Robust Factor Analysis Using the Multivariate t-Distribution

Jianchun Zhang, Jia Li, and Chuanhai Liu

Purdue University

Abstract: Factor analysis is a standard method for multivariate analysis. The sampling model in the most popular factor analysis is Gaussian and has thus often been criticized for its lack of robustness. A simple robust extension of the Gaussian factor analysis model is obtained by replacing the multivariate Gaussian distribution with a multivariate t-distribution. We develop computational methods for both maximum likelihood estimation and Bayesian estimation of the factor analysis model. The proposed methods include the ECME and PX-EM algorithms for maximum likelihood estimation and Gibbs sampling methods for Bayesian inference. Numerical examples show that use of multivariate t-distribution improves the robustness for the parameter estimation in factor analysis.

Key words and phrases: Bayesian Methods; EM-type Algorithms; Gibbs Sampling; Multivariate t-distribution; Robust Factor Analysis.

1. Introduction

Factor analysis (FA) as a popular statistical method to analyze the underlying relations among multivariate random variables has been extensively used in such areas as psychology, psychometrics, and educational testing. It has proven to be a useful tool in big data analysis. Examples of its most recent applications include Bossé et al. (2007), Banerjee and Gupta (2012), and Dickinson et al. (2011). It should be noted that the method of principal component analysis, which plays the role of exploratory data analysis for formal factor analysis, is used routinely in big data analysis (see, e.g., Witten et al. (2010)). The sampling model in the standard factor analysis is Gaussian and has often been criticized for its lack of robustness. It is of particular importance to develop simple robust alternatives for very high dimensional statistical problems.

Technically, the starting point is a linear model in which the observed variables are expressed as linear functions of a vector of unobservable factors and 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 written as, see Johnson and Wichern (2001) or Anderson (2003),

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

where y_i is the p-dimensional i^{th} observation, μ is a p-dimensional column vector, β is the $p \times q$ $(q < p)$ factor loading matrix, z_i is a q-dimensional vector of unobserved factor scores, and $z_i \sim \mathbf{N}_q(0, I_q)$, where I_q denotes the $q \times q$ identity matrix. The error term $\varepsilon_i \sim \mathbf{N}_p(0, \Psi)$, where $\Psi = \text{Diag}(\psi_1^2, ..., \psi_p^2)$ is a diagonal matrix whose components are called uniquenesses. 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 (ML) or Bayesian estimation is not robust to outliers in the data. The classical technique can be thought of as computing the sample covariance matrix or the sample correlation matrix and making inference based on the matrix obtained. This approach is not robust to outliers since they have a large effect on the estimate of the covariance matrix. There are two main streams of robust estimation methods for FA models: get robust estimates of the covariance matrix (see Hayashi and Yuan (2003) and Pison and Rousseeuw (2003)); replace the normal distribution by longer-tailed distributions to accommodate outliers (see Lee and Press (1998) and Polasek (2000)).

Lange, Little, and Taylor (1989) proposed replacing the normal distribution in linear regression models by the multivariate t-distribution for robust estimation. The use of t distribution for robust estimation dates back to Andrews and Mallows (1974) and Zellner (1976) and has been applied in various fields. Liu (1996) developed Bayesian robust multivariate linear regression with incomplete data, and Liu (2004) studied robust logistic regression. Pinheiro, Liu, and Wu (2001) worked on robust estimation in mixed-effects models. A Bayesian treatment of their model can be found in Lin and Lee (2007). The multivariate t-distribution in factor analysis has not been developed, although Yuan *et al.* (2002) mentioned its possible use.

We propose a multivariate-t factor analysis (TFA) model that replaces the normal assumption with the t-distribution. We show that the robustness is improved in TFA. We study ML estimation via the ECME (Liu and Rubin (1994)) and PX-EM (Liu et al. (1998)) algorithms. We also consider Bayesian estimation using the Gibbs sampling method (see Gelfand and Smith (1990)).

The remainder of this paper is arranged as follows. The TFA model is described in Section 2. Section 3 describes the EM-type algorithms for ML estimation of the TFA model. Section 4 considers the Bayesian estimation of the TFA model. Section 5 presents some numerical examples, including a simulation to examine the robustness of the TFA model and an application of the TFA model to a US bond indexes data. Different algorithms are compared in terms of computational efficiency. Conclusions and a few remarks are given in Section 6.

2. A Multivariate t Factor Analysis Model

The GFA model (1.1) can be written as:

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \overset{iid}{\sim} \mathbf{N}_{p+q}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} \beta\beta' + \Psi & \beta \\ \beta' & I_q \end{bmatrix} \right), \quad i = 1, \cdots, n, \tag{2.1}$$

where $(y_i', z_i')'$ is the i^{th} sample with z_i unobservable. For robust estimation of θ, we replace the multivariate normal distribution in (2.1) with the multivariate t-distribution:

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \overset{iid}{\sim} \mathbf{t}_{p+q}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} \beta\beta' + \Psi & \beta \\ \beta' & I_q \end{bmatrix}, v \right), \quad i = 1, \cdots, n, \tag{2.2}$$

where v is the degrees of freedom (df). This model can also be expressed in a hierarchical structure:

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \mid \tau_i \overset{ind}{\sim} \mathbf{N}_{p+q}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \frac{1}{\tau_i} \begin{bmatrix} \beta\beta' + \Psi & \beta \\ \beta' & I_q \end{bmatrix} \right), \tag{2.3}$$

$$\tau_i \overset{iid}{\sim} \text{Gamma}\left(\frac{v}{2}, \frac{v}{2} \right) \tag{2.4}$$

for $i = 1, ..., n$, where τ_i's are the weights. In (2.4), Gamma(a, b) is the gamma distribution with density

$$f(\tau) = b^a \tau^{(a-1)} \text{exp}(-b\tau)/\Gamma(a), \;\; \tau > 0, a > 0, b > 0, \tag{2.5}$$

where $\Gamma(a) = \int_0^\infty t^{a-1}\exp(-t)dt$ denotes the gamma function. The TFA model can then be written as

$$y_i = \mu + \beta z_i + \varepsilon_i, \quad i = 1, ..., n, \tag{2.6}$$

$$z_i \overset{ind}{\sim} t_q(0, I_q, v), \quad \varepsilon_i \overset{ind}{\sim} t_p(0, \Psi, v), \tag{2.7}$$

where z_i and ε_i are uncorrelated, but dependent. Thus, the GFA and TFA models have the structure (2.6), but have different distributions for the factor loading matrix and error term. When v goes to infinity, the TFA model reduces to the GFA model.

3. Efficient EM-type Algorithms For ML Estimation

Dempster, Laird, and Rubin (1977) and Rubin and Thayer (1982) use the EM algorithm for ML estimation of the GFA model, while Liu and Rubin (1998) use the ECME algorithm for ML estimation of the GFA model. Note that the missing data in the EM algorithm for ML estimation of GFA consist of the latent factors z_i's, while the EM algorithm for ML estimation of TFA involves both the missing weights τ_i's and the latent factors z_i's. As a result, the EM algorithm for TFA can be slow, especially when the number of degrees of freedom ν is to be estimated. In this section, we consider the ML estimation of the TFA model using the ECME and PX-EM algorithms.

The ECME algorithm 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 often substantially better than either EM or ECM, yet it retains the stable monotone convergence of EM, and can be only modestly more difficult to implement.

The ECME algorithm shares the simplicity of the EM and the efficiency of the Newton-Raphson. Like EM, ECME uses simple updates of parameters that can be very high dimensional and, thereby, can make it difficult to apply Newton-Raphson. Although stable, EM can be painfully slow. When Newton-Raphson is used to update the parameter over a low-dimensional subspace that dominates the rate of convergence of EM, ECME can have a dramatically improved rate of convergence.

To accelerate the EM algorithm, we consider the PX-EM algorithm; it shares the simplicity and stability of ordinary EM, but has a faster rate of convergence. Technically, the PX-EM algorithm is simply the EM algorithm applied to a parameter expanded model with the M-step followed by a reduction step that maps the estimate of the expanded parameter to the original parameter space.

3.1 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 $z^* = Q'z$ for all $q \times q$ orthogonal matrix Q. There are several ways to impose constraints on β to deal with the indeterminacy. One way is to add the restriction 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. Another way is to constrain β to be a block lower triangular matrix of full rank, with diagonal elements strictly positive (see, *e.g.*, Lopes and West (2004)). We use the latter when using the information matrix to estimate the standard errors of ML estimates. When using ECME and PX-EM, an unrestricted β is assumed, since without fully identifiable parameters EM-type algorithms converge to likelihood-equivalent points subject to an orthogonal transformation.

3.2 MLE With Unknown Degrees of freedom

Let $Y = [y_1, y_2, ..., y_n]'$ be the $n \times p$ data matrix and $Z = [z_1, z_2, \cdots, z_n]'$ be the $n \times q$ factor score matrix. If Z and $\tau = \{\tau_1, \tau_2, \cdots, \tau_n\}$ are observed, the log-likelihood function for the complete data with unknown degrees of freedom v is

$$\mathbf{L}(\mu, \beta, \Psi, v|Y, Z, \tau) = L_1(\mu, \beta, \Psi|Y, Z, \tau) + L_2(v \mid \tau) + \text{constant}, \qquad (3.1)$$

where

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

$$L_2(v \mid \tau) = \frac{vn}{2}\log\frac{v}{2} + \frac{v}{2}\sum_{i=1}^{n}\log\tau_i - \frac{v}{2}\sum_{i=1}^{n}\tau_i - n\log\Gamma(\frac{v}{2}). \tag{3.2}$$

The sufficient statistics for $L_1(\mu, \beta, \Psi|Y, \tau)$ 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 $S_{\tau ZZ} = \sum_{i=1}^n \tau_i z_i z_i'$. The conditional distribution of y_i given weight τ_i is

$$y_i \mid \tau_i \sim \mathbf{N}_p\left(\mu, \frac{1}{\tau_i}(\beta\beta' + \Psi)\right), \tag{3.3}$$

Applying Bayes theorem, the conditional distribution of τ_i given y_i is

$$\tau_i \mid y_i \sim \text{Gamma}\left(\frac{v+p}{2}, \frac{v + d(y_i, \mu, \beta\beta' + \Psi)}{2}\right), \tag{3.4}$$

where $d(y_i, \mu, \beta\beta' + \Psi) = (y_i - \mu)'(\beta\beta' + \Psi)^{-1}(y_i - \mu)$ denotes the Mahalanobis distance between y_i and its expectation μ. We then have that $E(\tau_i|y_i) = \frac{v+p}{v+d(y_i,\mu,\beta\beta'+\Psi)}$ and $E(\log\tau_i|y_i) = \psi\left(\frac{v+p}{2}\right) - \log\left(\frac{v+d(y_i,\mu,\beta\beta'+\Psi)}{2}\right)$, where $\psi(s)$ is the digamma function $\frac{\partial\Gamma(s)/\partial s}{\Gamma(s)}$.

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

$$\delta = (\frac{1}{\tau_i}\beta')[\frac{1}{\tau_i}(\Psi + \beta\beta')]^{-1} = \beta'(\Psi + \beta\beta')^{-1}, \tag{3.5}$$

$$\Delta_i = \frac{1}{\tau_i}I_q - \frac{1}{\tau_i}\beta'(\Psi + \beta\beta')^{-1}\beta = \frac{1}{\tau_i}\Delta. \tag{3.6}$$

For implementation, we note that the E-step of EM and ECME algorithms are the same. As well, the EM and ECM are the same if we partition the parameter as $\theta = (\theta_1, \theta_2)$, where $\theta_1 = (\mu, \beta, \Psi)$ and $\theta_2 = v$, since θ_1 and θ_2 are optimized independently of each other in the M-step.

EM and ECME algorithms:

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 conditional expectation of the sufficient statistics:

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

M step: Rewrite the FA model by combining the mean vector and the factor loading matrix so that

$$
y_i = \mu + \beta z_i + \varepsilon_i \Longrightarrow y_i = \begin{pmatrix} \mu & \beta \end{pmatrix} \begin{pmatrix} 1 \\ z_i \end{pmatrix} + \varepsilon_i \Longrightarrow y_i = \alpha x_i + \varepsilon_i, \qquad (3.7)
$$

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

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

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

CM step 1: By maximizing the conditional expectation of $L_1(\mu, \beta, \Psi|Y, Z, \tau)$,

$$vec(\alpha^{(t+1)}) = (\Psi^{(t)} \otimes \hat{T}^{(t)}_{\tau XX}) \cdot vec(A), \tag{3.8}$$

where $A = \hat{S}^{(t)}_{\tau XY} {\Psi^{(t)}}^{-1}$ and $\hat{T}^{(t)}_{\tau XX}$ is the inverse of $\hat{S}^{(t)}_{\tau XX}$, which is the conditional expectation of $S_{\tau XX}$ given $(Y, \theta^{(t)})$; $vec(X)$ denotes the vector formed by stacking the column vectors of the matrix X, and $\otimes$ stands for the Kronecker product operator. The uniquenesses are updated as

$$\begin{aligned} \Psi^{(t+1)} &= \frac{1}{n}\text{Diag}\left(E\Big[\sum_{i=1}^{n} \tau_i^{(t+1)} (y_i - \alpha^{(t+1)} x_i)(y_i - \alpha^{(t+1)} x_i)' | Y, \theta^{(t)} \Big] \right) \\ &= \frac{1}{n}\text{Diag}(\hat{S}^{(t)}_{\tau YY} - 2\alpha^{(t+1)} \hat{S}^{(t)}_{\tau XY} + \alpha^{(t+1)} \hat{S}^{(t)}_{\tau XX} \alpha^{(t+1)'}). \end{aligned}$$

CM step 2: Update $v^{(t+1)}$ by maximizing the conditional expectation of $L_2(v|\tau)$ over v to obtain

$$v^{(t+1)} = \arg\max_{v} \left\{ \frac{v}{2} \Big[\log \frac{v}{2} + \sum_{i=1}^{n} (\log \tau_i^{(t)} - \tau_i^{(t)})/n + \psi(\frac{v^{(t)} + p}{2}) - \log(\frac{v^{(t)} + p}{2}) \Big] - \log \Gamma(\frac{v}{2}) \right\}. \tag{3.9}$$

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. Alternatively, in this CM step, we can apply ECME by maximizing the actual log-likelihood over v with (α, Ψ) being fixed at their most recent estimates. Since $y_i \sim t_p(\mu, \beta\beta' + \Psi, v)$ independently for $i = 1, \cdots, n$, we have the following.

CML step 2: Update $v^{(t+1)}$ as

$$v^{(t+1)} = \arg\max_{v} \left\{ \sum_{i=1}^{n} \Big[\log \Gamma(\frac{v+p}{2}) - \log \Gamma(\frac{v}{2}) + \frac{v}{2} \log v - \frac{v+p}{2} \log(v + d_i) \Big] \right\}, \tag{3.10}$$

where $d_i = d(y_i, \mu^{(t+1)}, \beta^{(t+1)} \beta^{(t+1)'} + \Psi^{(t+1)})$. This step requires only a one-dimensional optimization.

To implement the PX-EM algorithm, we expand the covariance matrix of z_i to a diagonal matrix R. Then the TFA model becomes

$$y_i | z_i, \tau_i \sim N_p(\mu + \beta_* z_i, \frac{1}{\tau_i} \Psi), \tag{3.11}$$

$$z_i|\tau_i \sim N_p(0, \frac{1}{\tau_i}R), \tau_i \sim Gamma(\frac{v}{2}, \frac{v}{2}), i = 1, \ldots, n. \tag{3.12}$$

Thus, the many-to-one mapping ζ reduces to $\beta = \zeta(\beta_*, R)$, with a little abuse of notation.

PX-EM algorithm:

PX-E step: This is the E-step of EM algorithm except for a few changes. Specifically, the covariance matrix in Mahalanobis distance is $\beta_* R \beta_*' + \Psi$; δ in (3.5) is $R\beta_*'(\Psi + \beta_* R\beta_*')^{-1}$; and Δ in (3.6) is $R - R\beta_*'(\Psi + \beta_* R\beta_*')^{-1}\beta_* R$.

PX-M step: The computations for $\mu^{(t+1)}$ and $\Psi^{(t+1)}$ stay the same as those in EM. For $\beta^{(t+1)}$, obtain $\beta_*^{(t+1)}$ in the same way as that in EM, then update R as $R^{(t+1)} = diag(\hat{S}_{\tau ZZ}^{(t+1)})/n$ and set $\beta^{(t+1)}$ as $\beta_*^{(t+1)} {R^{(t+1)}}^{1/2}$.

A numerical comparison of the computational efficiencies of these algorithms is given in Section 5.3.

4. Bayesian Approach

We study the Bayesian estimation of the TFA model via the Gibbs sampling method. We first specify prior distributions for the parameters, and calculate the conditional posterior distributions. Then Gibbs sampling is applied to obtain the posterior distributions of the parameters.

4.1 Prior Distributions

We assume that the prior distribution for the parameters $\theta = (\mu, \beta, \Psi, v)$ has the form

$$Pr(\theta) = Pr(\mu, \beta, \Psi, v) = Pr(\mu)Pr(\beta|\Psi)Pr(\Psi)Pr(v). \tag{4.1}$$

This form of prior is usually adopted for highly structured models such as factor analysis model (see Rowe (2003)). Independent priors are used, partially at least for simplicity. The particular choice of the prior for β dependent on Ψ has also been used in the context of linear regression with censored data; see Hamada and Wu (1995) and Liu and Sun (2000). For specification of $Pr(\mu)$, $Pr(\beta|\Psi)$, and $Pr(\Psi)Pr(v)$, diffuse but conjugate priors are often preferred. Here a flat prior is used for the location parameter μ. For the factor loading matrix

$\beta = [\beta_1, \beta_2, \cdots, \beta_q]_{p\times q}$, we take

$$\beta_i|\Psi \overset{ind}{\sim} \mathbf{N}_p\left(\beta_0, \frac{\Psi}{\kappa}\right), i = 1, \cdots, q, \tag{4.2}$$

where the hyper-parameter κ, is usually taken to be as small as possible.

The uniquenesses are assumed to be iid inverse gamma IG$(a/2, b/2)$ *a priori*, with a, b taken to be small. The adoption of inverse gamma prior instead of the standard noninformative prior can help avoid the Bayesian analogue of the so-called Heywood cases (see Martin and McDonald (1981)).

For the degrees of freedom v, a brief discussion on how to choose a prior is given in Liu (1995). A recent critical discussion can be found in Fonseca, Goldberg, and Migon (2008). In order to obtain a proper posterior of v, the basic rule is that the prior should satisfy $Pr(v) = o(v^{-1})$ as $v \to +\infty$. We adopt the flat prior distribution for v^{-1}, $Pr(v) \propto v^{-2}I(v \geq 1)$.

4.2 Full Conditionals

The full conditional distributions are derived as follows. Given (2.3), the conditional posteriors for factor scores are independently normal,

$$z_i|y_i, \tau_i, \theta \sim \mathbf{N}_q(\beta'(\beta\beta' + \Psi)^{-1}(y_i - \mu), [I_q - \beta'(\beta\beta' + \Psi)^{-1}\beta]/\tau_i). \tag{4.3}$$

Similar to (3.4), the conditional posteriors for the weights τ_is are independently gamma,

$$\tau_i|y_i, z_i, \theta \sim \text{Gamma}(\frac{v+p+q}{2}, \frac{v + d([y_i', z_i']', [\mu', 0]', \Psi_{yz})}{2}), \tag{4.4}$$

where $\Psi_{yz} = \begin{bmatrix} \beta\beta' + \Psi & \beta \\ \beta' & I_q \end{bmatrix}$.

It is easy to derive that the conditional posterior for μ is multivariate normal distribution,

$$\mu|Y, Z, \tau, \beta, \Psi, v \sim N_p(\overline{\mu}, V) \tag{4.5}$$

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

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

With the joint prior distribution of the uniquenesses

$$Pr(\Psi) \propto |\Psi|^{-\frac{a+2}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Psi^{-1}A)\}, \tag{4.6}$$

where $A = bI_q$, the posterior distribution of Ψ conditional on the observations is

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

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 $d = a+n+q$. Then

$$\psi_i^{-2} \overset{ind}{\sim} \frac{\chi_d^2}{h_i^2} \quad (i = 1, \cdots, p), \tag{4.7}$$

which are independent inverse gamma distributions.

For the degrees of freedom v, the fact that $\tau_i|v \sim \mathrm{Gamma}(\frac{v}{2}, \frac{v}{2})$ leads to the conditional posterior

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

The implementation of Gibbs sampling is straightforward. All the unknown quantities except v can be drawn directly according to their conditional distributions. For drawing the degree of freedom v, a Metropolis sampler with truncated normal proposal performs well, where the standard deviation of the normal deviate is 0.1. The $(t + 1)$-th iteration of Gibbs sampler is as follows

Step 1: Draw $z_i^{(t+1)}$ independently from $f(z_i|y_i, \tau_i^{(t)}, \theta^{(t)})$ for $i = 1, \ldots, n$;
Step 2: Draw $\tau_i^{(t+1)}$ independently from $f(\tau_i|y_i, z_i^{(t+1)}, \theta^{(t)})$, for $i = 1, \ldots, n$;
Step 3: Draw $\mu^{(t+1)}$ from $f(\mu|Y, Z^{(t+1)}, \tau^{(t+1)}, \beta^{(t)}, \Psi^{(t)}, v^{(t)})$;
Step 4: Draw $\beta^{(t+1)}$ from $f(\beta|Y, Z^{(t+1)}, \tau^{(t+1)}, \mu^{(t+1)}, \Psi^{(t)}, v^{(t)})$;
Step 5: Draw $\psi_j^{2(t+1)}$ independently from $f(\psi_j^2|Y, Z^{(t+1)}, \tau^{(t+1)}, \mu^{(t+1)}, \beta^{(t+1)}, v^{(t)})$;
Step 6: Draw $v^{(t+1)}$ from $f(v|Y, Z^{(t+1)}, \tau^{(t+1)}, \mu^{(t+1)}, \beta^{(t+1)}, \Psi^{(t+1)})$,
where $f(\cdot|\cdot)$ denotes the corresponding conditional posterior distribution.

4.3 Partially Collapsed Gibbs Sampler

The ordinary Gibbs sampler is straightforward given that the corresponding conditional posterior distribution is available at each step. However, it can have slow convergence given the complex structured model. As a Bayesian counterpart of the ECME algorithm, we consider a version of the so-called partially collapsed Gibbs (PCG) sampler proposed recently by van Dyk and Park (2008). The idea of PCG is to replace some steps of the ordinary Gibbs sampler by drawing the components from the conditional distributions under some marginal distributions instead of that under the joint posterior distribution. Such changes often generate samples of a set of incompatible conditional distributions, but van Dyk and Park (2008) design a recipe through marginalization, permutation, and trimming of the modified sampler to achieve correct and faster convergence.

As an analogue to the ECME algorithm, we replace the draw of the degree-of-freedom v in Step 6 of the ordinary Gibbs sampler with a draw (together with Z and τ) from the posterior distribution conditioning only on Y and the other parameters

Modified-Step 6: Draw (Z, τ, v) from $f(Z, \tau, v|Y, \mu, \beta, \Psi)$.

We also replace Step 2 of the ordinary Gibbs sampler by not conditioning on Z

Modified-Step 2: Draw (Z, τ) from $f(Z, \tau|Y, \theta, v)$.

Note that this creates the draw of Z and τ in the first step of the ordinary Gibbs sampler because only the most recent values are conditioned upon for the next iteration. By moving Step 1 and the modified-Step 2 to the end and trimming the intermediate quantities Z and τ produced in the modified Step 2 and modified Step 6, and combining (blocking) the last three steps together, we obtain the following.

Step 1: Draw $\mu^{(t+1)}$ from $f(\mu|Y, Z^{(t)}, \tau^{(t)}, \beta^{(t)}, \Psi^{(t)}, v^{(t)})$;

Step 2: Draw $\beta^{(t+1)}$ from $f(\beta|Y, Z^{(t)}, \tau^{(t)}, \mu^{(t+1)}, \Psi^{(t)}, v^{(t)})$;

Step 3: Draw $\psi_j^{2(t+1)}$ independently from $f(\psi_j^2|Y, Z^{(t)}, \tau^{(t)}, \mu^{(t+1)}, \beta^{(t+1)}, v^{(t)})$;

Step 4: Draw $v^{(t+1)}$ from $f(v|Y, \mu^{(t+1)}, \beta^{(t+1)}, \Psi^{(t+1)})$,
Draw $\tau_i^{(t+1)}$ independently from $f(\tau_i|y_i, \theta^{(t+1)})$, for $i = 1, \ldots, n$;
Draw $z_i^{(t+1)}$ independently from $f(z_i|y_i, \tau_i^{(t+1)}, \theta^{(t+1)})$ for $i = 1, \ldots, n$.

Note that Step 4 of PCG is nothing but the joint distribution of (Z, τ, v) conditioning on the other quantities, leading to a blocked sampler. Similar to the ordinary Gibbs sampler, all the quantities except v can be drawn easily, while the degrees of freedom v is updated via a Metropolis sampler.

5. Numerical Examples

In this section, we use two numerical examples to show the improvement of robustness in TFA comparing with GFA. Computational efficiencies for MLE and Gibbs sampling are also discussed.

5.1 A Simulation Study

We generated the data Y from

$$(1-\pi)N(0,\beta\beta'+\Psi)+\pi N(0,w(\beta\beta'+\Psi))$$

for different combinations of π and f. We set the parameters as $\Psi = 0.1 * I_5$ and

$$\beta = \begin{bmatrix} 2 & 0 & 2 & 0 & 0 \\ 0 & 3 & 0 & 4 & 5 \end{bmatrix}'.$$

All eight combinations of $\pi = 0.05, 0.1, 0.15, 0.2$ and $w = 2, 5$ were used in the simulation study. The $w = 2$ case corresponds to a slight contamination pattern, while $w = 5$ illustrates a more distant contamination pattern. A total of 100 Monte Carlo replications were obtained for each combination. The sample size in each replication was $n = 300$. The ML estimates via ECME were obtained for GFA and TFA models.

We compare the estimated factor loadings with their respective true values by evaluating

$$r = \sum_{i=1}^{100}(\hat{\theta_{Ti}}-\theta_0)^2 / \sum_{i=1}^{100}(\hat{\theta_{Gi}}-\theta_0)^2, \tag{5.1}$$

where θ_0 is the true value of the parameter of interest, and $\hat{\theta_T}$ and $\hat{\theta_G}$ are the ML estimates under TFA and GFA models, respectively. A similar method of comparison can be found in Pinheiro *et al.* (2001). Figure 5.1 presents the simulation results for some selected factor loadings. There are substantial gains in terms of accuracy under the distant contamination pattern ($w = 5$) with respect to the slight contamination pattern ($w = 2$). As the chance of outliers increases, the TFA is favorable in the sense that the estimated factor loadings are closer to the true value under both contamination patterns. This demonstrates the robustness of the TFA model.

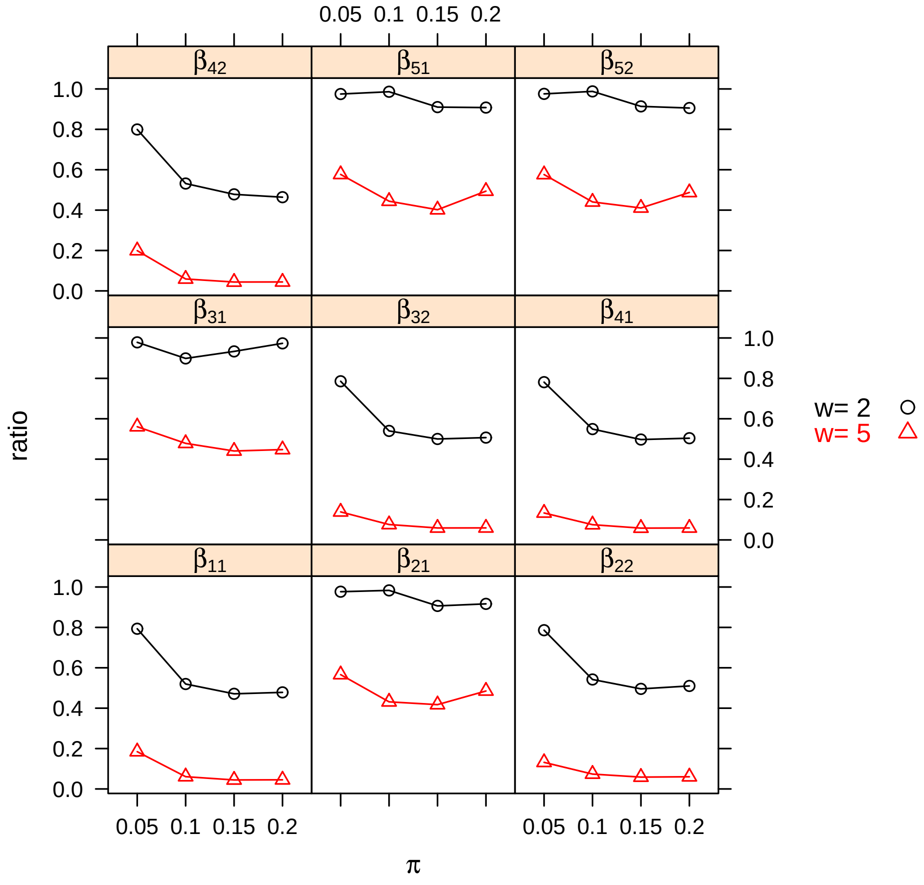
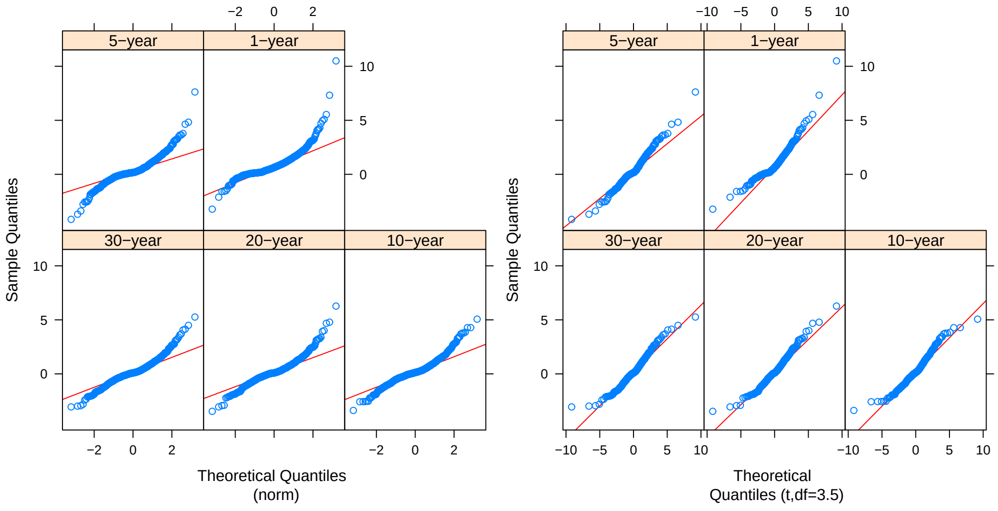


Figure 5.1: The ratio of empirical mean square error under TFA model with respect to GFA model for selected factor loadings.

5.2 US Bond Indexes Data Set

We considered 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 and argued that the original data could be used because the correlation matrix changed little after fitting a multivariate ARMA model. To be comparable with Tsay's results, we adjusted the data by dividing each component by its sample standard deviation. Figure 5.2 (a) shows the Q-Q normal plots of the five US bond indexes in terms of log-return. Heavy tails are clearly present in all five variables, and the p-value of the Shapiro-Wilk test is close to zero for each index. As a result, the normal distribution is not appropriate for this data set. Instead, we used the t-distribution to capture the pattern of heavy tails. Figure 5.2 (b) shows the Q-Q Student-t plots using the estimated degrees of freedom, evidently supporting the use of the t-distribution. In Section 4.2.2, we compare the ML estimates and Bayesian estimates based on the GFA and TFA models.



(a)

(b)

Figure 5.2: (a) Q-Q normal plots for log-return of each US-bond index, (b) Q-Q Student-t plots with degrees of freedom 3.5 for log-return of each US-bond index.

5.2.1 Application to Analyzing US Bond Indexes Data

We applied our method for robust factor analysis to the US bond indexes data set using Model N_5 where the observations are assumed to follow the Gaussian distribution and Model t_5 where the observations are assumed to follow the t-distribution. We found that simple exploratory data analysis supports the use of

Table 5.1: Test results of model N_5 and t_5 with different number of factors, where NA means the degrees-of-freedom is calculated as negative.

Model	# of factors	1	2	3	4
N_5	max log-likelihood	-2509.16	-2213.65	-2209.66	2209.64
	Likelihood Ratio	599.03(df=5)	8.02 (df=1)	0.04(df=NA)	0.04(df=NA)
	AIC	5048.32	4465.32	4463.33	4467.33
	BIC	5116.50	4551.68	4563.32	4576.42
t_5	max log-likelihood	-1842.05	-1605.97	-1601.62	-1601.62
	Likelihood Ratio	240.42(df=5)	8.78 (df=1)	0.02(df=NA)	0.02(df=NA)
	AIC	3716.10	3251.94	3249.24	3253.24
	BIC	3788.83	3342.85	3353.78	3362.87

distribution having heavier tails than the normal distribution, and that the use of Model t_5 results in a substantially improved fit to the observed data.

The likelihood ratio test can be used to help select the number of factors. The null hypothesis that the current factor analysis model has the covariance matrix structure $\Sigma = \beta\beta' + \Psi$ is tested against the alternative in which the covariance matrix structure is unconstrained. Under some regularity conditions, the likelihood ratio test statistic has chi-squared distribution asymptotically with degree-of-freedom $\max\{[(p-q)^2-(p+q)]/2, 0\}$. To assure the degrees-of-freedom is a positive integer, there is an upper bound for the number of factors (see Lopes and West (2004)). Usually, one starts with a small number of factors, say $q = 1$, testing goodness-of-fit until a nonsignificant result occurs, or the degree of freedom becomes non-positive. We refer to Jöreskog (1967) and Anderson (2003) for more details about this procedure. This procedure was criticized 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. With the given output, AIC preferred 3 factors under both normal and t assumptions, while BIC preferred 2 factors in both cases. Considering model parsimony, we chose to focus on 2 factor models. Tasy (2005) also fit the 2 factor GFA model.

5.2.2 Comparing the Gaussian and the Multivariate t MLEs

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

The ML estimates of two FA models are shown in Tables 5.2, 5.3, and 5.4. The estimate of the degrees of freedom of the model t_5 is 2.275 with standard deviation 0.1661. The associated variance-covariance matrix was computed via numerical differentiation. The mean vector shifts to left in the model t_5 because the data present a slight skewness to the left that cannot be accounted for by using symmetric distributions such as normal and t distributions. Table 5.3 shows a dramatic difference between ML estimates of the factor loading matrix under the two models. Although the estimated factor loading matrices are significantly different, the components in the estimated matrices have a 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.

Lange *et al.* (1989) considered diagnostics to check model assumptions. For 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 an asymptotic chi-squared distribution with degrees 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 a fourth-root transformation. Here, we use the fourth-root transformation (Hawkins and Wixley (1986)) because it performs well when the degrees-of-freedom p is small. For the TFA model, d_i^2/p has an asymptotic F-distribution with degrees of freedom 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 5.3 shows the normal quantile-quantile

Table 5.2: Estimation of mean and their standard deviation.

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 5.3: Estimation of the factor loading matrix 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 5.4: Estimation of the covariance matrix of the error terms and their standard deviation.

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

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.

5.2.3 Comparing the Results with those Obtained from Gibbs Sampling

Bayesian methods with the incorporation of proper prior information can eliminate the problem of indeterminacy. For example, the unimodality and the symmetry of the prior make the posterior distribution of β unimodal. To compare the results with those obtained via ECME, we obtain an identifiable pattern by converting the factor loading into a lower block triangle matrix in each iteration.

We followed Gelman and Rubin (1992) in monitoring the convergence of sampling. Starting from three initial points, we ran three independent chains simultaneously and stopped (at $\tilde{3}0000$ iterations) when all the parameters had

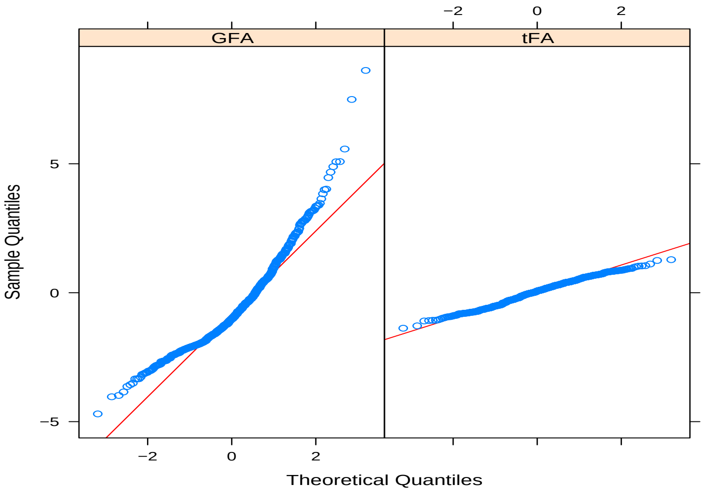


Figure 5.3: Normal Quantile-Quantile (QQ) plots for the GFA model (left), and the TFA model (right).

R close to 1. To be conservative, we used the second half of the samples for inference. The estimated values are listed in Table 5.5, 5.6 and 5.7, consistent with the ML results. The estimation of the degrees of freedom of Model t_5 is 2.3931 with standard deviation 0.1954.

5.3 Comparison of computational efficiency

5.3.1 Comparison of EM, ECME and PX-EM

We report some numerical results on computational efficiency of the three algorithms. Performance depends on such factors as data structure, initial value, and missing information. Meng and van Dyk (1997) reported that, in a multivariate-t model, the performance of ECME was better than that of the multi-cycle ECM in one dataset in term of the number of iterations, but similar in the other, given that the initial values were the same. However, the CPU time of ECME was at least 2 times of that of the multi-cycle ECM.

For our bond index dataset, we considered different initial values. Numerical experiments showed that the PX-EM generally performed best in terms of number of iterations and CPU time, while ECME was slightly better than EM

Table 5.5: Estimation of mean vector and its standard deviation(in parenthesis).

$\hat{\mu}$ model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
N_5	0.1637 (4.079e-2)	0.1781 (4.077e-2)	0.2189 (4.056e-2)	0.3212 (4.038e-2)	0.8225 (3.963e-2)
t_5	0.1115 (2.573e-2)	0.1253 (2.512e-2)	0.14820 (2.517e-2)	0.2231 (2.335e-2)	0.5752 (2.784e-2)

Table 5.6: Estimation of the factor loading matrix and its standard deviation(in parenthesis).

$\hat{\beta}$ model	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{31}$	$\hat{\beta}_{41}$	$\hat{\beta}_{51}$
N_5	1.0012 (3.013e-2)	0.9886 (3.148e-2)	0.9273 (3.240e-2)	0.8610 (3.356e-2)	0.6411 (3.618e-2)
t_5	0.5953 (2.463e-2)	0.5833 (2.443e-2)	0.5524 (2.443e-2)	0.4711 (2.394e-2)	0.3110 (2.665-2)
$\hat{\beta}$ model	-	$\hat{\beta}_{22}$	$\hat{\beta}_{32}$	$\hat{\beta}_{42}$	$\hat{\beta}_{52}$
N_5	0 -	0.0359 (1.488e-2)	0.2091 (2.566e-2)	0.5080 (4.757e-2)	0.5311 (5.437e-2)
t_5	0 -	0.0133 (0.576e-2)	0.1184 (2.888e-2)	0.2823 (5.920e-2)	0.3365 (7.637e-2)

Table 5.7: Estimation of the vector of uniquenesses and its standard deviation(in parenthesis).

$\hat{\Psi}$ model	$\hat{\psi}_1^2$	$\hat{\psi}_2^2$	$\hat{\psi}_3^2$	$\hat{\psi}_4^2$	$\hat{\psi}_5^2$
N_5	0.0098 (1.307e-3)	0.0335 (1.470e-3)	0.1080 (1.353e-3)	0.0123 (4.292e-2)	0.3163 (5.495e-2)
t_5	0.0072 (1.786e-3)	0.0051 (1.125e-3)	0.0328 (1.127e-3)	0.0149 (4.120e-2)	0.1533 (6.834e-2)

Table 5.8: Comparison of effective sample size (ESS) and effective sample size per second (ESS/s) between ordinary Gibbs sampler and PCG sampler

	Gibbs		PCG			Gibbs		PCG	
Par.	ESS	ESS/s	ESS	ESS/s	Par.	ESS	ESS/s	ESS	ESS/s
μ_1	37	0.06	69	0.13	β_{11}	56	0.09	38	0.07
μ_2	37	0.06	63	0.12	β_{21}	50	0.08	35	0.07
μ_3	25	0.04	56	0.11	β_{31}	34	0.06	27	0.05
μ_4	24	0.04	67	0.12	β_{41}	37	0.06	56	0.11
μ_5	85	0.14	148	0.28	β_{51}	116	0.20	139	0.26
ψ_1^2	343	0.58	244	0.46	β_{22}	914	1.55	576	1.09
ψ_2^2	263	0.45	188	0.35	β_{32}	2810	4.76	2111	4.0
ψ_3^2	583	0.99	385	0.73	β_{42}	2781	4.71	2543	4.82
ψ_4^2	13	0.02	24	0.05	β_{52}	3312	5.61	2671	5.06
ψ_5^2	503	0.85	567	1.07	v	621	1.05	276	0.52

in terms of number of iterations but slower than EM in running time. For example, when the initial values were set to the results from GFA using R function `factanal`, the mean was taken to be the sample mean μ, and the degrees-of-freedom v was set to be 100, the number of iterations were 2119, 2113, 1924 for EM, ECME and PX-EM, respectively, given the the absolute tolerance of log-likelihood to be 1e-4. The CPU times were 18.34s, 24.66s, and 17.11s, respectively, on a laptop with Intel Core Duo Processor T2350 and R version 2.10.

5.3.2 Comparison of the ordinary Gibbs sampler and the PCG sampler

We ran the two samplers 30000 iterations and used the second half of the samples to compare the effective sample size (ESS) and effective sample size per second. Table 5.8 shows the ESS and ESS per second for all the parameters. The two samplers performed similarly in terms of computational efficiency. In terms of running time, the PCG sampler was slightly faster.

6. Conclusion

A most recent EM-type algorithm, the dynamic 'expectation-conditional maximization either' (DECME) algorithm (He and Liu, (2012)), appears to efficient compared to existing EM-type algorithms. It would be interesting to investigate its performance for TFA and to develop its Markov chain Monte Carlo versions for Bayesian inference with the TFA models.

As in the case of the dataset, the TFA model cannot properly account for the skewness of the sample. The skew-elliptical distribution (Genton (2004)) could be considered. Even in the case of symmetry, the TFA model could be generalized by allowing different numbers of degrees of freedom for each variable.

Appendix: The Proof of Theorem 1

Let $vec(\beta) = [\beta_1', \beta_2', \cdots, \beta_q']'$. Then

$$vec(\beta)|\Psi \sim \mathbf{N}_{p\times q}\left(\overline{\beta}_0, I_{p\times q} \otimes \frac{\Psi}{n_1}\right), \tag{6.1}$$

where $\overline{\beta}_0 = [\beta_0', \beta_0', \cdots, \beta_0']'$. The posterior distribution of β conditional on the observations is of the form

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

We rewrite $Pr(Y|Z, \tau, \mu, \beta, \Psi)$ as

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

. With

$$R = \sum_{i=1}^{n} \tau_i (y_i - \mu - \beta z_i)(y_i - \mu - \beta z_i)', \tag{6.2}$$

$$S = \sum_{i=1}^{n} \tau_i (y_i - \mu - \hat{\beta} z_i)(y_i - \mu - \hat{\beta} z_i)', \tag{6.3}$$

then

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

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

$$
\begin{aligned}
Pr(\beta|Y,Z,\tau) &\propto Pr(Y,Z,\tau|\theta)Pr(\theta)\\
&\propto Pr(Y|Z,\tau,\mu,\beta,\Psi)Pr(\beta|\Psi)\\
&\propto exp\{-\frac{1}{2}vec(\beta-\hat{\beta})'(S'_{\tau ZZ}\otimes\Psi^{-1})vec(\beta-\hat{\beta})\}\\
&\cdot\ exp\{-\frac{1}{2}(vec(\beta)-\overline{\beta}_0)'(n_1I_{q\times q}\otimes\Psi^{-1})(vec(\beta)-\overline{\beta}_0)\}
\end{aligned}
$$

Let $D_1 = S'_{\tau ZZ}\otimes\Psi^{-1}$ and $D_2 = n_1I_{q\times q}\otimes\Psi^{-1}$. Then the mean of $vec(\beta)$ is $(D_1+D_2)^{-1}(D_1vec(\hat{\beta})+D_2\overline{\beta}_0)$ and the covariance is $(D_1+D_2)^{-1}$.

Acknowledgment

The authors are grateful to the Editors, an associate editor and the referees for helpful comments and suggestions, which greatly improved the presentation of this article.

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.

Banerjee, U.S. and Gupta, S. (2012). Assessment of significant sources influencing the variation of water quality of the river damodar through factor analysis. *Asian Journal of Water, Environment and Pollution* **9**, 87-94.

Bossé, Y., Després, J-P., Chagnon, Y. C., Rice, T., Rao, D. C., Bouchard, C., Pérusse, L., and Marie-Claude, M-C. (2007). Quantitative trait locus on 15q for a metabolic syndrome variable derived from factor analysis. *Obestity* **15**, 544-550.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society* B **39**, 1-38.

Dickinson, D. ,Goldberg, T. E, Gold, J. M., Elvevåg, B., and and Daniel R. Weinberger, D. R. (2011). Cognitive factor structure and invariance in people with schizophrenia, their unaffected siblings, and controls. *Schizophr. Bull.* **37**, 1157-1167.

Fonseca T.C.O., Ferreira, M.A.R., and Migon, H.S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika* **95**, 325-333.

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.

Genton, M.G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality.* Edited Volume, Chapman and Hall, Florida.

Hamada, M. and Wu, C. F. J. (1995). Analysis of censored data from fractionated experiment: a Bayesian approach. *J. Amer. Statist. Assoc.* **90**, 467–477.

Hayashi, K. and Yuan, K. (2003). Robust Bayesian factor analysis. *Structural Equation Modeling* **10**, 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. (2012). The dynamic 'expectation-conditional maximization either' algorithm. *Journal of the Royal Statistical Society* B, **74**, 313-336.

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

Johnson, R. A. and Wichern, D, W. (2001). *Applied Multivariate Statistical Analysis*, 5th edition. Prentice Hall.

Krzanowski, W. and Marriott, F. (1995). *Multivariate Analysis.* Kendalls Library of Statistics 2. Wiley, New York.

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. and Press, S. J. (1998). Robustness of Bayesian factor analysis estimates. *Communications in Statistics - Theory And Methods* **27**, 1871-1893.

Lin, T.I. and Lee, J.C. (2007) Bayesian analysis of hierarchical linear mixed modeling using the multivariate t distribution. *Journal of Statistical Planning and Inference* **137**, 484-495.

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. (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. (2004). Robit regression: a simple robust alternative to logistic and probit regression. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 227-238.

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* **8**, 729-747.

Liu, C. and Sun, D. X. (2000). Analysis of interval-censored data from fractional experiments using covariance adjustment. *Technometrics* **42**, 353-365.

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

Martin, J.L. and McDonald, R.P. (1981). Bayesian estimation in unrestricted factor analysis: a treatment for Heywood cases. *Psychometrika* **40**, 505-517.

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.

Rowe, D. B. (2003). *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing.* Chapman and Hall/CRC.

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

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

van Dyk, D and Park, T. (2008). Partially collapsed Gibbs samplers: theory and methods. *Journal of the American Statistical Association* **103**, 790-797.

Witten, D., Tibshirani, R., Gu, S. G., Fire, A., and Weng-Onn Lui, W-O. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology* **8**, doi:10.1186/1741-7007-8-58.

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.

Jianchun Zhang
Department of Statistics, Purdue University
West Lafayette, IN 47907 USA
E-mail: (zhangjc98@gmail.com)

Jia Li
Department of Mathematics, Purdue University
West Lafayette, IN 47907 USA
E-mail:(jialimath@gmail.com)

Chuanhai Liu
Department of Statistics, Purdue University
West Lafayette, IN 47907 USA
E-mail:(chuanhai@purdue.edu)