Robust Factor Analysis Using the Multivariate t-Distribution

by

J. Li, C. Liu and J. Zhang
Purdue University

Technical Report #07-11

Department of Statistics
Purdue University

October 2007

Robust Factor Analysis Using the Multivariate t-Distribution

Jia Li, Chuanhai Liu, and Jianchun Zhang

Department of Statistics, Purdue University

Abstract: Factor analysis model is often criticized for its lack of robustness. The most popular factor analysis model assumes normal distribution for the error terms, which makes fitted models sensitive to outliers. For robust estimation of factor analysis models, we replace the normal distribution by the multivariate t-distribution. The t-distribution provides a useful extension of the normal for modeling data sets involving errors with heavy tails. The extension of the Gaussian factor analysis model to the student-t factor analysis model is obtained in such a way that the joint distribution of the response variable is a multivariate t-distribution. We develop methods for both the maximum likelihood estimation and the Bayesian estimation of the factor analysis model using the multivariate t-distribution. The proposed methods include the ECME and PX-EM algorithms for maximum likelihood estimation and data augmentation in the Bayesian framework. Numerical examples show that use of multivariate t-distribution improves significantly not only the robustness but also the efficiency.

Key words and phrases: Bayesian Methods; Data Augmentation; EM-type Algorithms; Maximum Likelihood.

1. Introduction

Factor analysis (FA) as a popular statistical method to analyze the underlying relations among multivariate random variables has been extensively used in many areas. The starting point is a linear model in which the observed variables are expressed as linear functions of a vector of unobservable factors and the usual random "errors". The number of underlying factors is strictly less than the number of observed variables. The most commonly used FA model for continuous response variables, namely the Gaussian FA (GFA) model, can be described as follows (see, *e.g.*, Liu and Rubin (1998)):

$$y_i = \mu + \beta z_i + \varepsilon_i, \quad i = 1, ..., n, \tag{1.1}$$

where y_i is the p-dimensional column vector representing the i^{th} observation,

μ is a p-dimensional column vector playing the role of the location, β is the $p \times q$ $(q < p)$ factor loading matrix, z_i is a q-dimensional vector of unobserved factor scores, and $z_i \sim N_q(0, I_q)$, where I_q denotes the $q \times q$ identity matrix. The error term $\varepsilon_i \sim N_p(0, \Psi)$, and $\Psi = \text{Diag}(\psi_1^2, ..., \psi_p^2)$ is a diagonal matrix. The parameters to be estimated are $\theta = (\mu, \beta, \Psi)$.

Since the unobserved factor scores and errors in GFA are assumed to be Gaussian, the usual maximum likelihood or Bayesian estimation is not robust to outliers in the data. The classical technique can be summarized in two separate steps: (a) computing the sample covariance matrix or the sample correlation matrix; (b) making inference based on the matrix obtained in the first step. This approach is not robust to outliers since they have a large effect on the estimate of the covariance matrix obtained in the first step. To reduce the effects of the outliers, robust methods have been considered by researchers. There are two main streams of robust estimation methods for FA models. One is the classical approach, *i.e.*, to get robust estimates of the covariance matrix, and the other is to replace the normal distribution by longer-tailed distributions to accommodate outliers.

The idea in the first stream of robust estimation is to compute highly resistant matrices. Hayashi and Yuan (2003) combines the work of Press and Shigemasu (1997) and Yuan (2000). Before applying the Bayesian inference, they perform a robust transformation to get the M-estimator of (μ, Ψ). This procedure leads to a more accurate evaluation of the factor structure when data have significant skewness and kurtosis. Pison and Rousseeuw (2003) proposed another estimator for the covariance matrix named minimum covariance determinant (MCD) estimator, followed by a principal factor analysis. Actually, both of the estimators belong to the affine equivariant estimators with high breakdown point introduced in Rousseeuw (1983).

The idea in the second stream is to modify the normality assumption on the data. Lee and Press (1998) and Polasek (2000) considered the same idea of using the so-called $\epsilon-$contamination model by assuming that the contaminations follow a different normal distribution. That is, the model is a mixture factor model $f(Y) = (1 - \epsilon)p(Y|\theta) + (\epsilon)p(Y|\theta_0)$, where $p(\cdot)$ denotes the normal density function.

Lange *et al.* (1987) proposed an interesting method of dealing with the errors with longer-than-normal-tails distribution. The general idea is to replace the normal distribution by the multivariate t-distribution. The use of the student-t distribution for robust estimation dates back to Andrews and Mallows (1974) and Zellner (1976). In the last decade, the multivariate t-distribution is popular and works very well in practice for the robust estimations in various fields and applications. Liu (1996) studied the Bayesian robust multivariate linear regression with incomplete data by using the multivariate t-distribution. Pinheiro, Liu, and Wu (2001) worked on the robust estimation in the mixed-effects models by replacing the normal assumption for both the random effects vector and the within-subject errors with the t assumption. Liu (2002) extended the method to a robit regression model and showed that the robit model is a useful robust alternative to the probit and logistic models for analyzing binary response data. However, the multivariate t-distribution to factor analysis has not been developed in the literature, although Yuan *et al.* (2002) mentioned that t-distribution can be used for factor analysis.

We propose in this paper the student-t factor analysis (tFA) model that is obtained by replacing the normal assumption with the t-distribution. We show that not only robustness is gained but also the efficiency is improved. We study the maximum likelihood estimation via the ECME algorithm (Liu and Rubin (1994)), which is an extension of the EM algorithm (Dempster *et al.* (1977)) and the ECM algorithm (Meng and Rubin (1993)). Liu and Rubin (1995) showed that the ECME algorithm for the ML estimation of the t-distribution converges substantially faster than the EM algorithm when the degrees of freedom are to be estimated. Our numerical results agrees with the claim of Liu and Rubin (1995). We also consider the Bayesian estimation via the data augmentation (DA) algorithm (Tanner and Wong (1987)), which can be viewed as a stochastic version of the EM algorithm.

In Section 2, for a motivating example, we describe the US bond indexes data set in which the non-normality are clearly present. The tFA model is described in Section 3. In Section 4 we derive the efficient EM-type algorithms for maximum likelihood estimation of the tFA model. In Section 5, we consider the Bayesian estimation of the tFA model. We compare the robustness and efficiency under the tFA to the GFA model in Section 6. Conclusions and a few remarks are given in Section 7.

2. A Motivating Example: US bond indexes data

We consider monthly log-returns of US bond indexes with maturities in 30 years, 20 years, 10 years, 5 years, and 1 year. The data consist of 696 observations from Jan. 1942 to Dec. 1999. It is well-known that financial data are serially correlated. Tsay (2005) fitted the GFA model to the same data. He argued that the original data may be used because the correlation matrix only changes a little after fitting a multivariate ARMA model. To be comparable with Tsay's results, we adjust the data by dividing each component by its sample standard deviation. Figure 2.1 shows the Q-Q normal plots of the five US bond indexes in terms of log-return. Figure 2.3 is the smoothing density plot for log-return of each bond index. Heavy tails are clearly present in all five variables. Also, the p-value of the Shapiro-Wilk test is close to zero for each index. As a result, the normal distribution is not be appropriate for this data set. Instead, we will use the t-distribution in order to capture the pattern of heavy tails. Figure 2.2 shows the Q-Q student-t plots using the estimated degrees of freedom. Evidently, the Q-Q student-t plots support the use of the t-distribution. In Section 6, we compare both the maximum likelihood estimates and Bayesian estimates based on the GFA and tFA models.

3. A Multivariate t Factor Analysis Model

The GFA model (1.1) can be written as:

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \sim \mathbf{N}_{p+q}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} \beta\beta' + \Psi & \beta \\ \beta' & I_q \end{bmatrix} \right), \quad i = 1, \cdots, n. \tag{3.1}$$

The $(y_i', z_i')'$ is the i^{th} sample with z_i missing. For robust estimation of θ, we replace the multivariate normal distribution in (3.1) by the multivariate t-distribution:

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \sim \mathbf{t}_{p+q}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} \beta\beta' + \Psi & \beta \\ \beta' & I_q \end{bmatrix}, v_i \right), \quad i = 1, \cdots, n, \tag{3.2}$$

where v_i represents the multivariate t-distribution degrees-of-freedom (d.f.) for the i^{th} subject. We call model (3.2) the student-t factor analysis (tFA) model.

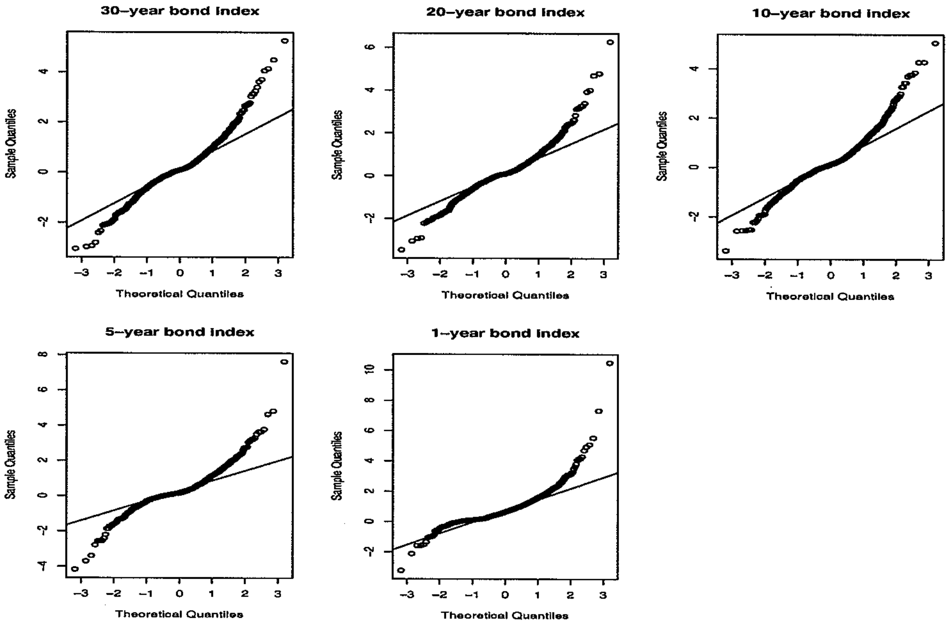

Figure 2.1: Q-Q normal plots for log-return of each US-bond index.

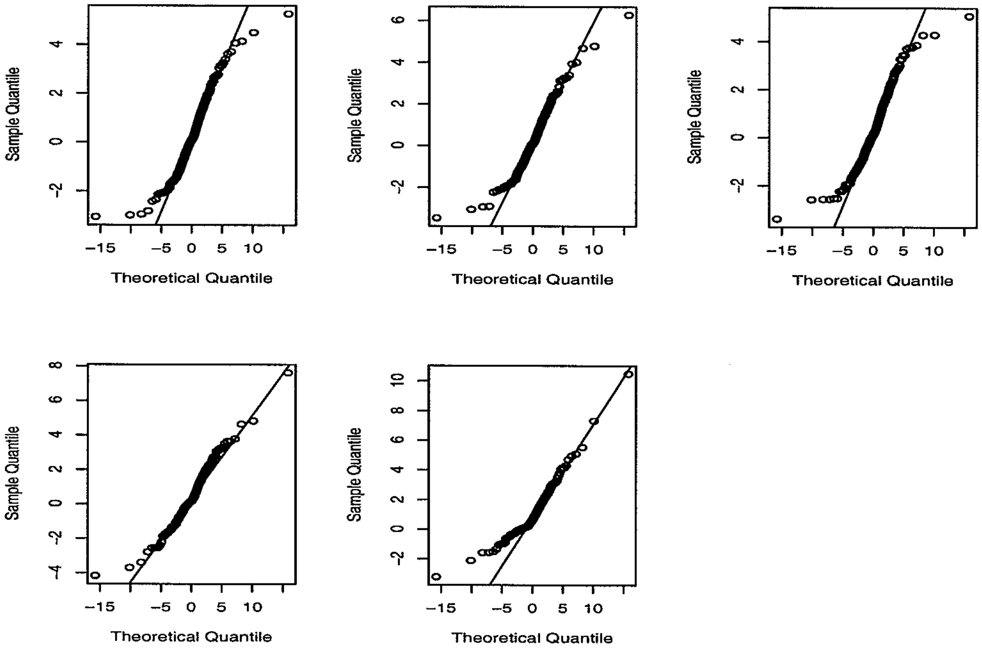

Figure 2.2: Q-Q student-t plots with degrees of freedom 2.5 for log-return of each US-bond index.

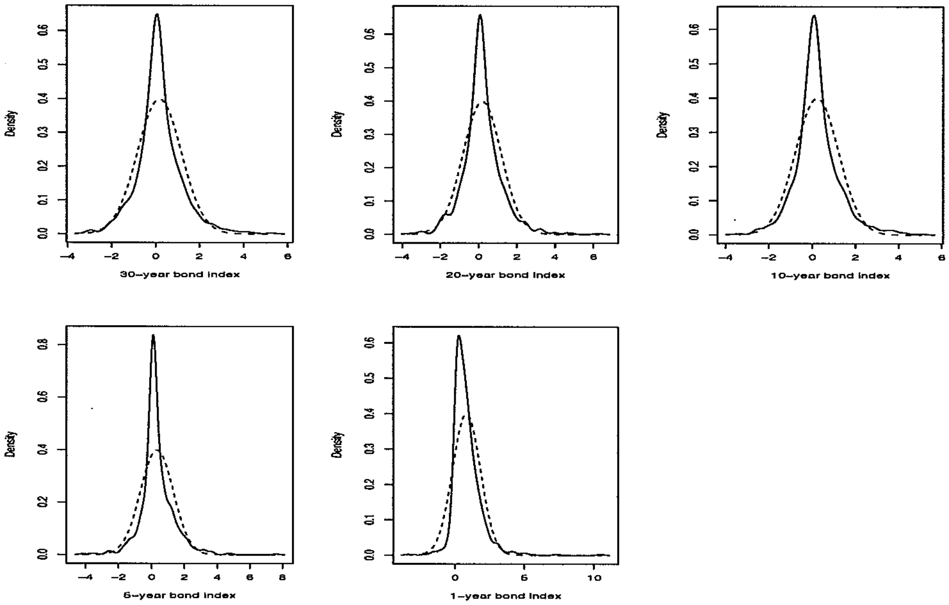

Figure 2.3: smoothing density (solid) for log-return of each US-bond index against normal density (dashed).

The tFA model can also be expressed using the following hierarchical model

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \mid \tau_i \sim \mathbf{N}_{p+q}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \frac{1}{\tau_i}\begin{bmatrix} \beta\beta' + \Psi & \beta \\ \beta' & I_q \end{bmatrix}\right), \quad i = 1, \cdots, n, \tag{3.3}$$

and

$$\tau_i \sim \text{Gamma}\left(\frac{v_i}{2}, \frac{v_i}{2}\right), \quad i = 1, ..., n. \tag{3.4}$$

Or

$$y_i \mid z_i, \tau_i \sim \mathbf{N}\left(\mu + \beta z_i, \frac{1}{\tau_i}\Psi\right), \tag{3.5}$$

$$z_i \mid \tau_i \sim N\left(0, \frac{I_q}{\tau_i}\right), \ \tau_i \sim \text{Gamma}\left(\frac{v_i}{2}, \frac{v_i}{2}\right), \quad i = 1, ..., n, \tag{3.6}$$

where Gamma(a, b) denotes the gamma distribution with shape and rate parameters a and b, defined by the probability density function

$$F(\tau) = b^a \tau^{(a-1)} \exp(-b\tau)/\Gamma(a), \ \ \tau > 0, a > 0, b > 0, \tag{3.7}$$

where $\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t) dt$ denotes the gamma function.

It follows that the tFA model (3.2) can be written as

$$y_i = \mu + \beta z_i + \varepsilon_i, \quad i = 1, ..., n, \tag{3.8}$$

and

$$z_i \sim t_q(0, I_q, v_i), \quad \varepsilon_i \sim t_p(0, \Psi, v_i). \tag{3.9}$$

We assume $v_i = v$ for all i. The marginal distribution of y_i is

$$y_i \mid \tau_i \sim \mathbf{N}\left(\mu, \frac{1}{\tau_i}(\beta\beta' + \Psi)\right), \tag{3.10}$$

and

$$\tau_i \sim \text{Gamma}\left(\frac{v}{2}, \frac{v}{2}\right), \ i = 1, ..., n. \tag{3.11}$$

A useful consequence is that the conditional distribution of τ_i given y_i is

$$\tau_i \mid y_i \sim \text{Gamma}\left(\frac{v + p}{2}, \frac{v + d(Y_i, \mu, \beta\beta' + \Psi)}{2}\right), \tag{3.12}$$

where $d(Y_i, \mu, \beta\beta' + \Psi) = (y_i - \mu)'(\beta\beta' + \Psi)^{-1}(y_i - \mu)$ denotes the Mahanobis distance between y_i and its expectation μ. In particular,

$$E(\tau_i \mid y_i) = \frac{v + p}{v + d(Y_i, \mu, \beta\beta' + \Psi)}. \tag{3.13}$$

4. Efficient ECME Algorithm For ML Estimation

In this section, we consider the maximum likelihood (ML) estimation of the tFA model (3.3). Liu and Rubin (1998) used the ECME algorithm for maximum likelihood estimation of the GFA model. ECME, sharing advantages with both EM and Newton-Raphson algorithms, is an extension of ECM (Meng and Rubin (1993)), which itself is an extension of the EM algorithm. The rate of convergence of ECME, at least judged by the number of iterations, is substantially faster than either EM or ECM, yet it retains stable monotone convergence of EM, and is only modestly more difficult to implement.

The ECM (Expectation, Conditional Maximization) algorithm modifies the EM algorithm by replacing its M step, which maximizes the current expected complete-data log-likelihood over the entire vector parameter θ, by a sequence of conditional maximization steps (indexed by $s = 1, \cdots, S$), each of which maximizes the expected complete-data log-likelihood but over a function of θ, say θ_s, subject to the rest of θ, say $\overline{\theta_s}$, being fixed at their previously estimated values. If the $(\theta_1, \cdots, \theta_S)$ span the parameter space of θ, the ECM algorithm will converge in the same way as EM to an ML estimate. ECME (Expectation, Conditional Maximization or Either) replaces each of one or more of ECM's final CM steps with a step that conditionally maximizes the actual log-likelihood function over θ_s rather than the expected complete-data log-likelihood as with ECM.

4.1 The Identifiability Problem

The factor loading matrix β is not fully identifiable, because it is invariant under transformation of the form $\beta^* = \beta Q$ and $z^* = Q'z$, where Q is any orthogonal $q \times q$ matrix. There are many ways to imposing constraints on β to deal with the indeterminacy. One way is to add the restrictions such that $\Gamma = \beta\Psi^{-1}\beta'$ is diagonal (see, *e.g.*, Anderson (2003)). If the diagonal elements of Γ are ordered and different, β is uniquely determined. Anderson and Rubin (1956) showed that the ML estimators are asymptotically normally distributed under these parameter restrictions. An alternative way is to constrain β so that it is a block lower triangular matrix, assumed to be full rank, with diagonal elements strictly positive (see, *e.g.*, Lopes and West (2004)). We will use the latter when using information matrix to estimate the standard errors of ML estimates. However, when performing ECME, an unrestricted β is assumed, since in the situation without fully identifiable parameters, EM-type algorithms converge to likelihood-equivalent points, which are subject to an orthogonal transformation.

4.2 MLE With Unknown Weights and v via ECME

Let $Y = [y_1, y_2, ..., y_n]'$ be the $n \times p$ data matrix, and let $Z = [z_1, z_2, \cdots, z_n]'$ be the missing data matrix. If Z and $\tau = \{\tau_1, \tau_2, \cdots, \tau_n\}$ were observed, the joint log-likelihood function for the complete data in the tFA model with unknown degrees-of-freedom v is:

$$\mathbf{L}(\mu, \beta, \Psi, v \mid Y, Z, \tau) = L_1(\mu, \beta, \Psi \mid Y, Z, \tau) + L_2(v \mid \tau) + \text{constant}, \qquad (4.1)$$

where

$$\begin{aligned} & L_1(\mu, \beta, \Psi \mid Y, Z, \tau) \\ = \; & -\frac{n}{2}\log|\Psi| - \frac{1}{2}\text{tr}(\Psi^{-1}\sum_{i=1}^{n}\tau_i y_i y_i') + \mu'\Psi^{-1}\sum_{i=1}^{n}\tau_i y_i + \text{tr}(\Psi^{-1}\beta\sum_{i=1}^{n}\tau_i z_i y_i') \\ & -\mu'\Psi^{-1}\beta(\sum_{i=1}^{n}\tau_i z_i) - \frac{1}{2}\text{tr}(\beta'\Psi^{-1}\beta\sum_{i=1}^{n}\tau_i z_i z_i') - \frac{1}{2}\mu'\Psi^{-1}\mu\sum_{i=1}^{n}\tau_i, \end{aligned}$$

and

$$L_2(v \mid \tau) = \frac{vn}{2}\log\frac{v}{2} + \frac{v}{2}\sum_{i=1}^{n}\log\tau_i - \frac{v}{2}\sum_{i=1}^{n}\tau_i - n\log\Gamma(\frac{v}{2}). \qquad (4.2)$$

The sufficient statistics for $L_1(\mu, \beta, \Psi \mid Y, \tau)$ are $S_\tau = \sum_{i=1}^n \tau_i$, $S_{\tau Y} = \sum_{i=1}^n \tau_i y_i$, $S_{\tau Z} = \sum_{i=1}^n \tau_i z_i$, $S_{\tau YY} = \sum_{i=1}^n \tau_i y_i y_i'$, $S_{\tau ZY} = \sum_{i=1}^n \tau_i z_i y_i'$, and $S_{\tau ZZ} = \sum_{i=1}^n \tau_i z_i z_i'$.

Given Ψ, μ and β, (y_i, z_i) are *iid* $(p+q)$-normal. Thus, given Ψ, μ and β, the conditional distribution of z_i given y_i is q-variate normal with mean $\delta(y_i - \mu)$ and covariance Δ, where the regression coefficient δ and residual covariance matrix Δ are given by:

$$\begin{aligned} \delta &= (\frac{1}{\tau_i}\beta')[\frac{1}{\tau_i}(\Psi + \beta\beta')]^{-1} = \beta'(\Psi + \beta\beta')^{-1}, & (4.3) \\ \Delta_i &= \frac{1}{\tau_i}I_q - \frac{1}{\tau_i}\beta'(\Psi + \beta\beta')^{-1}\beta = \frac{1}{\tau_i}\Delta. & (4.4) \end{aligned}$$

ECME algorithm:

E step: Let $\theta^{(t)} = (\mu^{(t)}, \beta^{(t)}, \Psi^{(t)}, v^{(t)})$ be the current estimate of θ. Then $\tau_i^{(t+1)} = E(\tau_i \mid \theta^{(t)}, Y) = \frac{v^{(t)}+p}{v^{(t)}+d(Y_i,\mu^{(t)},\beta^{(t)}\beta^{(t)\prime}+\Psi^{(t)})}$, $\delta^{(t+1)} = \beta^{(t)\prime}(\Psi^{(t)}+\beta^{(t)}\beta^{(t)\prime})^{-1}$, and $\Delta_i^{(t+1)} = \frac{1}{\tau_i^{(t+1)}}I_q - \frac{1}{\tau_i^{(t+1)}}\beta^{(t)\prime}(\Psi^{(t)} + \beta^{(t)}\beta^{(t)\prime})^{-1}\beta^{(t)} = \frac{\Delta^{(t+1)}}{\tau_i^{(t+1)}}$.

These lead to the expectation of the sufficient statistics:

$$
\begin{aligned}
\hat{S}_{\tau}^{(t+1)} &= E(S_\tau \mid \theta^{(t)}, Y) = \sum_{i=1}^{n} \tau_i^{(t+1)}, \\
\hat{S}_{\tau Y}^{(t+1)} &= E(S_{\tau Y} \mid \theta^{(t)}, Y) = \sum_{i=1}^{n} \tau_i^{(t+1)} y_i, \\
\hat{S}_{\tau Z}^{(t+1)} &= E(S_{\tau Z} \mid \theta^{(t)}, Y) \\
&= \sum_{i=1}^{n} \tau_i^{(t+1)} \delta^{(t+1)} (y_i - \mu^{(t)}) = \delta^{(t+1)} (\hat{S}_{\tau Y}^{(t+1)} - \hat{S}_{\tau}^{(t+1)} \mu^{(t)}), \\
\hat{S}_{\tau YY}^{(t+1)} &= E(S_{\tau YY} \mid \theta^{(t)}, Y) = \sum_{i=1}^{n} \tau_i^{(t+1)} y_i y_i', \\
\hat{S}_{\tau ZY}^{(t+1)} &= E(S_{\tau ZY} \mid \theta^{(t)}, Y) \\
&= \sum_{i=1}^{n} \tau_i^{(t+1)} \delta^{(t+1)} (y_i - \mu^{(t)}) y_i' = \delta^{(t+1)} (\hat{S}_{\tau YY}^{(t+1)} - \mu^{(t)} \hat{S'}_{\tau Y}^{(t+1)}),
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{S}_{\tau ZZ}^{(t+1)} &= E(S_{\tau ZZ} \mid \theta^{(t)}, Y) \\
&= \delta^{(t+1)} (\hat{S}_{\tau YY}^{(t+1)} - \hat{S}_{\tau Y}^{(t+1)} \mu^{(t)\prime} - \hat{S'}_{\tau Y}^{(t+1)} \mu^{(t)} + \hat{S}_{\tau}^{(t+1)} \mu^{(t)} \mu^{(t)\prime}) \delta^{(t+1)\prime} \\
&+ n \Delta^{(t+1)}.
\end{aligned}
$$

M step: Rewriting the FA model by combining the mean vector and the factor loading matrix, we get

$$
y_i = \mu + \beta z_i + \varepsilon_i \Longrightarrow y_i = \begin{pmatrix} \mu & \beta \end{pmatrix} \begin{pmatrix} 1 \\ z_i \end{pmatrix} + \varepsilon_i \Longrightarrow y_i = \alpha x_i + \varepsilon_i, \tag{4.5}
$$

where α is a $p \times (q+1)$ matrix, and x_i is a $(q+1) \times 1$ column vector. Then the log-likelihood becomes

$$
\begin{aligned}
& -\frac{n}{2} \log|\Psi| - \frac{1}{2} \mathrm{tr}(\sum_{i=1}^{n} \Psi^{-1} \tau_i (y_i - \alpha x_i)(y_i - \alpha x_i)') \\
= & -\frac{n}{2} \log|\Psi| - \frac{1}{2} \mathrm{tr}(\Psi^{-1} S_{\tau YY}) + \mathrm{tr}(\Psi^{-1} \alpha S_{\tau XY}) - \frac{1}{2} \mathrm{tr}(\Psi^{-1} \alpha S_{\tau XX} \alpha'),
\end{aligned}
$$

where

$$
S_{\tau XX} = \sum_{i=1}^{n} \tau_i x_i x_i' = \begin{pmatrix} S_\tau & S'_{\tau Z} \\ S_{\tau Z} & S_{\tau ZZ} \end{pmatrix}, \quad S_{\tau XY} = \sum_{i=1}^{n} \tau_i x_i y_i' = \begin{pmatrix} S'_{\tau Y} \\ S_{\tau ZY} \end{pmatrix}.
$$

From the results in E-step and the standard regression arguments, the MLE of μ, β, and Ψ are updated as follows.

CM step 1: Fix $\Psi^{(t)}$ and update α by maximizing $E[L_1(\mu, \beta, \Psi \mid Y, \tau)]$

$$vec(\alpha^{(t+1)}) = ((\Psi^{(t)} \otimes S^{-1}{}^{(t)}_{\tau XX}) \cdot vec(A))', \tag{4.6}$$

where $A = S^{(t)}_{\tau XY} {\Psi^{(t)}}^{-1}$ and $vec(X)$ denotes the vector formed by stacking the column vectors of the matrix X. The notation $\otimes$ stands for the Kronecker product operator. The detailed algebraic derivation is shown in Appendix.

CM step 2: Fix $\mu^{(t+1)}$, $\beta^{(t+1)}$ and update $\Psi^{(t+1)}$

$$\begin{aligned} \Psi^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} \tau_i^{(t+1)}(y_i - \alpha^{(t+1)}x_i)(y_i - \alpha^{(t+1)}x_i)' \\ &= \frac{1}{n}\text{Diag}(\hat{S}^{(t+1)}_{\tau YY} - 2\alpha^{(t+1)}\hat{S}^{(t+1)}_{\tau XY} + \alpha^{(t+1)}\hat{S}^{(t+1)}_{\tau XX}\alpha^{(t+1)\prime}). \end{aligned}$$

CM step 3: Update $v^{(t+1)}$ by maximizing $E[L_2(v \mid \tau)|Y, \theta^{(t+1)}]$ over v to obtain

$$v^{(t+1)} = \arg\max_v \left\{ \frac{v}{2}\left[n \log\frac{v}{2} + \sum_{i=1}^{n} E[\log \tau_i|y, \hat{\theta}^{(t+1)}] - \hat{S}^{(t)}_{\tau}\right] - n\log\Gamma(\frac{v}{2})\right\}, \tag{4.7}$$

where $\theta^{(t+1)} = \{\alpha^{(t+1)}, \Psi^{(t+1)}, v^{(t)}\}$. Note that finding $v^{(t+1)}$ only requires a one-dimensional search and can be done, for example, using the Newton-Raphson method or the Bisection Method. However, because $E[\log \tau_i|y, \hat{\theta}^{(t+1)}]$ has no closed-form expression, one can use ECME to ease calculation by maximizing the constrained likelihood over v with $\theta^{(t+1)} = \{\alpha^{(t+1)}, \Psi^{(t+1)}, v^{(t)}\}$ being fixed. Since $y_i \sim t_p(\mu^{(t+1)}, \beta^{(t+1)}\beta^{(t+1)\prime} + \Psi^{(t+1)}, v)$ independently for $i = 1, \cdots, n$, we have

CML step 3: Update $v^{(t+1)}$ as

$$v^{(t+1)} = \arg\max_v \left\{ \sum_{i=1}^{n} \left[\log\Gamma(\frac{v+p}{2}) - \log\Gamma(\frac{v}{2}) + \frac{v}{2}\log v - \frac{v+p}{2}\log(v + d_i)\right]\right\}, \tag{4.8}$$

where $d_i = d(y_i, \mu^{(t+1)}, \beta^{(t+1)}\beta^{(t+1)\prime} + \Psi^{(t+1)})$. This step requires only a one-dimensional search.

4.3 PX-ECME Algorithm

Liu, Rubin and Wu (1998) proposed the method of *Parameter Expansion* (PX) to accelerate EM-type algorithm and showed that the PX-EM algorithm shares the simplicity and stability of ordinary EM, but has a faster rate of convergence. Here, we will show how the PX-EM algorithm can be adopted in the context of the tFA model by expanding the covariance matrix of factor score. We call it PX-ECME. We expand the covariance matrix of z_i from the identity matrix to unrestricted covariance matrix (or positive definite matrix) γ. Then the model becomes

$$y_i \mid z_i, \tau_i \sim \mathbf{N}\left(\mu + \beta z_i, \frac{1}{\tau_i}\Psi\right), \tag{4.9}$$

and

$$z_i \mid \tau_i \sim N(0, \frac{\gamma}{\tau_i}),\ \tau_i \sim \text{Gamma}(\frac{v}{2}, \frac{v}{2}) \ ,\ i = 1, \cdots, n. \tag{4.10}$$

PX-ECME algorithm:

PX-E step: This is unchanged from ECME.

PX-CM step: The computations for $\mu^{(t+1)}$ and $\Psi^{(t+1)}$ stay the same as those in ECME. For $\beta^{(t+1)}$, obtain $\beta_*^{(t+1)}$ in the same way as that in ECME, let $\gamma^{(t+1)} = \frac{1}{n}\hat{S}_{\tau ZZ}^{(t+1)}$, and reduce $\beta_*^{(t+1)}$ to the original parameter by setting $\beta^{(t+1)} = \beta_*^{(t+1)} Chol(\gamma^{(t+1)})$, where $Chol(.)$ denotes the Cholesky decomposition.

The PX-ECME algorithm maintains the simplicity and stability properties of ECME algorithm. But it dominates ECME by its fast convergence. A numerical example for comparing PX-ECME and ECME will be given in section 6.2. We will see that PX-ECME still works very well for the tFA model.

5. A Bayesian Approach

In this section, we study the Bayesian estimation of the tFA model via the data augmentation (DA) algorithm (Tanner and Wong (1987)). We will first discuss proper prior distributions for the parameters, and then calculate the conditional posterior distributions. Furthermore, we use data augmentation algorithm to obtain the distributions of the parameters. DA, which can be viewed as a stochastic version of the EM algorithm, iterates between two steps as follows.

DA algorithm:

I step: Impute the missing values drawing from the predictive model given the observed data and the current estimated parameters.

P step: Draw the parameters from their posterior distributions given the current filled-in complete data.

First we assume that the prior distribution for the parameters has the form of

$$\begin{aligned} Pr(\theta) &= Pr(\mu, \beta, \Psi, v) \\ &= Pr(\mu)Pr(\beta, \Psi)Pr(v) \\ &= Pr(\mu)Pr(\beta|\Psi)Pr(\Psi)Pr(v). \end{aligned}$$

The joint density function for the complete data is decomposed as follows:

$$Pr(Y, Z, \tau|\theta) = Pr(Y|Z, \tau, \mu, \beta, \Psi)Pr(Z|\tau)Pr(\tau|v).$$

5.1 Priors and Posteriors of Parameters

We obtain the conditional posterior distributions of the parameters one by one. For the mean vector μ, which is a $p \times 1$ vector, we use the flat prior

$$Pr(\mu) \propto \text{constant}. \tag{5.1}$$

This yields

$$\begin{aligned} Pr(\mu|Y, Z, \tau) &\propto Pr(Y, Z, \tau|\theta)Pr(\theta) \\ &\propto Pr(Y|Z, \tau, \mu, \beta, \Psi)Pr(\mu) \\ &\propto |\Psi|^{-\frac{n}{2}}\exp\{-\frac{1}{2}\text{tr}\Psi^{-1}[\sum_{i=1}^{n} \tau_i(y_i - \mu - \beta z_i)(y_i - \mu - \beta z_i)']. \end{aligned}$$

The conditional posterior distribution of μ is distributed as multivariate normal with mean and covariance

$$\overline{\mu} = \frac{\sum_{i=1}^{n} \tau_i(y_i - \beta z_i)}{S_\tau} = \frac{S_{\tau Y} - \beta S_{\tau Z}}{S_\tau} \text{ and } V = \frac{\Psi}{S_\tau}. \tag{5.2}$$

For the factor loading matrix $\beta = [\beta_1, \beta_2, \cdots, \beta_q]_{p\times q}$, we let

$$\beta_i|\Psi \sim N\left(\beta_0, \frac{\Psi}{n_1}\right), i = 1, \cdots, q, \tag{5.3}$$

where n_1 is a constant. Let $vec(\beta) = [\beta_1', \beta_2', \cdots, \beta_q']'$, then we have

$$vec(\beta)|\Psi \sim N\left(\overline{\beta}_0, I_{q\times q} \otimes \frac{\Psi}{n_1}\right). \tag{5.4}$$

where $\overline{\beta}_0 = [\beta_0', \beta_0', \cdots, \beta_0']'$. And the density is

$$Pr(vec(\beta)|\Psi) \propto |\Psi|^{\frac{q}{2}}\exp\{-\frac{1}{2}(vec(\beta) - \overline{\beta}_0)'(I_{q\times q} \otimes \frac{\Psi}{n_1})^{-1}(vec(\beta) - \overline{\beta}_0)\}. \tag{5.5}$$

The posterior distribution of β conditional on the observations is of the form

$$\begin{aligned} Pr(\beta|Y,Z,\tau) &\propto Pr(Y,Z,\tau|\theta)Pr(\theta) \\ &\propto Pr(Y|Z,\tau,\mu,\beta,\Psi)Pr(\beta|\Psi). \end{aligned}$$

The following result provides details for updating β. The proof of the result is given in Appendix.

Theorem 1 *The conditional posterior distribution of $vec(\beta)$, given Y, Z, τ and Ψ is normal with mean $(D_1+D_2)^{-1}(D_1 vec(\hat{\beta})+D_2\overline{\beta}_0)$, and covariance $(D_1+D_2)^{-1}$, where $\hat{\beta} = (\sum_{i=1}^{n} \tau_i(y_i-\mu)z_i')(S_{\tau ZZ})^{-1} = (S'_{\tau ZY} - \mu S'_{\tau Z})(S_{\tau ZZ})^{-1}$, $D_1 = S'_{\tau ZZ} \otimes \Psi^{-1}$ and $D_2 = n_1 I_{q\times q} \otimes \Psi^{-1}$.*

For the prior distribution of Ψ, we use the inverse Wishart distribution

$$Pr(\Psi) \propto |\Psi|^{-\frac{m+1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Psi^{-1}A)\}, \tag{5.6}$$

where m is a scalar and A is a $(p \times p)$ non-negative definite matrix. If $m = p$ and $A = 0$, then it is the noninformative prior or Jeffrey's prior. If $m = -1$ and $A = 0$, then it is the flat prior. The posterior distribution of Ψ conditional on the observations is of the form

$$\begin{aligned} Pr(\Psi|Y,Z,\tau) &\propto Pr(Y,Z,\tau|\theta)Pr(\theta) \\ &\propto Pr(Y|Z,\tau,\mu,\beta,\Psi)Pr(\beta|\Psi)Pr(\Psi) \\ &\propto |\Psi|^{-\frac{m+n+q+1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Psi^{-1}(A+B+C))\}, \end{aligned}$$

where $B = \sum_{i=1}^{n} \tau_i(y_i-\mu-\beta z_i)(y_i-\mu-\beta z_i)'$, and $C = n_1 \sum_{i=1}^{q}(\beta_i-\beta_0)(\beta_i-\beta_0)'$.

Let $H = A + B + C$ with the diagonal elements $(h_1^2, \cdots, h_p^2)$, and let $d = m + n + q - p - 1$, then

$$\psi_i^{-2} \sim \frac{\chi_d^2}{h_i^2} \quad (i = 1, \cdots, p). \tag{5.7}$$

For the degrees of freedom v, it is either fixed or unknown. A brief discussion on drawing v is given in Liu (1995). In order to obtain a proper posterior of v, the basic rule is that the prior distribution of v should satisfy the condition that $Pr(v) = o(v^{-1})$ as $v \to +\infty$. There are several suggestions (Anscombe (1967), Relles and Rogers (1977), Gewek (1993), and Box and Tiao (1973)) about the prior of v such as $Pr(v) \propto (v+1)^{-\frac{3}{2}}$, $(v \geqslant 1)$, the flat prior distribution for v^{-1}, *i.e.* $Pr(v) \propto v^{-2}$, $(v \geqslant 1)$, and the exponential distribution $Pr(v) = \lambda e^{-\lambda v}$. And the application of Jeffrey's rule leads to the following noninformative prior distribution of the degree of freedom v of the form

$$Pr(v) \propto \left[\phi'(\frac{v}{2}) - \phi'(\frac{v+p}{2}) - \frac{2p(v+p+4)}{v(v+p)(v+p+2)}\right]^{1/2}, \qquad (5.8)$$

where $\phi'(x) = \frac{d^2}{d^2x}\ln(\Gamma(x))$ is the Trigamma function.

The fact that $\tau_i|v \sim \text{Gamma}(\frac{v}{2}, \frac{v}{2})$ leads to the conditional posterior of v:

$$\begin{aligned} Pr(v|Y,Z,\tau) &\propto Pr(Y,Z,\tau|\theta)Pr(\theta) \\ &\propto \exp\{\log(Pr(v)) + n\log\Gamma(\frac{v+p}{2}) - n\log\Gamma(\frac{v}{2}) + \frac{nv}{2}\log(v) \\ &\quad -\sum_{i=1}^{n}\frac{v+p}{2}\log\Gamma(v+d_i)\}. \end{aligned}$$

5.2 The DA Algorithm

This section presents the DA algorithm for taking draws of the parameters of the tFA model from their posterior distribution. For the sake of clarity, we present DA step by step with known and unknown weights and know and unknown degrees of freedom.

When the weights $\tau = \{\tau_1, \tau_2, \cdots, \tau_n\}$ are known, DA is straightforward:

I-step Impute $Z^{(t)}$ with a draw from $Pr(Z|Y, \theta^{(t)}, \tau)$, where $Z^{(t)}|\theta^{(t)}, \tau^{(t)} \sim N(\mu_{z|y}, \frac{\Delta}{\tau^{(t)}})$.

P-step Draw $\mu^{(t+1)} \sim Pr(\mu|Y, Z^{(t)}, \tau)$, $\beta^{(t+1)} \sim Pr(\beta|Y, Z^{(t)}, \tau)$ and draw $\Psi^{(t+1)} \sim Pr(\Psi|Y, Z^{(t)}, \tau)$.

If the weights τ are unknown, the missing data is (Z, τ). From Section 2, we know $\tau^{(t)}|Y, \theta^{(t)} \sim \text{Gamma}(\frac{v+p}{2}, \frac{v+d}{2})$, where d is the Mahanobis distance between y_i and its expectation.

I-step Impute $\tau^{(t)}$ from $Pr(\tau|Y, \theta^{(t)})$. Impute $Z^{(t)}$ from $Pr(Z|Y, \theta^{(t)}, \tau)$ where $Z^{(t)}|\theta^{(t)}, \tau^{(t)} \sim N(\mu_{z|y}, \frac{\Delta}{\tau^{(t)}})$.

P-step Draw $\mu^{(t+1)} \sim Pr(\mu|Y, Z^{(t)}, \tau^{(t)})$, $\beta^{(t+1)} \sim Pr(\beta|Y, Z^{(t)}, \tau^{(t)})$ and draw $\Psi^{(t+1)} \sim Pr(\Psi|Y, Z^{(t)}, \tau^{(t)})$.

When the weights τ are known but v is unknown, each iteration consists of two steps:

I-step Impute $Z^{(t)}$ the same as with known weights and known degree freedom.

P-step Draw $\mu^{(t+1)}$, $\beta^{(t+1)}$, and $\Psi^{(t+1)}$ in the same way as that in the case with known weights and known degree freedom, and draw $v^{(t+1)} \sim Pr(v|Y, Z^{(t)}, \tau)$.

When both the weights and the degree freedom are unknown, the missing data is (Z, τ), the parameters are μ, β, Ψ and v. Comparing with the case with unknown weights and known v, we need an additional step to take a draw of v.

I-step Impute $\tau^{(t)}$ and $Z^{(t)}$ in the same way as the case with unknown weights and known v.

P-step Draw $\mu^{(t+1)}$, $\beta^{(t+1)}$, and $\Psi^{(t+1)}$ in the same way as with unknown weights and known degree freedom. In addition, take a draw $v^{(t+1)} \sim Pr(v|Y, Z^{(t)}, \tau^{(t)})$.

6. Application to analyzing US bond indexes data

In this section, we apply our method for robust factor analysis of the US bond indexes data set, which is described in Section 2. To illustrate our methodology, consider two multivariate exploratory factor analysis models:

- Model N_5, the errors are assumed to follow the Gaussian distribution;
- Model t_5, the errors are assumed to follow the multivariate t-distribution.

To choose appropriate number of factors, we consider the results in Table 6.1. Likelihood ratio test can be used to help select the number of factors. The null hypothesis is the current factor analysis model. It is tested against the alternative in which no factor analysis model is considered, that is, the likelihood is calculated based on the marginal distribution of the observable variable. Under some regularity conditions, the likelihood ratio test statistic has the chi-squared distribution asymptotically with degree-of-freedom $[(p-q)^2 - (p+q)]/2$. Usually, one starts with a small number of factors, say $q = 1$, testing goodness-of-fit until a nonsignificant result occurs. We refer to Jöreskog (1967) and Anderson (2003) for more details about this procedure. However, this procedure is criti-

Model	number of factors	2	3
N_5	max log-likelihood	-2213.66	-2209.66
	Likelihood Ratio	8.02 (d.f.=1)	0.04
	AIC	4465.32	4463.33
	BIC	4551.68	4563.32
t_5	max log-likelihood	-1605.97	-1601.64
	Likelihood Ratio	8.72 (d.f.=1)	0.06
	AIC	3249.94	3247.28
	BIC	3336.28	3347.28

Table 6.1: Test results of model N_5 and t_5 with different number of factors.

cized by Krzanowski and Marriott (1995) because *no adjustment is made to the significance level to allow for its sequential nature.* AIC and BIC are in general considered to be better by taking into account the trade-off between goodness-of-fit and number of parameters. While BIC is thought to be better than AIC because the penalty imposed being related to the sample size. In the above output, the Likelihood ratio and AIC prefer 3 factors under both normal and t assumptions, while BIC prefers 2 factors in both cases. Considering the model parsimony, we choose to focus on 2 factor models. Tasy (2005) also fitted the 2 factor GFA model.

6.1 Comparing the Gaussian and the Multivariate t MLEs

To run ECME, we choose the following initial values: $\mu^{(0)}$ is the sample mean of observed data, $\beta^{(0)}$ is a $p \times q$ matrix with all the components being 1, $\psi^{(0)}$ is the $p \times p$ identity matrix I_p, and $v^{(0)} = 20$. The convergence criterion is that the difference of the log-likelihood between two iterations is less than 10^{-4}. For identifiability of the factor loading β, the estimate of β is rotated in such a way that the upper-right triangle is 0 and the diagonal elements are positive. This rotation makes the comparison meaningful.

The ML estimates of two different FA models are shown in Tables 6.2-6.4. In addition, the estimation of the degree of freedom of the model t_5 is 2.3005 with standard deviation 0.1682. The associated variance-variance matrix was computed via numerical differentiation. Alternative methods (see, *e.g.*, He and Liu (2007)) can be used. The mean vector shifts to left in the model t_5 because the data has heavier tails on the right than on the left. Table 6.3 shows dramatic

N_5	0.1719	0.1865	0.2273	0.3301	0.8298
S.d.	3.808e-2	3.808e-2	3.808e-2	3.809e-2	3.807e-2
t_5	0.1135	0.1269	0.1500	0.2234	0.5706
S.d.	2.490e-2	2.439e-2	2.464e-2	2.353e-2	2.761e-2

Table 6.2: Estimation of mean and their standard deviation.

N_5		S.d.		t_5		S.d.	
0.9979	0	2.742e-2	0	0.5839	0	2.456e-2	0
0.9893	0.0291	2.764e-2	3.072e-2	0.5731	0.0107	2.387e-2	2.365e-2
0.9285	0.2034	1.064e-2	2.101e-2	0.5432	0.1165	0.570e-2	1.524e-2
0.8636	0.5158	2.915e-2	3.435e-2	0.4645	0.2992	2.414e-2	2.572e-2
0.6434	0.5244	1.552e-2	2.960e-2	0.3083	0.3308	1.046e-2	2.385e-2

Table 6.3: Estimation of the factor loading matrix and their standard deviation.

difference between ML estimates of the factor loading matrix under the two different models. Although the estimated factor loading matrices are significantly different in two models, the components in the estimated matrices have a very similar pattern. The factor loadings for the first factor are roughly proportional to the time to bond maturity, whereas the factor loadings of the second factor are inversely proportional to the time to bond maturity.

To make the MLEs comparable, we consider the variance of the error term that is given by $var(\psi) = [v/(v-2)]\sigma^2$. We see that, except for the last three components of ψ, the corresponding approximate standard errors of all the components of μ, all the components of β and the first two components of ψ in model t_5 are consistently smaller than the estimated standard errors in the model N_5. It indicates that the estimation is more accurate under the model t_5.

Lange *et al.* (1989) considered diagnostics to check model assumption. For

N_5	0.0135	0.0299	0.1066	0.0006	0.3192
S.d.	4.478e-3	4.229e-3	6.152e-3	1.541e-2	2.215e-2
t_5	0.0074	0.0044	0.0303	0.0002	0.1463
S.d.	3.963e-3	3.422e-3	7.059e-3	1.709e-2	3.413e-2

Table 6.4: Estimation of the covariance matrix of the error terms and their standard deviation.

the GFA model, a natural measure is the Mahalanobis-like distance $\delta_i^2 = (y_i - \hat{\mu}_i)'(\hat{\beta}\hat{\beta}' + \hat{\Psi})^{-1}(y_i - \hat{\mu}_i)$ which has the asymptotic chi-squared distribution with degree of freedom p. The normality assumption can be checked by transforming each δ_i^2 to an asymptotically standard normal deviate using the well-known cube-root of Wilson and Hilferty or fourth-root transformation. Here, we use the fourth-root transformation (Hawkins and Wixley (1986)), because it performs well when the degree-of-freedom is small. For the tFA model, d_i^2/p has the asymptotic F-distribution with degree-of-freedoms p and v, where $d_i^2 = (y_i - \hat{\mu}_i)'(\hat{\beta}\hat{\beta}' + \hat{\Psi})^{-1}(y_i - \hat{\mu}_i)$. The normality approximation is available by first transforming the numerator and denominator chi-squared deviates in the F-statistic using fourth-root transformation into normal-like deviates, then applying Geary's (1930) approximation to the ratio of normal deviates. The explicit formula is given by Little (1990). Figure 6.4 shows the normal quantile-quantile plots of the two distances under normal and t distributions, respectively. The left panel suggests that the GFA model is inadequate. The plot for the tFA model, with most of the points lying close to the reference line, is much better than that for the normal model. The improvement is apparent. These results support the use of the tFA model.

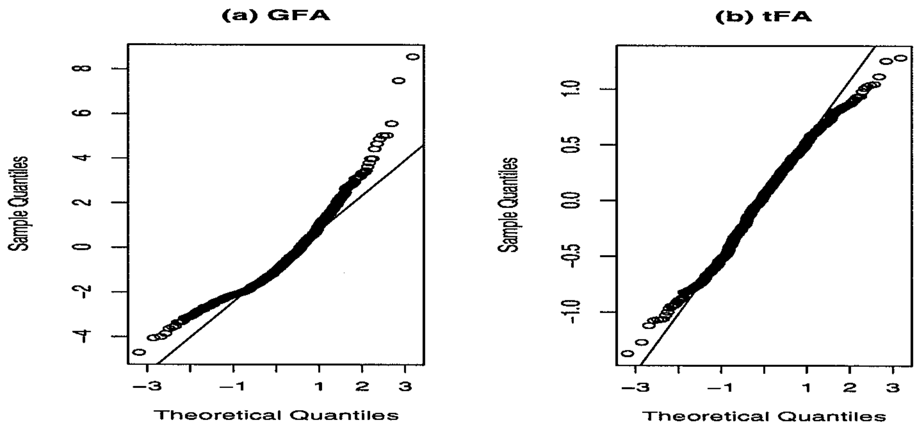

Figure 6.4: Normal Quantile-Quantile (QQ) plots for the GFA model (left), and the tFA model (right). The Sample Quantiles are those of d_i's in the fourth-root scale.

6.2 Comparison of PX-ECME and ECME

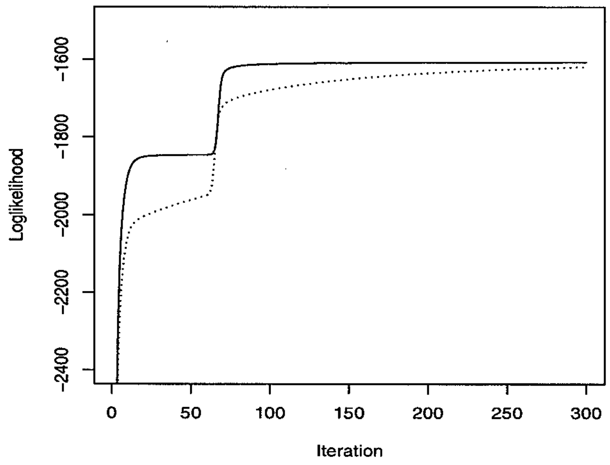

Figure 6.5: Convergence of PX-ECME (solid line) and ECME (dotted line).

As we mentioned in Section 4.3, PX-ECME can accelerate ECME without changing its original advantages. We implemented of PX-ECME with the same initial values and convergence criterion as ECME. PX-ECME converged in 1700 iterations while ECME converged in 2300 iterations, improved 35 percent of the efficiency. Figure 6.5 shows us that, PX-ECME (solid line) dominates the ECME (dotted line) in terms of likelihood values.

6.3 Comparing the Results Obtained from Data Augmentation

In Section 4.1, we discussed the identifiability problem in the FA models. Here we note that Bayesian methods with the incorporation of proper prior information can also eliminate the problem of indeterminacies. For example, in equation (5.5), the unimodality and the symmetry of the prior make the posterior distribution of β unimode. To compare the results we obtained by ECME, we use the same way to obtain the identifiable pattern by converting the factor loading into the lower block triangle matrix in each iteration.

We chose the hyperparameters $\beta_0 = \mathbf{0}$, $n_1 = 1$, $m = p$ and $A = 0$. We also placed the exponential distribution with $\lambda = 1$ as the prior for the degree of freedom v. We run DA in the case where both weights and the degree of freedom are unknown for 30000 iterations. The sample means from the last 10000 iterations are given in Table 6.5-6.7, which are consistent with the results of MLEs. The estimation of the degree of freedom of the model t_5 is 2.3760 with

N_5	0.1742	0.1889	0.2298	0.3326	0.8319
S.d.	3.738e-2	3.756e-2	3.819e-2	3.881e-2	3.946e-2
t_5	0.1010	0.1152	0.1385	0.2134	0.5647
S.d.	1.984e-2	1.937e-2	1.974e-2	1.940e-2	2.554e-2

Table 6.5: Estimation of mean and their standard deviation.

N_5		S.d.		t_5		S.d.	
0.9317	0	2.370e-2	0	0.5369	0	1.983e-2	0
0.9294	0.0172	2.373e-2	1.068e-2	0.5242	0.0156	1.916e-2	0.562e-2
0.8689	0.1739	2.580e-2	8.559e-2	0.4951	0.1209	1.963e-2	1.492e-2
0.8050	0.4007	2.809e-2	1.786e-1	0.4176	0.2746	1.961e-2	2.693e-2
0.5963	0.4831	3.329e-2	2.222e-1	0.2692	0.3511	2.260e-2	3.743e-2

Table 6.6: Estimation of the factor loading matrix and their standard deviation.

standard deviation 0.1767. The corresponding approximate standard errors of all the components of μ, β and ψ in the model t_5 are strictly smaller than the estimated standard errors in the model N_5.

The estimated posterior density functions of the parameters yielded from the model N_5 (dotted lines) are plotted against the estimates obtained from the model t_5 (solid lines) in Figures 6.6-6.9. Apparently, the posterior distributions under the model t_5 have smaller tail probabilities than those under the model N_5. This implies that assuming the data follow the t-distribution may lead to more accurate and efficient estimation.

7. Conclusion

We proposed in this article a robust method for factor analysis. The method is obtained from the usual Gaussian factor analysis model by replacing the Gaussian distribution with the multivariate t-distribution. Both EM-type and DA

N_5	0.02281	0.02673	0.11519	0.09452	0.35437
S.d.	2.767e-3	2.820e-3	3.355e-2	1.519e-1	2.375e-1
t_5	0.00762	0.00673	0.03077	0.01569	0.13418
S.d.	2.448e-3	2.081e-3	1.325e-2	4.472e-2	8.439e-2

Table 6.7: Estimation of the covariance matrix of the error terms and their standard deviation.

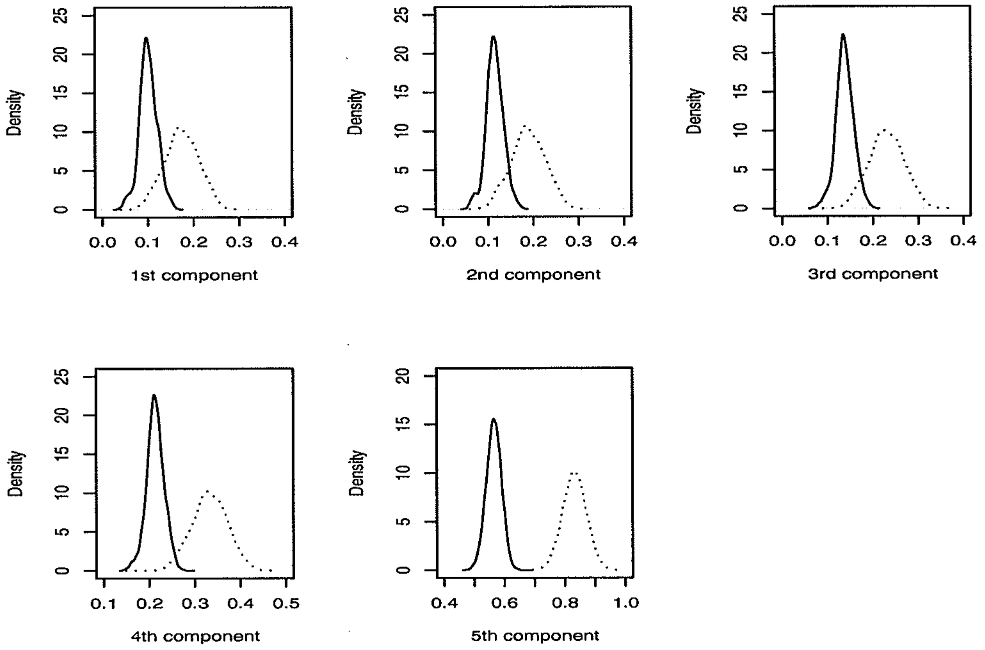

Figure 6.6: Estimated posterior density function of mean vector.

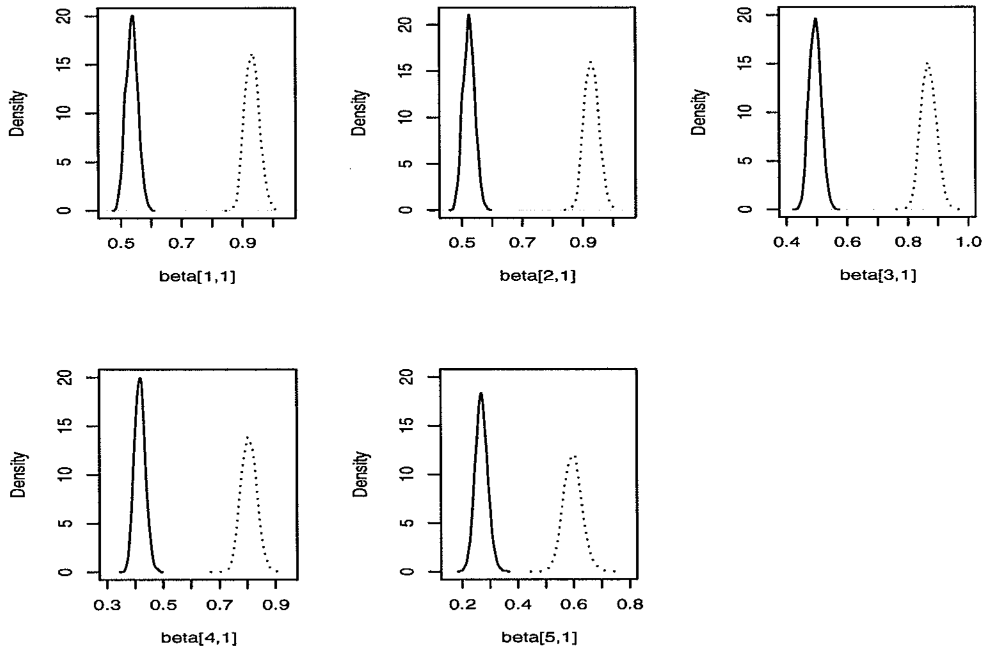

Figure 6.7: Estimated posterior density function of first column of factor loading matrix.

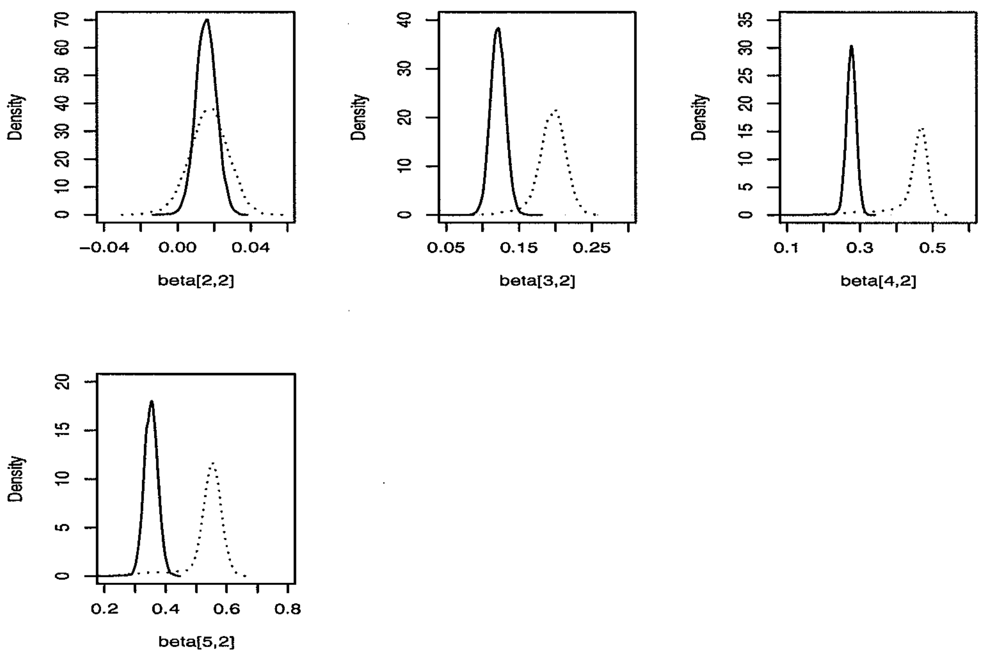

Figure 6.8: Estimated posterior density function of second column of factor loading matrix.

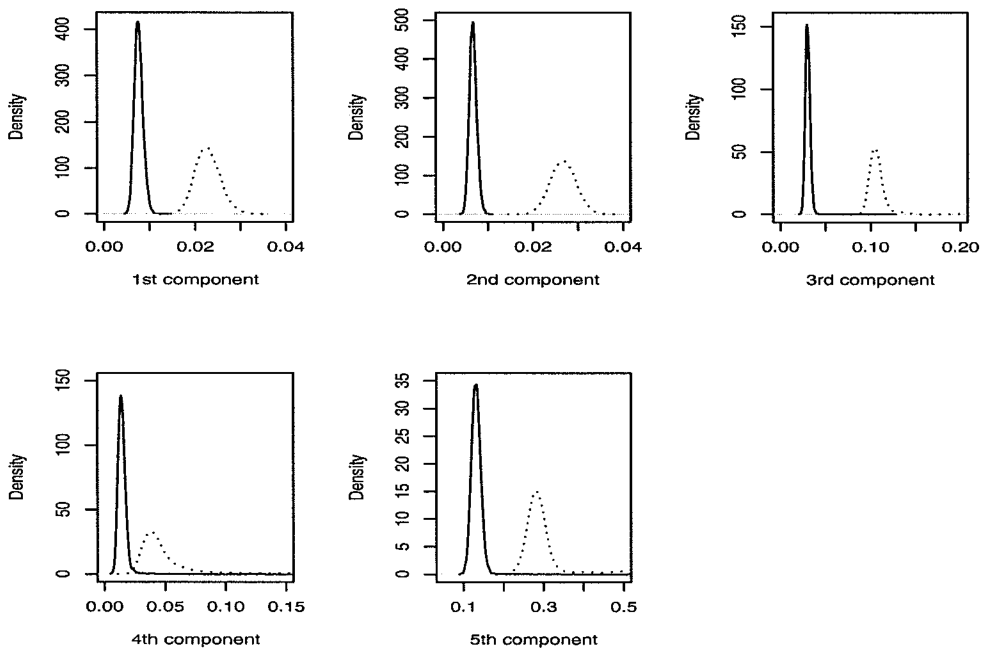

Figure 6.9: Estimated posterior density function of standard deviations of the errors.

algorithms are developed to deal with the situation when the weights τ and degree of freedom are unknown. The closed form expressions for both the E- and M- steps in ECME, and as well as the proportional posterior distribution in Bayesian method are obtained. This robust approach is demonstrated by its application to analyzing the US bond return data. The identifiability problem of the loading matrix is conquered by converting it into the block lower triangle matrix with 0 on the up-right corner, or using the QR decomposition, which makes the comparison meaningful. Smaller standard deviations of the parameters under the t-distribution assumption show both the robustness and efficiency of our approach. We also considered implementation of PX-ECME algorithm for the factor analysis model with the multivariate t-distribution, which converges even faster than ECME.

8. Appendix

8.1 Calculation of CM-Step 1

The log-likelihood of L_1 only has two terms involving α.

$$\begin{aligned} & \mathrm{tr}(\alpha S_{\tau XY}\Psi^{-1} - \frac{1}{2}\alpha'\Psi^{-1}\alpha S_{\tau XX}) \\ = & \mathrm{tr}(A\alpha) - \frac{1}{2}\mathrm{tr}(S_{\tau XX}\alpha'\Psi^{-1}\alpha) \\ = & (vec(\alpha'))'vec(A) - \frac{1}{2}(vec(\alpha'))'vec(S_{\tau XX}\alpha'\Psi^{-1}) \\ = & (vec(\alpha'))'vec(A) - \frac{1}{2}(vec(\alpha'))'[\Psi^{-1} \otimes S_{\tau XX}]vec(\alpha'), \end{aligned}$$

where $A = S_{\tau XY}\Psi^{-1}$. Taking derivative with respect to $vec(\alpha')$, we get

$$vec(\alpha') = [(\Psi^{-1})' \otimes S_{\tau XX}]^{-1}vec(A) = [\Psi \otimes (S_{\tau XX})^{-1}]vec(A), \tag{8.1}$$

and

$$vec(\alpha) = (B \cdot vec(A))' \text{ where } B = \Psi \otimes (S_{\tau XX})^{-1}. \tag{8.2}$$

8.2 Proof of Theorem 1

The conditional posterior distribution of β can be split into two parts. We rewrite $Pr(Y|Z,\tau,\mu,\beta,\Psi)$ into the form:

$$Pr(Y|Z,\tau,\mu,\beta,\Psi) \propto |\Psi|^{-\frac{n}{2}}\exp\{-\frac{1}{2}\mathrm{tr}\Psi^{-1}\sum_{i=1}^{n}\tau_i(y_i-\mu-\beta z_i)(y_i-\mu-\beta z_i)'\}.$$

Let

$$R = \sum_{i=1}^{n} \tau_i (y_i - \mu - \beta z_i)(y_i - \mu - \beta z_i)', \tag{8.3}$$

and let

$$S = \sum_{i=1}^{n} \tau_i (y_i - \mu - \hat{\beta} z_i)(y_i - \mu - \hat{\beta} z_i)'. \tag{8.4}$$

Then

$$R = S + \sum_{i=1}^{n} \tau_i (\beta z_i - \hat{\beta} z_i)(\beta z_i - \hat{\beta} z_i)'.$$

$$\begin{aligned} & \mathrm{tr}\Psi^{-1} \sum_{i=1}^{n} \tau_i (\beta z_i - \hat{\beta} z_i)(\beta z_i - \hat{\beta} z_i)' \\ = \; & \mathrm{tr}\Psi^{-1} (\beta - \hat{\beta}) S_{\tau ZZ} (\beta - \hat{\beta})' \\ = \; & vec(\beta - \hat{\beta})'(S'_{\tau ZZ} \otimes \Psi^{-1}) vec(\beta - \hat{\beta}). \end{aligned}$$

Thus

$$\begin{aligned} Pr(\beta|Y, Z, \tau) \; & \propto \; Pr(Y, Z, \tau|\theta) Pr(\theta) \\ & \propto \; Pr(Y|Z, \tau, \mu, \beta, \Psi) Pr(\beta|\Psi) \\ & \propto \; \exp\{-\frac{1}{2} vec(\beta - \hat{\beta})'(S'_{\tau ZZ} \otimes \Psi^{-1}) vec(\beta - \hat{\beta})\} \\ & \cdot \; \exp\{-\frac{1}{2}(vec(\beta) - \overline{\beta}_0)'(n_1 I_{q \times q} \otimes \Psi^{-1})(vec(\beta) - \overline{\beta}_0)\}. \end{aligned}$$

Let $D_1 = S'_{\tau ZZ} \otimes \Psi^{-1}$ and $D_2 = n_1 I_{q \times q} \otimes \Psi^{-1}$. Then the mean of $vec(\beta)$ is $(D_1 + D_2)^{-1}(D_1 vec(\hat{\beta}) + D_2 \overline{\beta}_0)$ and the covariance is $(D_1 + D_2)^{-1}$.

References

Anderson, D. R. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society* **B 36**, 99-102.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. *Proc. 3rd Berkley Symp. Math. Statist. Prob.* **5**, 111-150.

Anderson, T. W. (2003). An introduction to multivariate statistical analysis. Third edition. Wiley, New York.

Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. *Journal of the Royal Statistical Society* **B 29**, 1-52.

Box, G. E. P. and Tiao, G. C. (1973). Bayesian inference in statistical analysis. Addison-Wesley, Reading. MA.

Geary, R. C. (1930). The frequency distribution of the quotient of two normal variates. *Journal of the Royal Statistical Society* **93**, 442-446.

Gelfand, A. E. and Smith, F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics* **8**, 519-540.

Hayashi, K. and Yuan, K. (2003). Robust Bayesian factor analysis. *Structural Equation Modeling* **10(4)**, 525-533.

Hawkins, D. M., and Wixley, R. A. J. (1986). A note on the transformation of chi-squared variables to normality. *The American Statistician* **40**, 296-298.

He, Y. and Liu, C. (2007). Computation of Fisher information : new methods and a comparative study. *Technical Report*, Department of Statistics, Purdue University.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32(4)**, 443-482.

Krzanowski, W. and Marriott, F. (1995). *Multivariate Analysis.* First edition. Kendalls Library of Statistics 2. Arnold.

Lange, K. L., Little, R. J. and Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881-896.

Lee, S. E., Press, S. J. (1998). Robustness of Bayesian factor analysis estimates. *Communications in Statistics - Theory And Methods* **27(8)**.

Little, R. J. A. (1990). Editing and imputation of multivariate data: issues and new approaches. *Data Quality Control: Theory and Pragmatics* (Edited by Liepens, G. and Uppuluru, V. R. R.), CRC press.

Liu, C. (1993). Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *Journal of Multivariate Analysis* **46**, 198-206.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis* **53**, 139-158.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association* **91**, 1219-1227.

Liu, C. (2002). Robit regression: a simple robust alternative to logistic and probit regression. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* Wiley, New York.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633-648.

Liu, C. and Rubin, D. B. (1995). ML estimation of the multivariate t distribution with unknown degrees of freedom. *Statistica Sinica* **5**, 19-39.

Liu, C. and Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica* **8**, 729-747.

Liu, C., Rubin, D. B. and Wu, Y. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **85**, 755-770.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41-67.

Meng, X. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267-278.

Pinheiro, J. C., Liu, C. and Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* **10**, 249-276.

Pison, G., Rousseeuw, P. J. and Croux, C. (2003). Robust factor analysis *Journal of Multivariate Analysis* **84**, 145-172.

Polasek, W. (2000). Factor analysis and outliers: a Bayesian approach. *Technical Report*, Institute of Statistics and Econometrics, University of Basel, Switzerland.

Press, S. J. and Shigemasu, K. (1997). Bayesian inference in factor analysis-revised. *Technical Report* **243**, Department of Statistics, University of California, Riverside.

Relles, D. A. and Roger, W. H. (1977). Statistics are fairly robust estimators of location. *Journal of the American Statistical Association* **72**, 107-111.

Rousseeuw, P. J. (1983). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* (Edited by W. Grossmann, G. Pflug, I. Vincze, W. Wertz), 283-297.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47**, 69-76.

Ruey, S. T. (2005). *Analysis of Financial Time Series.* Second edition, Wiley, New York.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528-540.

Yuan, K., Chan, W. and Bentler, P. M. (2000). Robust transformation with applications to structural equation modeling. *British Journal of Mathematical and Statistical Psychology* **53**, 31-50.

Yuan, K., Marshall, L. L. and Bentler, P. M. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika* **67**, 95-122.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate student-t error terms. *Journal of the American Statistical Association* **71**, 400-405.

Jia Li
Department of Mathematics, Purdue University
West Lafayette, IN 47907 USA
E-mail:(jiali@math.purdue.edu)
Chuanhai Liu
Department of Statistics, Purdue University
West Lafayette, IN 47907 USA
E-mail:(chuanhai@stat.purdue.edu)
Jianchun Zhang
Department of Statistics, Purdue University
West Lafayette, IN 47907 USA
E-mail: (zhang10@stat.purdue.edu)