-0.001	-0.001
	-0.003	0.014	-0.002	0.014
Φ (22)	-0.002	0.000	-0.002	0.000
	-0.003	0.011	-0.002	0.012
Φ (23)	-0.003	-0.003	-0.003	-0.003
	0.001	0.013	0.001	0.014
Φ (24)	-0.003	-0.001	-0.003	-0.001
	-0.003	0.015	-0.003	0.015
Φ (25)	-0.002	0.001	-0.002	0.001
	-0.003	0.013	-0.003	0.013
Φ (26)	0.001	0.001	0.001	0.001
	0.004	0.012	0.004	0.012
Φ (27)	-0.001	-0.001	-0.001	-0.001
	0.000	0.013	0.001	0.014
Σ	1.52E-03	-2.28E-05	1.45E-03	-2.24E-05
	-2.28E-05	1.91E-03	-2.24E-05	1.91E-03
DF	1.469		1.409	

Figure 4.3.2 Response to price shocks, ML estimation

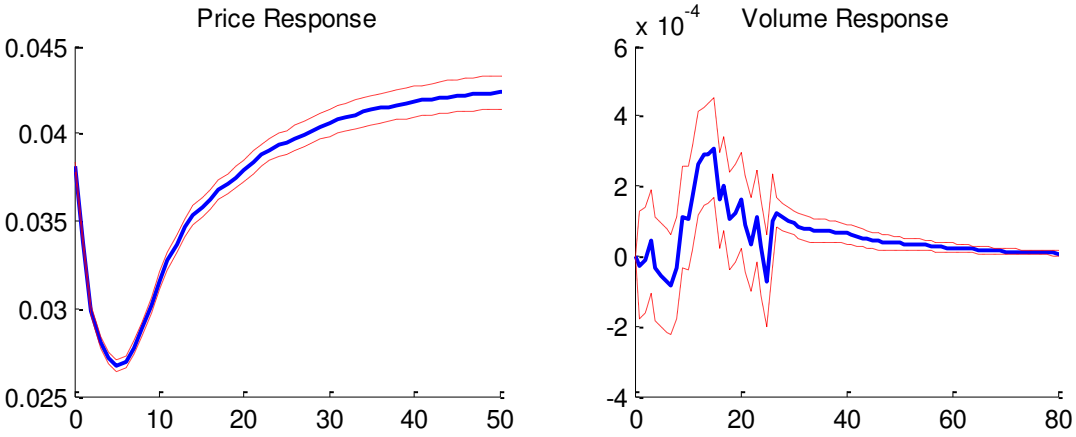


Note:

The left panel is the response of price to the price shocks.

The right panel is the response of volume to the price shocks.

Figure 4.3.3 Response to price shocks, Bayesian estimation

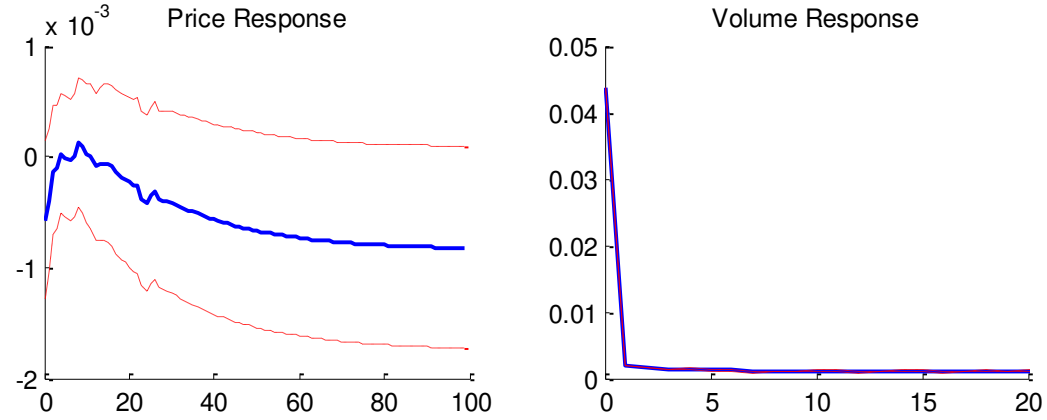


Note:

The left panel is the response of price to the price shocks.

The right panel is the response of volume to the price shocks.

Figure 4.3.4 Response to volume shocks, ML estimation

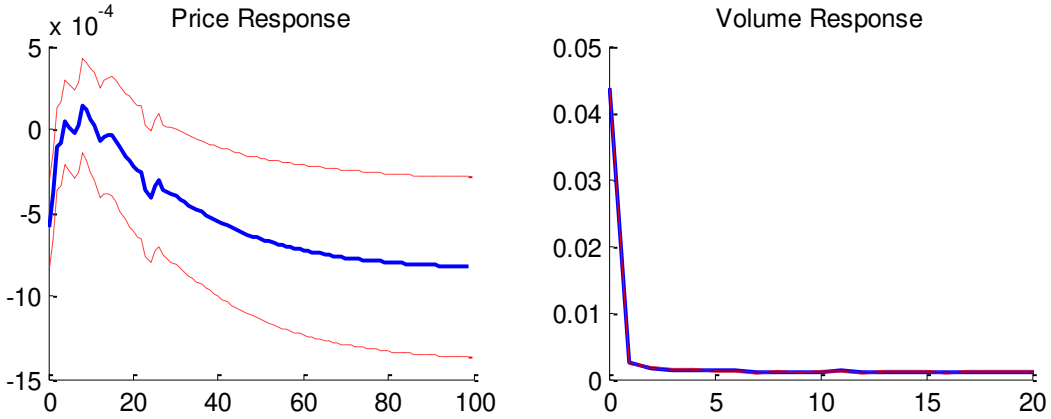


Note:

The left panel is the response of price to the volume shocks.

The right panel is the response of volume to the volume shocks.

Figure 4.3.5 Response to volume shocks, Bayesian estimation



Note:

The left panel is the response of price to the volume shocks.

The right panel is the response of volume to the volume shocks.

5. Extensions

In this section, we extend the scale-mixture SUR Tobit model in two directions. Firstly, we keep the scale-mixture assumption and extend the model to incidental truncation, namely sample selection models. Secondly, we keep the basic Tobit specification, but extend the disturbance terms to finite Gaussian regime mixtures, which in principle can imitate any shape of distribution.

5.1 Sample selection model

In the benchmark model in Section 2.1, the censoring mechanism and activity levels are governed by the same stochastic process, that is, the variable is censored to zero whenever its value falls below zero. However, sometimes it is of interest to consider the separation of participation equation and the activity level equation. In other words, one variable could be incidentally censored whenever another related variable falls below zero.

The sample selection SUR Tobit with scale mixtures can be written as:

- Participation equations:
  $Z^*_{1t} = \mathbf{w_{1t}}\,\boldsymbol{\gamma_1} + u_{1t}$
  … …
  $Z^*_{mt} = \mathbf{w_{mt}}\,\boldsymbol{\gamma_m} + u_{mt}$
- Activity level equations:
  $Y^*_{1t} = \mathbf{x_{1t}}\,\boldsymbol{\beta_1} + \varepsilon_{1t}$
  … …
  $Y^*_{m\,t} = \mathbf{x_{m\,t}}\,\boldsymbol{\beta_m} + \varepsilon_{m\,t}$
- Selection mechanism:
  $Z_{k\,t} = I(Z^*_{k\,t} > 0)$
  $Y_{k\,t} = Y^*_{k\,t} \cdot Z_{k\,t}$
  $k = 1, \dots, m$
- Error specification:
  $(u_{1t} \dots u_{mt}, \dots, \varepsilon_{1t} \dots \varepsilon_{m\,t})' \,|\tau_t \sim \mathrm{MVN}(\mathbf{0}, \tau_t\,\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is $2m \times 2m$ positive definite matrix
  $\tau_t \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$

The $Z_{k\,t}$ and $Y_{k\,t}$ are observable, while the latent $Z^*_{k\,t}$, $Y^*_{k\,t}$, τ_t are not.

Essentially, the censoring rules are: $Y^*_{k\,t}$ will be censored to zero whenever $Z^*_{k\,t}$ falls below zero. In most applications, $Z^*_{k\,t}$ is an artifact constructed from the knowledge whether the observable $Y_{k\,t}$ is zero or not.

For future reference, stack the variables in the obvious way,

$$\mathbf{Z^*_t} = \begin{pmatrix} Z^*_{1t} \\ \dots \\ Z^*_{mt} \end{pmatrix},\ \mathbf{Z_t} = \begin{pmatrix} Z_{1t} \\ \dots \\ Z_{mt} \end{pmatrix},\ \mathbf{Y^*_t} = \begin{pmatrix} Y^*_{1t} \\ \dots \\ Y^*_{mt} \end{pmatrix},\ \mathbf{Y_t} = \begin{pmatrix} Y_{1t} \\ \dots \\ Y_{mt} \end{pmatrix},$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma_1} \\ \dots \\ \boldsymbol{\gamma_m} \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \dots \\ \boldsymbol{\beta_m} \end{pmatrix},\ \mathbf{u_t} = \begin{pmatrix} u_{1t} \\ \dots \\ u_{mt} \end{pmatrix},\ \boldsymbol{\varepsilon_t} = \begin{pmatrix} \varepsilon_{1t} \\ \dots \\ \varepsilon_{mt} \end{pmatrix}$$

$$\mathbf{W_t} = \mathrm{diag}(\mathbf{w_{1t}}, \dots, \mathbf{w_{m\,t}}),\ \mathbf{X_t} = \mathrm{diag}(\mathbf{x_{1t}}, \dots, \mathbf{x_{m\,t}})$$

$$\mathbf{Z^*} = \begin{pmatrix} \mathbf{Z^*_1} \\ \dots \\ \mathbf{Z^*_T} \end{pmatrix},\ \mathbf{Z} = \begin{pmatrix} \mathbf{Z_1} \\ \dots \\ \mathbf{Z_T} \end{pmatrix},\ \mathbf{Y^*} = \begin{pmatrix} \mathbf{Y^*_1} \\ \dots \\ \mathbf{Y^*_T} \end{pmatrix},\ \mathbf{Y} = \begin{pmatrix} \mathbf{Y_1} \\ \dots \\ \mathbf{Y_T} \end{pmatrix},\ \mathbf{W} = \begin{pmatrix} \mathbf{W_1} \\ \dots \\ \mathbf{W_T} \end{pmatrix},\ \mathbf{X} = \begin{pmatrix} \mathbf{X_1} \\ \dots \\ \mathbf{X_T} \end{pmatrix},\ \boldsymbol{\tau} = \begin{pmatrix} \tau_1 \\ \dots \\ \tau_T \end{pmatrix}$$

To discuss the estimation, we need some further notations.

Without loss of generality, for some individual t, rearrange the order of variables so that the first m_u elements of $\mathbf{Y^*_t}$ are intact but the last $m_c \equiv m - m_u$ variates are censored to zero. We use the subscript "u" to denote the non-censored block, and subscript "c" to denote the censored block. So the

variables are partitioned to be:

$$\mathbf{Z_t^*} = \begin{pmatrix} \mathbf{Z_{ut}^*} \\ \mathbf{Z_{ct}^*} \end{pmatrix}, \mathbf{Z_t} = \begin{pmatrix} \mathbf{Z_{ut}} \\ \mathbf{Z_{ct}} \end{pmatrix}, \ \mathbf{Y_t^*} = \begin{pmatrix} \mathbf{Y_{ut}^*} \\ \mathbf{Y_{ct}^*} \end{pmatrix}, \mathbf{Y_t} = \begin{pmatrix} \mathbf{Y_{ut}} \\ \mathbf{Y_{ct}} \end{pmatrix}, \ \mathbf{W_t} = \begin{pmatrix} \mathbf{W_{ut}} & \\ & \mathbf{W_{ct}} \end{pmatrix}, \mathbf{X_t} = \begin{pmatrix} \mathbf{X_{ut}} & \\ & \mathbf{X_{ct}} \end{pmatrix},$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma_u} \\ \boldsymbol{\gamma_c} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_u} \\ \boldsymbol{\beta_c} \end{pmatrix}, \ \mathbf{u_t} = \begin{pmatrix} \mathbf{u_{ut}} \\ \mathbf{u_{ct}} \end{pmatrix}, \ \boldsymbol{\varepsilon_t} = \begin{pmatrix} \boldsymbol{\varepsilon_{ut}} \\ \boldsymbol{\varepsilon_{ct}} \end{pmatrix}$$

$$\mathrm{Var}\begin{pmatrix} \mathbf{u_{ut}} \\ \mathbf{u_{ct}} \\ \boldsymbol{\varepsilon_{ut}} \\ \boldsymbol{\varepsilon_{ct}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} & \boldsymbol{\Sigma_{13}} & \boldsymbol{\Sigma_{14}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} & \boldsymbol{\Sigma_{23}} & \boldsymbol{\Sigma_{24}} \\ \boldsymbol{\Sigma_{31}} & \boldsymbol{\Sigma_{32}} & \boldsymbol{\Sigma_{33}} & \boldsymbol{\Sigma_{34}} \\ \boldsymbol{\Sigma_{41}} & \boldsymbol{\Sigma_{42}} & \boldsymbol{\Sigma_{43}} & \boldsymbol{\Sigma_{44}} \end{pmatrix} \equiv \widetilde{\boldsymbol{\Sigma}}$$

$\widetilde{\boldsymbol{\Sigma}}$ is some permutation of $\boldsymbol{\Sigma}$ conformable to the above variable rearrangement.

The log likelihood is $\log L = \sum_{t=1}^{T} p(\mathbf{Z_{ut}}, \mathbf{Z_{ct}}, \mathbf{Y_{ut}}, \mathbf{Y_{ct}})$, where

$$p(\mathbf{Z_{ut}}, \mathbf{Z_{ct}}, \mathbf{Y_{ut}}, \mathbf{Y_{ct}}) = \int_{-\infty}^{\infty} \int_{-\infty}^{0} \int_{0}^{\infty} p(\mathbf{Z_{ut}^*}, \mathbf{Z_{ct}^*}, \mathbf{Y_{ut}^*}, \mathbf{Y_{ct}^*}) \, d\mathbf{Z_{ut}^*} \, d\mathbf{Z_{ct}^*} \, d\mathbf{Y_{ct}^*}$$

$$= \int_{-\infty}^{0} \int_{0}^{\infty} p(\mathbf{Z_{ut}^*}, \mathbf{Z_{ct}^*}, \mathbf{Y_{ut}^*}) \, d\mathbf{Z_{ut}^*} \, d\mathbf{Z_{ct}^*}$$

$$= p(\mathbf{Y_{ut}^*}) \cdot \int_{-\infty}^{0} \int_{0}^{\infty} p(\mathbf{Z_{ut}^*}, \mathbf{Z_{ct}^*} | \mathbf{Y_{ut}^*}) \, d\mathbf{Z_{ut}^*} \, d\mathbf{Z_{ct}^*}$$

$$= p(\mathbf{Y_{ut}^*}) \cdot \int_{0}^{\infty} \int_{-\infty}^{0} \int_{0}^{\infty} p(\tau_t | \mathbf{Y_{ut}^*}) \, p(\mathbf{Z_{ut}^*}, \mathbf{Z_{ct}^*} | \mathbf{Y_{ut}^*}, \tau_t) \, d\mathbf{Z_{ut}^*} \, d\mathbf{Z_{ct}^*} \, d\tau_t$$

As a notation, variables and intervals have the order like $\int_{x_1 \in A_1} \int_{x_{2 \in A_2}} \int_{x_{3 \in A_3}} f(x_1, x_2, x_3) \, dx_3 \, dx_2 \, dx_1$

It can be shown that

$\mathbf{Y_{ut}^*} \sim \mathrm{MVT}(\mathbf{X_{ut}} \, \boldsymbol{\beta_u} \, ; \, \boldsymbol{\Sigma_{33}}, \nu)$

$\tau_t | \mathbf{Y_{ut}^*} \sim \mathrm{IG}\left\{ \frac{\nu + m_u}{2}, \left[\frac{\nu}{2} + \frac{1}{2} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}} \, \boldsymbol{\beta_u})' \boldsymbol{\Sigma_{33}^{-1}} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}} \, \boldsymbol{\beta_u}) \right]^{-1} \right\}$

$\mathbf{Z_{ut}^*}, \mathbf{Z_{ct}^*} | \mathbf{Y_{ut}^*}, \tau_t \sim \mathrm{MVN}(\bar{\boldsymbol{\mu}}, \ \bar{\boldsymbol{\Sigma}})$

Where $\bar{\boldsymbol{\mu}} \equiv \begin{pmatrix} \mathbf{W_{ut}} \, \boldsymbol{\gamma_u} \\ \mathbf{W_{ct}} \, \boldsymbol{\gamma_c} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Sigma_{13}} \\ \boldsymbol{\Sigma_{23}} \end{pmatrix} \boldsymbol{\Sigma_{33}^{-1}} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}} \, \boldsymbol{\beta_u})$

$$\bar{\boldsymbol{\Sigma}} \equiv \tau_t \left[\begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Sigma_{13}} \\ \boldsymbol{\Sigma_{23}} \end{pmatrix} \boldsymbol{\Sigma_{33}^{-1}} \begin{pmatrix} \boldsymbol{\Sigma_{31}} & \boldsymbol{\Sigma_{32}} \end{pmatrix} \right]$$

Obviously, simulation techniques must be used to evaluate the likelihood function. GHK simulator adapted for MVT c.d.f., as is described in Section 2.2, can be used to evaluate the integral.

Alternatively, the above sample selection model can also be estimated via EM or Bayesian approach. We briefly outline the steps as follows.

E-M algorithm:

The complete data log likelihood can be written as

$\ln p(\mathbf{Z}, \mathbf{Z^*}, \mathbf{Y}, \mathbf{Y^*}, \boldsymbol{\tau} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \nu)$

$= \sum_{t=1}^{T} [\ln p(\tau_t | \nu) + \ln p(\mathbf{Z_t^*}, \mathbf{Y_t^*} | \tau_t, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})] + \text{constant}$

$$= \sum_{t=1}^{T} \begin{bmatrix} \ln \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \ln\left(\frac{\nu}{2}\right) - \left(\frac{\nu}{2} + 1\right) \ln \tau_t - \frac{\nu}{2} \tau_t^{-1} \\ -\frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \tau_t^{-1} \begin{pmatrix} \mathbf{Z_t^*} - \mathbf{W_t} \, \boldsymbol{\gamma} \\ \mathbf{Y_t^*} - \mathbf{X_t} \, \boldsymbol{\beta} \end{pmatrix}' \boldsymbol{\Sigma^{-1}} \begin{pmatrix} \mathbf{Z_t^*} - \mathbf{W_t} \, \boldsymbol{\gamma} \\ \mathbf{Y_t^*} - \mathbf{X_t} \, \boldsymbol{\beta} \end{pmatrix} \end{bmatrix} + \text{constant}$$

We use the informal notation “constant” to represent those additive terms which do not interact with parameters $\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu$

The joint posterior distribution of $\mathbf{Z_t^*}, \mathbf{Y_t^*}, \tau_t$ does not have a recognizable form, but the conditional distribution does.

$\mathbf{Z_{ut}^*}|\tau_t, \mathbf{Z_{ut}}, \mathbf{Z_{ct}^*}, \mathbf{Y_{ut}^*}, \mathbf{Y_{ct}^*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu \sim \mathrm{MVN}_{(\mathbf{0},\infty)}(\bar{\boldsymbol{\mu}}_{\mathbf{zu}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{zu}})$

$\mathbf{Z_{ct}^*}|\tau_t, \mathbf{Z_{ut}^*}, \mathbf{Z_{ct}}, \mathbf{Y_{ut}^*}, \mathbf{Y_{ct}^*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu \sim \mathrm{MVN}_{(-\infty,\mathbf{0})}(\bar{\boldsymbol{\mu}}_{\mathbf{zc}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{zc}})$

$\mathbf{Y_{ut}^*} \equiv \mathbf{Y_{ut}}$

$\mathbf{Y_{ct}^*}|\tau_t, \mathbf{Z_{ut}^*}, \mathbf{Z_{ct}^*}, \mathbf{Y_{ut}^*}, \mathbf{Y_{ct}}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu \sim \mathrm{MVN}(\bar{\boldsymbol{\mu}}_{\mathbf{yc}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{yc}})$

$$\tau_t|\mathbf{Y_t^*}, \mathbf{Z_t^*}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu \sim \mathrm{IG}\left\{\frac{\nu+m}{2}, \left[\frac{\nu}{2} + \frac{1}{2}\begin{pmatrix}\mathbf{Z_t^*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y_t^*} - \mathbf{X_t}\,\boldsymbol{\beta}\end{pmatrix}' \boldsymbol{\Sigma}^{-1} \begin{pmatrix}\mathbf{Z_t^*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y_t^*} - \mathbf{X_t}\,\boldsymbol{\beta}\end{pmatrix}\right]^{-1}\right\}$$

where

$$\bar{\boldsymbol{\mu}}_{\mathbf{zu}} \equiv \mathbf{W_{ut}}\,\boldsymbol{\gamma_u} + (\boldsymbol{\Sigma_{12}} \quad \boldsymbol{\Sigma_{13}} \quad \boldsymbol{\Sigma_{14}})\begin{pmatrix}\boldsymbol{\Sigma_{22}} & \boldsymbol{\Sigma_{23}} & \boldsymbol{\Sigma_{24}} \\ \boldsymbol{\Sigma_{32}} & \boldsymbol{\Sigma_{33}} & \boldsymbol{\Sigma_{34}} \\ \boldsymbol{\Sigma_{42}} & \boldsymbol{\Sigma_{43}} & \boldsymbol{\Sigma_{44}}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{Z_{ct}^*} - \mathbf{W_{ct}}\,\boldsymbol{\gamma_u} \\ \mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\boldsymbol{\beta_u} \\ \mathbf{Y_{ct}^*} - \mathbf{X_{ct}}\boldsymbol{\beta_c}\end{pmatrix}$$

$$\bar{\boldsymbol{\mu}}_{\mathbf{zc}} \equiv \mathbf{W_{ct}}\,\boldsymbol{\gamma_c} + (\boldsymbol{\Sigma_{21}} \quad \boldsymbol{\Sigma_{23}} \quad \boldsymbol{\Sigma_{24}})\begin{pmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{13}} & \boldsymbol{\Sigma_{14}} \\ \boldsymbol{\Sigma_{31}} & \boldsymbol{\Sigma_{33}} & \boldsymbol{\Sigma_{34}} \\ \boldsymbol{\Sigma_{41}} & \boldsymbol{\Sigma_{43}} & \boldsymbol{\Sigma_{44}}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{Z_{ut}^*} - \mathbf{W_{ut}}\,\boldsymbol{\gamma_u} \\ \mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\boldsymbol{\beta_u} \\ \mathbf{Y_{ct}^*} - \mathbf{X_{ct}}\boldsymbol{\beta_c}\end{pmatrix}$$

$$\bar{\boldsymbol{\mu}}_{\mathbf{yc}} \equiv \mathbf{X_{ct}}\,\boldsymbol{\beta_c} + (\boldsymbol{\Sigma_{41}} \quad \boldsymbol{\Sigma_{42}} \quad \boldsymbol{\Sigma_{43}})\begin{pmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} & \boldsymbol{\Sigma_{13}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} & \boldsymbol{\Sigma_{23}} \\ \boldsymbol{\Sigma_{31}} & \boldsymbol{\Sigma_{32}} & \boldsymbol{\Sigma_{33}}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{Z_{ut}^*} - \mathbf{W_{ut}}\,\boldsymbol{\gamma_u} \\ \mathbf{Z_{ct}^*} - \mathbf{W_{ct}}\,\boldsymbol{\gamma_u} \\ \mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\boldsymbol{\beta_u}\end{pmatrix}$$

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{zu}} \equiv \tau_t\left[\boldsymbol{\Sigma_{11}} - (\boldsymbol{\Sigma_{12}} \quad \boldsymbol{\Sigma_{13}} \quad \boldsymbol{\Sigma_{14}})\begin{pmatrix}\boldsymbol{\Sigma_{22}} & \boldsymbol{\Sigma_{23}} & \boldsymbol{\Sigma_{24}} \\ \boldsymbol{\Sigma_{32}} & \boldsymbol{\Sigma_{33}} & \boldsymbol{\Sigma_{34}} \\ \boldsymbol{\Sigma_{42}} & \boldsymbol{\Sigma_{43}} & \boldsymbol{\Sigma_{44}}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{\Sigma_{21}} \\ \boldsymbol{\Sigma_{31}} \\ \boldsymbol{\Sigma_{41}}\end{pmatrix}\right]$$

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{zc}} \equiv \tau_t\left[\boldsymbol{\Sigma_{22}} - (\boldsymbol{\Sigma_{21}} \quad \boldsymbol{\Sigma_{23}} \quad \boldsymbol{\Sigma_{24}})\begin{pmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{13}} & \boldsymbol{\Sigma_{14}} \\ \boldsymbol{\Sigma_{31}} & \boldsymbol{\Sigma_{33}} & \boldsymbol{\Sigma_{34}} \\ \boldsymbol{\Sigma_{41}} & \boldsymbol{\Sigma_{43}} & \boldsymbol{\Sigma_{44}}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{32}} \\ \boldsymbol{\Sigma_{42}}\end{pmatrix}\right]$$

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{yc}} \equiv \tau_t\left[\boldsymbol{\Sigma_{44}} - (\boldsymbol{\Sigma_{41}} \quad \boldsymbol{\Sigma_{42}} \quad \boldsymbol{\Sigma_{43}})\begin{pmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} & \boldsymbol{\Sigma_{13}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} & \boldsymbol{\Sigma_{23}} \\ \boldsymbol{\Sigma_{31}} & \boldsymbol{\Sigma_{32}} & \boldsymbol{\Sigma_{33}}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{\Sigma_{14}} \\ \boldsymbol{\Sigma_{24}} \\ \boldsymbol{\Sigma_{34}}\end{pmatrix}\right]$$

Make sure that $\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu$ are evaluated at estimators of last iteration.

With the posterior conditionals, Gibbs Sampler can be used to obtain draws:

$\{\mathbf{Z_{ut(r)}^*}, \mathbf{Z_{ct(r)}^*}, \mathbf{Y_{ut\,(r)}^*}, \mathbf{Y_{ct\,(r)}^*}, \tau_{t\,(r)}\}_{r=1}^{R}$, or more compactly $\{\mathbf{Z_{t\,(r)}^*}, \mathbf{Y_{t\,(r)}^*}, \tau_{t\,(r)}\}_{r=1}^{R}$, $t = 1, \ldots, T$.

The expected complete data likelihood then can be approximated by the ergodic mean:

$\mathrm{E}[\ln p(\mathbf{Y}, \mathbf{Y}^*, \boldsymbol{\tau}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)]$

$$\approx \frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T}\begin{bmatrix} -\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) - \left(\frac{\nu}{2}+1\right)\ln\tau_{t\,(r)} - \frac{\nu}{2}\tau_{t\,(r)}^{-1} \\ -\frac{m}{2}\ln\tau_{t\,(r)} - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\tau_{t\,(r)}^{-1}\begin{pmatrix}\mathbf{Z_{t\,(r)}^*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y_{t\,(r)}^*} - \mathbf{X_t}\,\boldsymbol{\beta}\end{pmatrix}'\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\mathbf{Z_{t\,(r)}^*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y_{t\,(r)}^*} - \mathbf{X_t}\,\boldsymbol{\beta}\end{pmatrix}\end{bmatrix} + \text{constant}$$

The conditional M-steps have the following cycling update rule of estimators:

$$\begin{pmatrix}\boldsymbol{\gamma} \\ \boldsymbol{\beta}\end{pmatrix} = \left[\sum_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\right)\begin{pmatrix}\mathbf{W_t} & \\ & \mathbf{X_t}\end{pmatrix}'\boldsymbol{\Sigma}^{-1}\cdot\begin{pmatrix}\mathbf{W_t} & \\ & \mathbf{X_t}\end{pmatrix}\right]^{-1}$$
$$\cdot\sum_{t=1}^{T}\left[\begin{pmatrix}\mathbf{W_t} & \\ & \mathbf{X_t}\end{pmatrix}'\cdot\boldsymbol{\Sigma}^{-1}\cdot\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\begin{pmatrix}\mathbf{Z_{t\,(r)}^*} \\ \mathbf{Y_{t\,(r)}^*}\end{pmatrix}\right)\right]$$

$$\boldsymbol{\Sigma} = \frac{1}{\mathrm{T\,R}} \sum_{r=1}^{R} \sum_{t=1}^{T} \left[\tau_{t\,(r)}^{-1} \begin{pmatrix} \mathbf{Z}_{\mathbf{t}\,(\mathbf{r})}^{*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y}_{\mathbf{t}\,(\mathbf{r})}^{*} - \mathbf{X_t}\,\boldsymbol{\beta} \end{pmatrix} \begin{pmatrix} \mathbf{Z}_{\mathbf{t}\,(\mathbf{r})}^{*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y}_{\mathbf{t}\,(\mathbf{r})}^{*} - \mathbf{X_t}\,\boldsymbol{\beta} \end{pmatrix}' \right]$$

$$\nu = \arg\max \sum_{t=1}^{T} \left[-\ln \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \ln\left(\frac{\nu}{2}\right) - \left(\frac{\nu}{2} + 1\right) \left(\frac{1}{R} \sum_{r=1}^{R} \ln \tau_{t\,(r)} \right) - \frac{\nu}{2} \left(\frac{1}{R} \sum_{r=1}^{R} \tau_{t\,(r)}^{-1} \right) \right]$$

As for the Bayesian approach to estimate the model,
The proper priors are assumed to be:

$\boldsymbol{\gamma} \sim \mathrm{MVN}\left(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \mathbf{V}_{\boldsymbol{\gamma}}\right)$
$\boldsymbol{\beta} \sim \mathrm{MVN}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}}\right)$
$\boldsymbol{\Sigma}^{-1} \sim \mathrm{Wishart}\ (\boldsymbol{\Omega}, \mathrm{k})$
$\nu \sim \mathrm{Uniform}\left(\underline{\nu}, \bar{\nu}\right)$
The full set of posterior conditionals are:

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix} | \cdot \sim \mathrm{MVN}(\mathbf{D}\mathbf{d}, \mathbf{D})$$

$$\text{where } \mathbf{D} = \left[\sum_{t=1}^{T} \tau_t^{-1} \cdot \begin{pmatrix} \mathbf{W_t} & \\ & \mathbf{X_t} \end{pmatrix}' \cdot \boldsymbol{\Sigma}^{-1} \cdot \begin{pmatrix} \mathbf{W_t} & \\ & \mathbf{X_t} \end{pmatrix} + \begin{pmatrix} \mathbf{V}_{\boldsymbol{\gamma}} & \\ & \mathbf{V}_{\boldsymbol{\beta}} \end{pmatrix}^{-1} \right]^{-1}$$

$$\mathbf{d} = \left[\sum_{t=1}^{T} \tau_t^{-1} \cdot \begin{pmatrix} \mathbf{W_t} & \\ & \mathbf{X_t} \end{pmatrix}' \cdot \boldsymbol{\Sigma}^{-1} \cdot \begin{pmatrix} \mathbf{Z}_{\mathbf{t}}^{*} \\ \mathbf{Y}_{\mathbf{t}}^{*} \end{pmatrix} + \begin{pmatrix} \mathbf{V}_{\boldsymbol{\gamma}} & \\ & \mathbf{V}_{\boldsymbol{\beta}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\gamma}} \\ \boldsymbol{\mu}_{\boldsymbol{\beta}} \end{pmatrix} \right]$$

$$\boldsymbol{\Sigma}^{-1} | \cdot \sim \mathrm{Wishart} \left\{ \left[\boldsymbol{\Omega}^{-1} + \sum_{t=1}^{T} \tau_t^{-1} \cdot \begin{pmatrix} \mathbf{Z}_{\mathbf{t}}^{*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y}_{\mathbf{t}}^{*} - \mathbf{X_t}\,\boldsymbol{\beta} \end{pmatrix} \begin{pmatrix} \mathbf{Z}_{\mathbf{t}}^{*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y}_{\mathbf{t}}^{*} - \mathbf{X_t}\,\boldsymbol{\beta} \end{pmatrix}' \right]^{-1}, \ \mathrm{T} + \mathrm{k} \right\}$$

$\mathbf{Z}_{\mathbf{ut}}^{*} | \cdot \sim \mathrm{MVN}_{(\mathbf{0},\infty)}(\bar{\boldsymbol{\mu}}_{\mathbf{zu}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{zu}})$
$\mathbf{Z}_{\mathbf{ct}}^{*} | \cdot \sim \mathrm{MVN}_{(-\infty,\mathbf{0})}(\bar{\boldsymbol{\mu}}_{\mathbf{zc}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{zc}})$
$\mathbf{Y}_{\mathbf{ut}}^{*} \equiv \mathbf{Y}_{\mathbf{ut}}$
$\mathbf{Y}_{\mathbf{ct}}^{*} | \cdot \sim \mathrm{MVN}(\bar{\boldsymbol{\mu}}_{\mathbf{yc}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{yc}})$
where $\bar{\boldsymbol{\mu}}_{\mathbf{zu}}$, $\bar{\boldsymbol{\mu}}_{\mathbf{zc}}$, $\bar{\boldsymbol{\mu}}_{\mathbf{yc}}$, $\bar{\boldsymbol{\Sigma}}_{\mathbf{zu}}$, $\bar{\boldsymbol{\Sigma}}_{\mathbf{zc}}$, $\bar{\boldsymbol{\Sigma}}_{\mathbf{yc}}$ are defined above.

$$\tau_t | \cdot \sim \mathrm{IG} \left\{ \frac{\nu + \mathrm{m}}{2}, \left[\frac{\nu}{2} + \frac{1}{2} \begin{pmatrix} \mathbf{Z}_{\mathbf{t}}^{*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y}_{\mathbf{t}}^{*} - \mathbf{X_t}\,\boldsymbol{\beta} \end{pmatrix}' \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \mathbf{Z}_{\mathbf{t}}^{*} - \mathbf{W_t}\,\boldsymbol{\gamma} \\ \mathbf{Y}_{\mathbf{t}}^{*} - \mathbf{X_t}\,\boldsymbol{\beta} \end{pmatrix} \right]^{-1} \right\}$$

$$\mathrm{p}(\nu | \cdot) \propto \prod_{t=1}^{T} \left[\Gamma\left(\frac{\nu}{2}\right) \cdot \left(\frac{2}{\nu}\right)^{\nu/2} \right]^{-1} (\tau_t)^{-\left(\frac{\nu}{2}+1\right)} \mathrm{e}^{-\frac{\nu}{2}\tau_t^{-1}} \cdot \mathrm{I}\left(\underline{\nu} < \nu < \bar{\nu}\right)$$

The last one can be drawn by Metropolis-within-Gibbs procedure.

It should be noted that the sample selection model has an identification issue. Similar to Probit, the participation equation should have disturbance variance normalized to unity, i.e. $\boldsymbol{\Sigma}_{\mathbf{11}}$ and $\boldsymbol{\Sigma}_{\mathbf{22}}$ has unit diagonal elements. That is not a big problem for ML—in fact, it makes ML estimation easier because we do not need to estimate those main diagonal elements. However, for EM and Bayesian sampler, it does matter. In Bayesian Gibbs sampler, we might resolve the problem by taking draws from restricted Wishart distribution, as is described by Nobile (2000). But for EM algorithm, that means the condition M-step cannot yield a close-form update rule for $\boldsymbol{\Sigma}$. How to address the issue is left for future research.

5.2 Finite Gaussian Regime mixtures

In the benchmark model outlined in Section 2, the disturbances are scale mixture with marginal distribution as multivariate t. Although it is more general than the normality assumption and can accommodate leptokurtosis, it does not help to describe skewness, multi-modality etc.

Finite Gaussian Regime mixtures model has a long history, maybe as early as Newcomb (1886). Finite Gaussian mixture models provide a flexible representation that could mimic virtually all shapes of density, including multimodality, skewness and Leptokurtosis, etc. But the model also has a drawback— invariance of relabeling, i.e. the permutation of the parameter vector across regimes will not change the likelihood function. In that case, estimation could exhibits unusual properties and interpretation of estimators is difficult. We do not discuss the identification issues here, but the limitation should always be born in mind.

Suppose there are S potential regimes which could generate the data. Denote the latent variable $\tau_t = s$ whenever regime "s" occurs, with probability π_s. That is, $p(\tau_t) = \sum_{s=1}^{S} \pi_s \cdot I(\tau_t = s)$

The SUR in Regime "s": $\mathbf{Y_t^*} = \mathbf{x_t}\,\boldsymbol{\beta_s} + \boldsymbol{\varepsilon_t}$, $s = 1, \dots, S$

where $\mathbf{Y_t^*} = \begin{pmatrix} Y_{1t}^* \\ \dots \\ Y_{mt}^* \end{pmatrix}$, $\mathbf{X_t} = \mathrm{diag}(\mathbf{x_{1t}}, \dots, \mathbf{x_{m\,t}})$, $\boldsymbol{\varepsilon_t} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma_s})$, $\boldsymbol{\Sigma_s}$ is m*m positive definite

The observables are $\mathbf{Y_t} = \mathbf{max}(\mathbf{0}, \mathbf{Y_t^*})$; the "max" operator works element-wise.

The ML estimation of the model is straightforward.

Again, for individual t, partition the data into the non-censored block "u", and censored block "c"

$$\mathbf{Y_t^*} = \begin{pmatrix} \mathbf{Y_{ut}^*} \\ \mathbf{Y_{ct}^*} \end{pmatrix}, \mathbf{Y_t} = \begin{pmatrix} \mathbf{Y_{ut}} \\ \mathbf{Y_{ct}} \end{pmatrix}, \mathbf{X_t} = \begin{pmatrix} \mathbf{X_{ut}} & \\ & \mathbf{X_{ct}} \end{pmatrix}, \boldsymbol{\beta_s} = \begin{pmatrix} \boldsymbol{\beta}_{s,u} \\ \boldsymbol{\beta}_{s,c} \end{pmatrix}, \boldsymbol{\Sigma_s} = \begin{pmatrix} \boldsymbol{\Sigma}_{s,uu} & \boldsymbol{\Sigma}_{s,uc} \\ \boldsymbol{\Sigma}_{s,cu} & \boldsymbol{\Sigma}_{s,cc} \end{pmatrix}$$

The log likelihood is $\log L = \sum_{t=1}^{T} \ln p(\mathbf{Y_{ut}}, \mathbf{Y_{ct}})$, where

$$p(\mathbf{Y_{ut}}, \mathbf{Y_{ct}}) = \sum_{s=1}^{S} \left[\pi_s \cdot p(\mathbf{Y_{ut}^*} | \tau_t = s) \cdot \int_{-\infty}^{\mathbf{0}} p(\mathbf{Y_{ct}^*} | \mathbf{Y_{ut}^*}, \tau_t = s)\, d\, \mathbf{Y_{ct}^*} \right]$$

Apparently,

$\mathbf{Y_{ut}^*} | \tau_t = s \sim \mathrm{MVN}(\mathbf{X_{ut}}\boldsymbol{\beta}_{s,u}, \boldsymbol{\Sigma}_{s,uu})$

$\mathbf{Y_{ct}^*} | \mathbf{Y_{ut}^*}, \tau_t = s \sim \mathrm{MVN}\left[\mathbf{X_{ct}}\,\boldsymbol{\beta}_{s,c} + \boldsymbol{\Sigma}_{s,cu}\,\boldsymbol{\Sigma}_{s,uu}^{-1}\left(\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\,\boldsymbol{\beta}_{s,u}\right), \boldsymbol{\Sigma}_{s,cc} - \boldsymbol{\Sigma}_{s,cu}\,\boldsymbol{\Sigma}_{s,uu}^{-1}\,\boldsymbol{\Sigma}_{s,uc}\right]$

So essentially the likelihood is a weighted averge of a MVN p.d.f. multiplied by a MVN c.d.f.

GHK simulators can be directly applied to evaluate the likelihood function.

Alternatively, the model can be estimated via EM or Bayesian approach.

For EM algorithm,

The complete data log likelihood can be written as

$$\ln p(\mathbf{Y}, \mathbf{Y}^*, \boldsymbol{\tau} | \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)$$

$$= \sum_{t=1}^{T} \ln\left[\sum_{s=1}^{S} \pi_s \cdot \phi\left(\mathbf{Y_t^*}; \mathbf{X_t}\boldsymbol{\beta_s}, \boldsymbol{\Sigma_s}\right) \cdot I(\tau_t = s)\right] + \text{constant}$$

The joint posterior of latent variables $\mathbf{Y_t^*}$ and τ_t does not have a close form, but the direct Monte Carlo (not MCMC!) is available.

$$p(\tau_t | \mathbf{Y_t}) \propto p(\tau_t) \cdot p(\mathbf{Y_{ut}} | \tau_t) \cdot p(\mathbf{Y_{ct}} | \mathbf{Y_{ut}}, \tau_t)$$

$$= \sum_{s=1}^{S} \left\{ \begin{array}{c} \pi_s \cdot I(\tau_t = s) \cdot \phi\left(\mathbf{Y_{ut}^*};\, \mathbf{X_{ut}}\boldsymbol{\beta}_{s,u}, \boldsymbol{\Sigma}_{s,uu}\right) \cdot \\ \Phi\left[0\,;\, \mathbf{X_{ct}}\,\boldsymbol{\beta}_{s,c} + \boldsymbol{\Sigma}_{s,cu}\,\boldsymbol{\Sigma}_{s,uu}^{-1}\left(\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\,\boldsymbol{\beta}_{s,u}\right), \boldsymbol{\Sigma}_{s,cc} - \boldsymbol{\Sigma}_{s,cu}\,\boldsymbol{\Sigma}_{s,uu}^{-1}\,\boldsymbol{\Sigma}_{s,uc}\right] \end{array} \right\}$$

This is a discrete distribution where $\text{pr}(\tau_t = s|\mathbf{Y_t}) \propto \pi_s\, \phi(\cdot)\, \Phi(\cdot)$

The "$\propto$ " can be replaced by "=" when we impose the constraint such that the sum of probability equal to one.

$\mathbf{Y^*_{ut}}|\tau_t, \mathbf{Y_t} \equiv \mathbf{Y_{ut}}$

$\mathbf{Y^*_{ct}}|\tau_t, \mathbf{Y_t} \sim \text{MVN}_{(-\infty,\mathbf{0})}(\bar{\boldsymbol{\mu}}_\mathbf{c}, \bar{\boldsymbol{\Sigma}}_\mathbf{c})$

where $\bar{\boldsymbol{\mu}}_\mathbf{c} = \sum_{s=1}^{S}\left[\mathbf{X_{ct}}\,\boldsymbol{\beta}_{\mathbf{s,c}} + \boldsymbol{\Sigma}_{\mathbf{s,cu}}\,\boldsymbol{\Sigma}^{-1}_{\mathbf{s,uu}}\left(\mathbf{Y^*_{ut}} - \mathbf{X_{ut}}\,\boldsymbol{\beta}_{\mathbf{s,u}}\right)\right] \cdot I(\tau_t = s)$

$\bar{\boldsymbol{\Sigma}}_\mathbf{c} = \sum_{s=1}^{S}\left(\boldsymbol{\Sigma}_{\mathbf{s,cc}} - \boldsymbol{\Sigma}_{\mathbf{s,cu}}\,\boldsymbol{\Sigma}^{-1}_{\mathbf{s,uu}}\,\boldsymbol{\Sigma}_{\mathbf{s,uc}}\right) \cdot I(\tau_t = s)$

Draws of $\{\mathbf{Y_t^*}, \tau_t\}_{t=1}^{T}$ can be obtained by method of composition. That is, first take a draw from $\tau_t|\mathbf{Y_t}$, then use this draw to obtain a draw from $\mathbf{Y^*_{ct}}|\tau_t, \mathbf{Y_t}$ .Repeat for 1,…, T.

Of course, it is possible to use Gibbs sampler and cycle through the posterior conditionals. In that case, we simply use $p(\tau_t|\mathbf{Y_t^*}) \propto \sum_{s=1}^{S}[\pi_s \cdot \phi(\mathbf{Y_t^*};\, \mathbf{X_t}\boldsymbol{\beta}_\mathbf{s}, \boldsymbol{\Sigma}_\mathbf{s}) \cdot I(\tau_t = s)]$. Each method has its own pros and cons. On the one hand, draws obtained from MCMC are not independent, and therefore more draws are needed to achieve the same level of accuracy. On the other hand direct Monte Carlo requires evaluation the normal c.d.f.. For high dimension problems, it is too demanding.

Repeat the above sampling for R time, we will obtain $\left\{\left\{\mathbf{Y^*_{t(r)}}, \tau_{t(r)}\right\}_{t=1}^{T}\right\}_{r=1}^{R}$

The expected complete data likelihood then can be approximated by:

$E[\ln p(\mathbf{Y}, \mathbf{Y}^*, \boldsymbol{\tau}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)]$

$\approx \frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T}\left\{\ln\left[\sum_{s=1}^{S}\pi_s \cdot \phi\left(\mathbf{Y^*_{t(r)}}; \mathbf{X_t}\boldsymbol{\beta}_\mathbf{s}, \boldsymbol{\Sigma}_\mathbf{s}\right) \cdot I(\tau_{t(r)} = s)\right]\right\} + \text{constant}$

The conditional M-steps have the following cycling update rule of estimators:

For each $s = 1, \ldots, S$

$$\boldsymbol{\beta}_\mathbf{s} = \left[\sum_{t=1}^{T}\mathbf{x}_\mathbf{t}' \cdot \boldsymbol{\Sigma}^{-1} \cdot \mathbf{x_t}\right]^{-1} \cdot \sum_{t=1}^{T}\left[\mathbf{x}_\mathbf{t}' \cdot \boldsymbol{\Sigma}^{-1} \cdot \left(\frac{1}{R_{s,t}}\sum_{r=1}^{R}\mathbf{Y^*_{t\,(r)}} \cdot I(\tau_{t(r)} = s)\right)\right]$$

$$\boldsymbol{\Sigma}_\mathbf{s} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{R_{s,t}}\sum_{r=1}^{R}\left(\mathbf{Y^*_{t\,(r)}} - \mathbf{x_t}\,\boldsymbol{\beta}\right)\left(\mathbf{Y^*_{t\,(r)}} - \mathbf{x_t}\,\boldsymbol{\beta}\right)' \cdot I(\tau_{t(r)} = s)\right]$$

$$\pi_s = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{R_{s,t}}{R}\right)$$

Where $R_{s,t} = \sum_{r=1}^{R} I(\tau_{t(r)} = s)$

Essentially, $\boldsymbol{\beta}_\mathbf{s}$ is GLS that uses data falling into regime "s".

$\boldsymbol{\Sigma}_\mathbf{s}$ is the sample covariance matrix of residuals with data falling into regime "s".

π_s is the average frequency of regime "s" occurs.

Last but not least, we briefly outline the Bayesian approach to estimate the model.

The proper priors are assumed to be:

$\boldsymbol{\beta}_\mathbf{s} \sim \text{MVN}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_\mathbf{s}}, \mathbf{V}_{\boldsymbol{\beta}_\mathbf{s}}\right)$

$\boldsymbol{\Sigma}_\mathbf{s}^{-1} \sim \text{Wishart}\,(\boldsymbol{\Omega}_\mathbf{s}, k_s)$

$\boldsymbol{\pi} \sim \text{Dirichlet}\,(\alpha_1, \ldots, \alpha_m)$

The full set of posterior conditionals can be shown to take the following form:

For each $s = 1, \ldots, S$

$\boldsymbol{\beta}_\mathbf{s}|\cdot \sim \text{MVN}(\mathbf{Dd}, \mathbf{D})$

where $\mathbf{D} = \left[\sum_{t=1}^{T}\mathbf{x}_\mathbf{t}'\,\boldsymbol{\Sigma}^{-1}\,\mathbf{x_t} \cdot I(\tau_t = s) + \mathbf{V}_\boldsymbol{\beta}^{-1}\right]^{-1}$

$\mathbf{d} = \left[\sum_{t=1}^{T}\mathbf{x}_\mathbf{t}'\,\boldsymbol{\Sigma}^{-1}\,\mathbf{Y^*_t} \cdot I(\tau_t = s) + \mathbf{V}_\boldsymbol{\beta}^{-1}\,\boldsymbol{\mu}_\boldsymbol{\beta}\right]$

$$\boldsymbol{\Sigma_s^{-1}}|\cdot \sim \text{Wishart}\left\{\left[\boldsymbol{\Omega_s^{-1}} + \sum_{t=1}^{T}(\mathbf{Y_t^*} - \mathbf{x_t\beta})(\mathbf{Y_t^*} - \mathbf{x_t\beta})' \cdot I(\tau_t = s)\right]^{-1},\ k + T_s\right\}$$

where $T_s = \sum_{t=1}^{T} I(\tau_t = s)$

$$\mathbf{Y_{ut}^*}|\cdot \equiv \mathbf{Y_{ut}}$$

$$\mathbf{Y_{ct}^*}|\cdot \sim \text{MVN}_{(-\infty,\mathbf{0})}(\bar{\boldsymbol{\mu}}_\mathbf{c}, \bar{\boldsymbol{\Sigma}}_\mathbf{c})$$

where $\bar{\boldsymbol{\mu}}_\mathbf{c} = \sum_{s=1}^{S}\left[\mathbf{X_{ct}}\,\boldsymbol{\beta}_{\mathbf{s,c}} + \boldsymbol{\Sigma}_{\mathbf{s,cu}}\,\boldsymbol{\Sigma}_{\mathbf{s,uu}}^{-1}\left(\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\,\boldsymbol{\beta}_{\mathbf{s,u}}\right)\right] \cdot I(\tau_t = s)$

$$\bar{\boldsymbol{\Sigma}}_\mathbf{c} = \sum_{s=1}^{S}\left(\boldsymbol{\Sigma}_{\mathbf{s,cc}} - \boldsymbol{\Sigma}_{\mathbf{s,cu}}\,\boldsymbol{\Sigma}_{\mathbf{s,uu}}^{-1}\,\boldsymbol{\Sigma}_{\mathbf{s,uc}}\right) \cdot I(\tau_t = s)$$

$\tau_t|\cdot$ is discrete, $\text{pr}(\tau_t = s|\cdot) = \frac{\pi_s \cdot \phi(\mathbf{Y_t^*};\, \mathbf{X_t\beta_s}, \boldsymbol{\Sigma_s})}{\sum_{j=1}^{S} \pi_j \cdot \phi(\mathbf{Y_t^*};\, \mathbf{X_t\beta_j}, \boldsymbol{\Sigma_j})}$

$$\boldsymbol{\pi}|\cdot \sim \text{Dirichlet}\,[\alpha_1 + T_1, \ldots, \alpha_m + T_m]$$

6. Conclusions

In this paper, we studied generalized SUR model with two features: firstly, some of the dependent variables are censored; secondly the disturbance terms deviate from normality. The key ingredients added to make the model tractable are two types of latent variables: latent dependent variables and latent scale of disturbances. Three approaches are developed to estimate the model, namely Maximum Simulated Likelihood, Expectation-Maximization, and Bayesian Posterior Simulators. All of them involve some simulation techniques and thus non-trivial in computation. The ML estimation requires evaluation of the c.d.f. of a type of generalized multivariate t distribution. An adapted GHK simulator is developed to approximate that integral. The EM algorithm leaves complete-data likelihood without close-form expectation. A Markov Chain based Monte Carlo E-step is designed to approximate the expectation. In fact, it turns out that the same MCMC procedure can be lent to Bayesian posterior simulators to obtain draws of latent variables. The only difference between the EM and Bayesian estimators is that the M-step produces a point (updated) estimate of regressor coefficients, covariance matrix and the degree of freedom parameter, but the Bayesian approach offers posterior distribution of every variate in the model.

The three methods are compared in a series of generated data experiments. The main findings are Bayesian approach is exceedingly accurate, fast and stable. Examined both individually and as a whole, the Bayesian estimators are smallest in terms of MSE, and almost never exhibit anomalies even though the starting values wander far away from the truth. Another advantage of Bayesian estimator over ML and EM is computation speed — not only very fast but also controllable and predictable.

An application of the model is provided as well. High frequency financial data, featuring censoring and non-normality, are used to study to dynamic interaction between security price and trading volume. Using our model, the price, to some extent, is indeed predictable, and therefore it justifies the usefulness of technical analysis in high frequency data.

As a matter of fact, the SUR Tobit model with mixture error distributions has many variants. In the main text, we only discuss the scale-mixtures, and in the Extensions, we briefly outline the finite Gaussian regime mixtures. And much more… leaves unexplored yet. The mixture model could generalize the existing model in various directions and provides better descriptions to the data, and might be interesting for future studies.

References

Amemiya, Takeshi, "Regression Analysis When the Dependent Variable Is Truncated Normal," Econometrica 41 (1973), 997-1016.

Amemiya, Takeshi, "Multivariate Regression and Simultaneous Equation Models When the Dependent Variables Are Truncated Normal," Econometrica 42:6 (1974), 999-1012.

Arias, Carlos, and Thomas Cox, "Estimation of a US Dairy Sector Model by Maximum Simulated Likelihood," Applied Economics 33 (2001), 1201-1211.

Arslan, Olcay, "Family of Multivariate Generalized T Distributions," Journal of Multivariate Analysis 89 (2004), 329-337.

Birnbaum, Z.W., and P.L. Meyer, "On the Effect of Truncation in Some or All Co-ordinates of a Multi-normal Population," Journal of the Indian Society of Agricultural Statistics 5 (1953), 17-28.

Borsch-Supan, Axel, and Vassilis Hajivassiliou, "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models," Journal of Econometrics 58 (1993), 347-368.

Brock, William, Josef Lakonishok, and Blake LeBaron, "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns," Journal of Finance 47:5 (1992), 1731-64.

Butler, J.S., and Robert Moffitt, "a Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model," Econometrica 50 (1982), 761-764.

Campbell, John, Andrew Lo, and Craig MacKinlay, the Econometrics of Financial Markets (Princeton: Princeton University Press, 1997).

Chib, Siddhartha, and Edwards Greenberg, "Hierarchical Analysis of SUR Models with Extensions to Correlated Serial Errors and Timevarying Parameter Models," Journal of Econometrics 68 (1995), 339-60.

Chib, Siddhartha, and Edwards Greenberg, "Bayesian Analysis of the Multivariate Probit Model," Biometrika 85 (1998), 347-361.

Cornick, Jorge, Thomas Cox, and Brain Gould, "Fluid Milk Purchases: a Multivariate Tobit Analysis," American Journal of Agricultural Economics 76:1 (1994), 74-82.

Dempster, Arthur, Nan Laird, and Donald Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," Journal of the Royal Statistical Society 39:1 (1977), 1-38.

Fama, E., "the Behavior of Stock Market Prices," Journal of Business 38 (1965), 34-105.

Genz, Alan, "Numerical Computation of Rectangular Bivariate and Trivariate Normal and T Probabilities," Statistics and Computing. Vol. 14, No 3 (2004), 251-260.

Geweke, Johns, "Efficient Simulation from the Multivariate Normal and Student T Distribution Subject to Linear Constraints". In Computer Science and Statistics Proceedings of the Twenty-Third Symposium on the Inference, American Statistical Association, (1991), 571-578.

Heien, Dale, and Cathy Wessells, "Demand Systems Estimation with Microdata: a Censored Regression Approach," Journal of Business and Economic Statistics 8 (1990), 365-71.

Huang, Cliff, Frank Sloan, and Killard Adamache, "Estimation of Seemingly Unrelated Tobit Regressions via the EM Algorithm," Journal of Business and Economic Statistics 5 (1987), 425-430.

Huang, River, "Estimation of the SUR Tobit Model via the MCECM Algorithm," Economics Letters 64 (1999), 25-30.

Huang, River, "Bayesian Analysis of the SUR Tobit Model," Applied Economics Letters 8 (2001), 617- 622.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud, "Simulation of Multivariate Normal Rectangle Probabilities and Their Derivatives Theoretical and Computational Results," Journal of Econometrics 72 (1996), 85-134.

Hamilton, James, Time Series Analysis (Princeton: Princeton University Press, 1994).

Kao, Chihwa, Lung-fei Lee, and Mark Pitt, "Simulated Maximum Likelihood Estimation of the Linear Expenditure System with Binding Non-Negativity Constraints," Annals of Economics and Finance 2 (2001), 215-55.

Keane, Michael, "A Computationally Practical Simulation Estimator for Panel Data," Econometrica 62 (1994), 95-116.

Lee, Lung-Fei, and Mark Pitt, "Microeconomic Demand Systems with Binding Nonnegativity Constraints: the Dual Approach," Econometrica 5 (1986), 1237-42.

Lo, Andrew, Harry Mamaysky, and Jiang Wang, "Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation," Journal of Finance 55:4 (2000), 1705-1770.

Newcomb, S., "A Generalized Theory of the Combination of Observations So as to Obtain the Best Result," American Journal of Mathematics 8 (1886), 343-366.

Nobile, Agostino, "Comment: Bayesian Multinomial Probit Models with a Normalization Constraint," Journal of Econometrics 99:2 (2000), 335-345.

Peiro, Amado, "the Distribution of Stock Returns: International Evidence," Applied Financial Economics 4:6 (1994), 431-39.

Praetz, Peter, "The Distribution of Share Price Changes," Journal of Business 45 (1972), 49-55.

Pudney, Stephen, Modelling Individual Choice: the Econometrics of Corners, Kinks and Holes (New York: Basil-Blackwell, 1989).

Schnedler, Wendelin, "Likelihood Estimation for Censored Random Vectors," Econometric Reviews 24:2 (2005), 195-217.

Shonkwiler, Scott, and Steven Yen, "Two-step Estimation of a Censored System of Equations," American Journal of Agricultural Economics 81 (1999), 972-82.

Su, Shew-Jiuan, and Steven Yen, "A Censored System of Cigarette and Alcohol Consumption," Applied Economics 32:6 (2000), 729-37.

Tallis, G.M., "The Moment Generating Function of the Truncated Multinormal Distribution," Journal of the Royal Statistical Society 23 (1961), 223-229.

Taylor, Mark, and Helen Allen, "The Use of Technical Analysis in the Foreign Exchange Market," Journal of International Money and Finance 11:3 (1992), 304-314.

Wales, T.J., and Alan Woodland, "Estimation of Consumer Demand Systems with Binding Non-Negativity Constraints," Journal of Econometrics 21 (1983), 263-285.

Yen, Steven, and Biing-Hwan Lin, "Quasi-maximum Likelihood Estimation of a Censored Equation System with a Copula Approach: Meat Consumption by US Individuals," Agricultural Economics 39 (2008), 207-217.

Appendix

A1: Latent scale mixtures: conditionals and marginals

All of the three estimation methods proposed in this paper are based on the following lemma.

Lemma:

Let τ be a scalar variate such that $\tau \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$

Let $\mathbf{Y}$ be a *m*-dimension random vector such that $\mathbf{Y}|\tau \sim \mathrm{MVN}(\boldsymbol{\mu}, \tau\boldsymbol{\Sigma})$

Then $\tau|\mathbf{Y} \sim \mathrm{IG}\left\{\frac{\nu+m}{2}, \left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\right]^{-1}\right\}$

$\mathbf{Y} \sim \mathrm{MVT}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$

Proof:

$$p(\tau|\mathbf{Y}) \propto p(\tau)\cdot p(\mathbf{Y}|\tau)$$

$$\propto \tau^{-\left(\frac{\nu}{2}+1\right)}\cdot \exp\left(-\frac{\nu}{2}\tau^{-1}\right)\cdot \tau^{-\frac{m}{2}}\cdot \exp\left[-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\tau^{-1}\right]$$

$$\propto \tau^{-\left(\frac{\nu+m}{2}+1\right)}\cdot \exp\left\{-\left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\right]\cdot\tau^{-1}\right\}$$

So we recognize that $\tau|\mathbf{Y} \sim \mathrm{IG}\left\{\frac{\nu+m}{2}, \left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\right]^{-1}\right\}$

$$p(\mathbf{Y}) \propto \int p(\tau)\cdot p(\mathbf{Y}|\tau)\cdot d\tau$$

$$\propto \int \tau^{-\left(\frac{\nu+m}{2}+1\right)}\cdot \exp\left\{-\left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\right]\cdot\tau^{-1}\right\}\cdot d\tau$$

The integrand is proportional to aforementioned IG density, so the integral must be the reciprocal of the proportional constant.

$$p(\mathbf{Y}) \propto \left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\right]^{-\frac{\nu+m}{2}}$$

$$\propto \left[1 + \frac{1}{\nu}(\mathbf{Y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}-\boldsymbol{\mu})\right]^{-\frac{\nu+m}{2}}$$

So we recognize that $\mathbf{Y} \sim \mathrm{MVT}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$

A2: Proof of MSL components in Section 2.2

We want to show the following:

$\mathbf{Y}^*_{\mathbf{ut}} \sim \mathrm{MVT}(\mathbf{X}_{\mathbf{ut}}\,\boldsymbol{\beta}_{\mathbf{u}}\,;\,\boldsymbol{\Sigma}_{\mathbf{uu}}, \nu)$

$\tau_t|\mathbf{Y}^*_{\mathbf{ut}} \sim \mathrm{IG}\left\{\frac{\nu+m_u}{2}, \left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y}^*_{\mathbf{ut}} - \mathbf{X}_{\mathbf{ut}}\,\boldsymbol{\beta}_{\mathbf{u}})'\boldsymbol{\Sigma}^{-1}_{\mathbf{uu}}(\mathbf{Y}^*_{\mathbf{ut}} - \mathbf{X}_{\mathbf{ut}}\,\boldsymbol{\beta}_{\mathbf{u}})\right]^{-1}\right\}$

$\mathbf{Y}^*_{\mathbf{ct}}|\,\mathbf{Y}^*_{\mathbf{ut}}, \tau_t \sim \mathrm{MVN}\left[\,\mathbf{X}_{\mathbf{ct}}\,\boldsymbol{\beta}_{\mathbf{c}} + \boldsymbol{\Sigma}_{\mathbf{cu}}\boldsymbol{\Sigma}^{-1}_{\mathbf{uu}}(\mathbf{Y}^*_{\mathbf{ut}} - \mathbf{X}_{\mathbf{ut}}\,\boldsymbol{\beta}_{\mathbf{u}}), \tau_t\left(\boldsymbol{\Sigma}_{\mathbf{cc}} - \boldsymbol{\Sigma}_{\mathbf{cu}}\boldsymbol{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\Sigma}_{\mathbf{uc}}\right)\,\right]$

Proof:

Since $\begin{pmatrix}\mathbf{Y_{ut}}\\ \mathbf{Y_{ct}}\end{pmatrix} | \tau_t \sim \mathrm{MVN}\left[\begin{pmatrix}\mathbf{X_{ut}}\,\boldsymbol{\beta_u}\\ \mathbf{X_{ct}}\,\boldsymbol{\beta_c}\end{pmatrix}, \tau_t \begin{pmatrix}\boldsymbol{\Sigma_{uu}} & \boldsymbol{\Sigma_{uc}}\\ \boldsymbol{\Sigma_{cu}} & \boldsymbol{\Sigma_{cc}}\end{pmatrix}\right]$

So we have $\mathbf{Y_{ut}} | \tau_t \sim \mathrm{MVN}(\mathbf{X_{ut}}\,\boldsymbol{\beta_u}, \tau_t \boldsymbol{\Sigma_{uu}})$

The latent scale $\tau_t \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$. By the previous Lemma in Section A1, we have

$\mathbf{Y_{ut}^*} \sim \mathrm{MVT}(\mathbf{X_{ut}}\,\boldsymbol{\beta_u}\,;\, \boldsymbol{\Sigma_{uu}}, \nu)$

$\tau_t | \mathbf{Y_{ut}^*} \sim \mathrm{IG}\left\{\frac{\nu + m_u}{2}, \left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\,\boldsymbol{\beta_u})' \boldsymbol{\Sigma_{uu}^{-1}} (\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\,\boldsymbol{\beta_u})\right]^{-1}\right\}$

As for $\mathbf{Y_{ct}^*} | \mathbf{Y_{ut}^*}, \tau_t$, note that conditional on τ_t , $\mathbf{Y_{ut}}, \mathbf{Y_{ct}}$ are jointly normal. Apply the familiar formula for conditional normal distribution, we have

$\mathbf{Y_{ct}^*} | \mathbf{Y_{ut}^*}, \tau_t \sim \mathrm{MVN}\left[\mathbf{X_{ct}}\,\boldsymbol{\beta_c} + \boldsymbol{\Sigma_{cu}}\boldsymbol{\Sigma_{uu}^{-1}}(\mathbf{Y_{ut}^*} - \mathbf{X_{ut}}\,\boldsymbol{\beta_u}), \tau_t(\boldsymbol{\Sigma_{cc}} - \boldsymbol{\Sigma_{cu}}\boldsymbol{\Sigma_{uu}^{-1}}\boldsymbol{\Sigma_{uc}})\right]$

A3: Proof of M-Step formulae in Section 2.3

The MCMC E-Step is essentially part of the posterior conditionals of Bayesian sampler, which will be shown in Section A4. Here we prove the conditional M-Step formulae.

We have shown the expectation of complete data likelihood.

$$\mathrm{E}[\ln p(\mathbf{Y}, \mathbf{Y}^*, \boldsymbol{\tau} | \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)]$$

$$\approx \frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T}\begin{bmatrix} -\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) - \left(\frac{\nu}{2}+1\right)\ln\tau_{t\,(r)} - \frac{\nu}{2}\tau_{t\,(r)}^{-1} \\ -\frac{m}{2}\ln\tau_{t\,(r)} - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\tau_{t\,(r)}^{-1}\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)'\boldsymbol{\Sigma^{-1}}\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right) \end{bmatrix} + \text{constant}$$

Treat $\boldsymbol{\Sigma}, \nu$ as given, taking derivatives w.r.t. , we have

$$\frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T}\left[-\frac{1}{2}\tau_{t\,(r)}^{-1}\,\mathbf{x_t'}\,\boldsymbol{\Sigma^{-1}}\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)\right] = 0$$

$$\text{i.e. } \left[\sum_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\right)\mathbf{x_t'}\cdot\boldsymbol{\Sigma^{-1}}\cdot\mathbf{x_t}\right]^{-1}\cdot\boldsymbol{\beta} = \sum_{t=1}^{T}\left[\mathbf{x_t'}\cdot\boldsymbol{\Sigma^{-1}}\cdot\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\,\mathbf{Y_{t\,(r)}^*}\right)\right]$$

$$\boldsymbol{\beta} = \left[\sum_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\right)\mathbf{x_t'}\cdot\boldsymbol{\Sigma^{-1}}\cdot\mathbf{x_t}\right]^{-1}\cdot\sum_{t=1}^{T}\left[\mathbf{x_t'}\cdot\boldsymbol{\Sigma^{-1}}\cdot\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\,\mathbf{Y_{t\,(r)}^*}\right)\right]$$

Treat $\boldsymbol{\beta}, \nu$ as given, taking derivatives w.r.t. $\boldsymbol{\Sigma}$ (a symmetric matrix), we have

$$\frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T}\left[-\frac{1}{2}\boldsymbol{\Sigma^{-1}} - \frac{1}{2}\tau_{t\,(r)}^{-1}(-1)\boldsymbol{\Sigma^{-1}}\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)'\boldsymbol{\Sigma^{-1}}\right] = 0$$

$$\text{i.e. } \mathrm{T\,R} = \sum_{r=1}^{R}\sum_{t=1}^{T}\left[\tau_{t\,(r)}^{-1}\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)'\right]\cdot\boldsymbol{\Sigma^{-1}}$$

$$\boldsymbol{\Sigma} = \frac{1}{T\,R}\sum_{r=1}^{R}\sum_{t=1}^{T}\left[\tau_{t\,(r)}^{-1}\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)\left(\mathbf{Y_{t\,(r)}^*} - \mathbf{x_t}\,\boldsymbol{\beta}\right)'\right]$$

However, Taking derivatives w.r.t. ν, we cannot get explicit solution. We must use numerical methods to optimize:

$$\nu = \arg\max \sum_{t=1}^{T}\left[-\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) - \left(\frac{\nu}{2}+1\right)\left(\frac{1}{R}\sum_{r=1}^{R}\ln\tau_{t\,(r)}\right) - \frac{\nu}{2}\left(\frac{1}{R}\sum_{r=1}^{R}\tau_{t\,(r)}^{-1}\right)\right]$$

A4: Proof of Bayesian posterior conditionals in Section 2.4

$$p(\boldsymbol{\beta}|\, \mathbf{Y}, \boldsymbol{\Sigma}, \nu, \mathbf{Y}^*, \boldsymbol{\tau})$$
$$\propto p(\boldsymbol{\beta}) \cdot p(\mathbf{Y}^*|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$$
$$\propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})' \mathbf{V}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})\right] \cdot \exp\left[\sum_{t=1}^{T}\left(-\frac{1}{2}\right)\tau_t^{-1}(\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})\right]$$

Complete the squares w.r.t. , the resulting expression is still the kernel of multivariate normal density.
$$\boldsymbol{\beta}|\, \mathbf{Y}, \boldsymbol{\Sigma}, \nu, \mathbf{Y}^*, \boldsymbol{\tau} \sim \mathrm{MVN}(\mathbf{Dd}, \mathbf{D})$$

where $= \left[\sum_{t=1}^{T} \tau_t^{-1} \cdot \mathbf{x}_t' \boldsymbol{\Sigma}^{-1} \mathbf{x}_t + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\right]^{-1}$, $\mathbf{d} = \left[\sum_{t=1}^{T} \tau_t^{-1} \cdot \mathbf{x}_t' \boldsymbol{\Sigma}^{-1} \mathbf{Y}_t^* + \mathbf{V}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}}\right]$

The result is in line with Lindley and Smith (1972).

Next, consider $\boldsymbol{\Sigma}^{-1}$
$$p\left(\boldsymbol{\Sigma}^{-1}|\, \mathbf{Y}, \boldsymbol{\beta}, \nu, \mathbf{Y}^*, \boldsymbol{\tau}\right)$$
$$\propto p\left(\boldsymbol{\Sigma}^{-1}\right) \cdot p(\mathbf{Y}^*|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$$
$$\propto \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{k-m-1}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}^{-1}\right)\right] \cdot \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{T}{2}} \cdot \exp\left[\sum_{t=1}^{T}\left(-\frac{1}{2}\right)\tau_t^{-1}(\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})\right]$$
$$= \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{k+T-m-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\left(\boldsymbol{\Omega}^{-1} + \sum_{t=1}^{T} \tau_t^{-1}(\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})(\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})'\right) \cdot \boldsymbol{\Sigma}^{-1}\right]\right\}$$

So we recognize that
$$\boldsymbol{\Sigma}^{-1}|\, \mathbf{Y}, \boldsymbol{\beta}, \nu, \mathbf{Y}^*, \boldsymbol{\tau} \sim \text{Wishart}\left\{\left[\boldsymbol{\Omega}^{-1} + \sum_{t=1}^{T} \tau_t^{-1} \cdot (\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})(\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})'\right]^{-1}, \ T+k\right\}$$

Next, simply apply the Lemma in Section A.1
$$\tau_t|\, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu, \mathbf{Y}^* \sim \mathrm{IG}\left\{\frac{\nu+m}{2}, \left[\frac{\nu}{2} + \frac{1}{2}(\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t^*-\mathbf{x}_t\boldsymbol{\beta})\right]^{-1}\right\}$$

As for $\mathbf{Y}_{\mathbf{ct}}^*|\cdot$, note two points. Firstly, conditional on τ_t , $\mathbf{Y}_{\mathbf{ut}}^*, \mathbf{Y}_{\mathbf{ct}}^*$ are jointly normal. So we can use the formula for conditional normal distribution. Secondly, the knowledge of $\mathbf{Y}^*$ make the sign of $\mathbf{Y}^*$ clear. $\mathbf{Y}_{\mathbf{ct}}^*$ must be non-positive due to censoring at zero. So we have:
$$\mathbf{Y}_{\mathbf{ct}}^*|\, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu, \boldsymbol{\tau} \sim \mathrm{MVN}_{(-\infty, \mathbf{0})}\left[\mathbf{X}_{\mathbf{ct}}\boldsymbol{\beta}_{\mathbf{c}} + \boldsymbol{\Sigma}_{\mathbf{cu}}\boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}(\mathbf{Y}_{\mathbf{ut}}^* - \mathbf{X}_{\mathbf{ut}}\boldsymbol{\beta}_{\mathbf{u}}) \ , \ \tau_t\left(\boldsymbol{\Sigma}_{\mathbf{cc}} - \boldsymbol{\Sigma}_{\mathbf{cu}}\boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\Sigma}_{\mathbf{uc}}\right)\right]$$

Lastly, consider $\nu|\cdot$
$$p(\nu|\, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu, \mathbf{Y}^*, \boldsymbol{\tau})$$
$$\propto p(\nu) \cdot p(\boldsymbol{\tau}|\, \nu)$$
$$\propto \prod_{t=1}^{T}\left[\Gamma\left(\frac{\nu}{2}\right) \cdot \left(\frac{2}{\nu}\right)^{\nu/2}\right]^{-1} (\tau_t)^{-\left(\frac{\nu}{2}+1\right)} e^{-\frac{\nu}{2}\tau_t^{-1}} \cdot \mathrm{I}\left(\underline{\nu} < \nu < \bar{\nu}\right)$$

The expression is so complicated that it does not belong to any known distribution family. Rejection Sampling or Metropolis-Hasting sampler is needed to take draws from that posterior.