Efficient Factor Analysis Using the Multivariate t-Distribution via PX-EM

ZHOU Rui

1 Preliminary Knowledge

1.1 Student-t Distribution

The multivariate student-t distribution with notation $\mathbf{t}_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right)$ has the density

$$f\left(\mathbf{x}\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{p}{2}}\pi^{\frac{p}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}\left[1+\frac{1}{\nu}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{-\frac{\nu+p}{2}}$$

where ν is the degrees of freedom, $\boldsymbol{\Sigma}$ is a $p\times p$ matrix, $\boldsymbol{\mu}$ is a p-dimensional constant vector, $\Gamma\left(a\right) = \int_0^\infty t^{(a-1)}\exp\left(-t\right)dt$ is the gamma function. Note here the covariance matrix should be $\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$.

1.2 Gamma Distribution

The standard $\text{Gamma}\left(a, b\right)$ distribution has density function

$$f\left(\tau\right) = b^a\tau^{(a-1)}\frac{\exp\left(-b\tau\right)}{\Gamma\left(a\right)}$$

According to the properties of Gamma distribution, the expectation of τ is $\mathrm{E}\left(\tau\right) = a/b$ and logarithmic expectation is $\mathrm{E}\left(\log\tau\right) = \phi\left(a\right) - \log b$, where $\phi\left(x\right) = d\log\left(\Gamma\left(x\right)\right)/dx$ is the digamma function.

1.3 Hierarchical Structure of Student-t

A classical student-t distribution can be represented in hierarchical structure as

$$\begin{aligned}\mathbf{x} &\overset{i.i.d}{\sim} \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{\tau}\boldsymbol{\Sigma}\right)\\ \tau &\overset{i.i.d}{\sim} \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)\end{aligned}$$

1.4 PX-EM

The classical EM algorithm is usually criticised for its slow convergence rate. Liu [1] proposed a method called parameter expansion to accelerate EM, i.e., PX-EM. As presented in [1], it can be used in multivariate student-t situation as follows:

$$\begin{aligned}\mathbf{x} &\overset{i.i.d}{\sim} \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{\tau}\boldsymbol{\Sigma}\right)\\ \tau &\overset{i.i.d}{\sim} \alpha\cdot\text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)\end{aligned} \tag{1}$$

where α is a expanded scale parameter to be estimated. Note that when the $\alpha = 1$, this model can be reduced to classical hierarchical structure of student-t distribution.

2 Introduction

In this report, we consider the observed p-dimension data $\mathbf{y}_i, i = 1, \dots, n$ follows a student-t distribution, i.e., $\mathbf{y}_i \overset{i.i.d}{\sim} \mathbf{t}_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right)$. Note here we assume $\mathbf{y}_i$ to be mean-centered. According to factor analysis, the high-dimensional observed data can also be represented by the sum of low-dimensional factors influence and some residuals as $\mathbf{y}_i = \boldsymbol{\mu}+\mathbf{L}\mathbf{f}_i+\boldsymbol{\varepsilon}_i$. The $\mathbf{L}$ is a $p\times r$ matrix, $\mathbf{f}_i$ is a r-dimensional vector and $\boldsymbol{\varepsilon}_i$ is a p-dimensional vector. Besides, $\mathbf{f}_i$ and $\boldsymbol{\varepsilon}_i$ are usually assumed to be uncorrelated and $\boldsymbol{\varepsilon}_i$ be inter-elements uncorrelated. This structure implies that the parameter $\boldsymbol{\Sigma}$ should have the certain structure as

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi},$$

where $\mathbf{B} \in \mathbb{R}^{p\times r}$ and $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_p) \geq \epsilon\mathbf{I}$. To simplify the discussion in the following Section, we denote $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \mathbf{B}, \boldsymbol{\Psi}, \nu\}$. And we can also express the distribution of $\mathbf{y}$ as

$$\mathbf{y}_i \overset{i.i.d}{\sim} \mathbf{t}_p\left(\boldsymbol{\mu}, \mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi}, \nu\right).$$

3 Estimation of $\boldsymbol{\theta}$ Using PX-ML

Considering the hierarchical structure of student-t distribution, the $\boldsymbol{\theta}$ can be estimated through PX-EM. Assuming the $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_n\}$ is observed, we can express the negative log likelihood of complete data $(\mathbf{Y}, \boldsymbol{\tau})$, ignoring the constants, is

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\theta}_*|\mathbf{Y}, \boldsymbol{\tau}\right) &= -\sum_{i=1}^{n} \log f\left(\mathbf{y}_i|\tau_i, \boldsymbol{\theta}_*\right) - \sum_{i=1}^{n} \log f\left(\tau_i|\boldsymbol{\theta}_*\right)\\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\tau_i\left(\mathbf{y}_i-\boldsymbol{\mu}\right)\left(\mathbf{y}_i-\boldsymbol{\mu}\right)^T\right)\\ &\quad -\frac{n\nu}{2}\log\frac{\nu}{2} - \frac{\nu}{2}\sum_{i=1}^{n}\left(\log\frac{\tau_i}{\alpha} - \frac{\tau_i}{\alpha}\right) + n\log\Gamma\left(\frac{\nu}{2}\right)\end{aligned}$$

where $\boldsymbol{\theta}_* = \{\boldsymbol{\mu}, \mathbf{B}, \boldsymbol{\Psi}, \nu, \alpha\}$ is the expanded parameters set.

3.1 E-step

The E-step of the EM algorithm is to find the conditional expectation of $\mathcal{L}\left(\boldsymbol{\theta}_*|\mathbf{Y}, \boldsymbol{\tau}\right)$ under the observed $\mathbf{Y}$ and the current estimation of $\boldsymbol{\theta}_*$, i.e., $\boldsymbol{\theta}_*^{(t)}$. Applying the Bayes theorem, the conditional distribution of τ_i given $\mathbf{y}_i$ is

$$\mathbf{y}_i \sim \alpha^{(t)}\mathrm{Gamma}\left(\frac{\nu^{(t)}+p}{2}, \frac{\nu^{(t)} + \left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)^T\left(\boldsymbol{\Sigma}^{(t)}/\alpha^{(t)}\right)^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)}{2}\right)$$

whose proof is similar to last report. Therefore, we can have the follows

$$\mathrm{E}\left(\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right) = \alpha^{(t)}\frac{\nu^{(t)}+p}{\nu^{(t)} + \left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)^T\left(\boldsymbol{\Sigma}^{(t)}/\alpha^{(t)}\right)^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)}$$

$$\mathrm{E}\left(\log\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right) = \log\alpha^{(t)} + \phi\left(\frac{\nu^{(t)}+p}{2}\right) - \log\frac{\nu^{(t)} + \left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)^T\left(\boldsymbol{\Sigma}^{(t)}/\alpha^{(t)}\right)^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)}{2}$$

Denote the approximation function under the current parameters estimation as $Q\left(\boldsymbol{\theta}_*|\boldsymbol{\theta}_*^{(t)}\right)$, which can be written as:

$$\begin{aligned}Q\left(\boldsymbol{\theta}_*|\boldsymbol{\theta}_*^{(t)}\right) &= \mathrm{E}_{\boldsymbol{\tau}}\left(\mathcal{L}\left(\boldsymbol{\theta}_*|\mathbf{Y}, \boldsymbol{\tau}\right)|\mathbf{Y}, \boldsymbol{\theta}_*^{(t)}\right)\\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| + \frac{n}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\mathbf{S}}^{(t)}_{\tau YY}\right)\\ &\quad -\frac{n\nu}{2}\log\frac{\nu}{2} - \frac{n\nu}{2}\hat{S}_{\tau*\tau}\left(\alpha\right) + n\log\Gamma\left(\frac{\nu}{2}\right)\end{aligned}$$

where $\hat{\mathbf{S}}^{(t)}_{\tau YY} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right)\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)^T$ and
$\hat{S}^{(t)}_{\tau*\tau}\left(\alpha\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}\left(\log\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right) - \log\alpha - \mathrm{E}\left(\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right)/\alpha\right)$.

3.2 M-step

The M-step of EM algorithm is to find the minimum solution to the approximation function, i.e.,

$$\begin{aligned}
\underset{\boldsymbol{\theta}}{\mathsf{minimize}} \quad & Q\left(\boldsymbol{\theta}_*|\boldsymbol{\theta}_*^{(t)}\right) \\
\mathsf{subject\ to} \quad & \boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi} \\
& \boldsymbol{\Psi} = \operatorname{diag}\left(\psi_1, \ldots, \psi_p\right) \geq \epsilon \mathbf{I}
\end{aligned}$$

Note that the optimal value of variables $\boldsymbol{\mu}$ and α can be directly derived by taking its derivative be zero, i.e.,

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^n \mathrm{E}\left(\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right)\mathbf{y}_i}{\sum_{i=1}^n \mathrm{E}\left(\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right)}$$

$$\alpha^{(t+1)} = \frac{1}{N}\sum_{i=1}^n \mathrm{E}\left(\tau_i|\mathbf{y}_i, \boldsymbol{\theta}_*^{(t)}\right)$$

Besides, because the optimization variables $(\mathbf{B}, \boldsymbol{\Psi})$ and ν are fully decoupled in objective function and constraint. The above problem can be easily rewritten into two problem, the one w.r.t ν is

$$\underset{\nu}{\mathsf{minimize}} - \nu \log \frac{\nu}{2} - \nu \hat{S}_{\tau * \tau}^{(t)}\left(\alpha^{(t+1)}\right) + 2 \log \Gamma\left(\frac{\nu}{2}\right) \tag{2}$$

which is a unconstrainted problem and can be easily solved by taking the derivative to be zero, i.e., by solving

$$-\log \frac{\nu}{2} - 1 - \hat{S}_{\tau * \tau}^{(t)}\left(\alpha^{(t+1)}\right) + \phi\left(\frac{\nu}{2}\right) = 0$$

The above equation is proved to have a unique solution and happen to be the optimal solution to problem (2) when $\hat{S}_{\tau * \tau}\left(\alpha^{(t+1)}\right) + 1 < 0$ is satisfied. Similarly, we can also get the subproblem w.r.t $(\mathbf{B}, \boldsymbol{\Psi})$ as

$$\begin{aligned}
\underset{\mathbf{B},\boldsymbol{\Psi}}{\mathsf{minimize}} \quad & -\log|\boldsymbol{\Sigma}^{-1}| + \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\mathbf{S}}_{\tau YY}^{(t)}\right) \\
\mathsf{subject\ to} \quad & \boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi} \\
& \boldsymbol{\Psi} = \operatorname{diag}\left(\psi_1, \ldots, \psi_p\right) \geq \epsilon \mathbf{I}
\end{aligned} \tag{3}$$

which has been efficiently solved by algorithm proposed in [2], which could be regarded as MM iterations.

Algorithm 1 Efficient algorithm to problem (3)

```
**Require** $\boldsymbol{\Psi}^{(0)}$ and set $\boldsymbol{\Phi}^{(0)} = \left(\boldsymbol{\Psi}^{(0)}\right)^{-1}$, $t = 0$
**repeat**
    1. compute $\triangledown_i = \left(\left(\boldsymbol{\Phi}^{(t)}\right)^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_1\mathbf{U}^T\left(\boldsymbol{\Phi}^{(t)}\right)^{\frac{1}{2}}\mathbf{S}\right)_{ii}$ for $i = 1, \ldots, p$, where
       $\mathbf{U}\operatorname{diag}\left(\lambda_1^*, \ldots, \lambda_p^*\right)\mathbf{U}^T = \left(\boldsymbol{\Phi}^{(t)}\right)^{\frac{1}{2}}\mathbf{S}\left(\boldsymbol{\Phi}^{(t)}\right)^{\frac{1}{2}}$ and $\mathbf{D}_1 = \operatorname{diag}\left(\delta_1, \ldots, \delta_p\right)$ with
       $\delta_i = \begin{cases} \max\left\{0, 1 - \frac{1}{\lambda_i^*}\right\} & 1 \leq i \leq r \\ 0 & \text{otherwise} \end{cases}$
    2. update $\Phi_{ii}^{(t+1)} = \min\left\{\frac{1}{S_{ii} - \triangledown_i}, \frac{1}{\epsilon}\right\}$ for $i = 1, \ldots, p$ and $t \leftarrow t + 1$
**until** convergence
recovery $\boldsymbol{\Psi}^{\star} = \left(\boldsymbol{\Phi}^{(t)}\right)^{-1}$
compute $\mathbf{B}^{\star} = \left(\boldsymbol{\Psi}^{\star}\right)^{\frac{1}{2}}\left[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_r\right]$ where $\mathbf{z}_i$ are the largest $r$ eigenvectors of $\left(\boldsymbol{\Psi}^{\star}\right)^{-\frac{1}{2}}\mathbf{S}\left(\boldsymbol{\Psi}^{\star}\right)^{-\frac{1}{2}}$
rescaled by corresponding largest $r$ eigenvalues $\lambda_i^*, i \leq r$, with $\|\mathbf{z}_i\|^2 = \max\left\{1, \lambda_i^*\right\} - 1$
```

3.3 Complete EM algorithm

The complete PX-EM algorithm is presented as Algorithm 1. It should be mentioned that the EM algorithm is a particular case of MM algorithm. Therefore, to avoid the heavy computation incurred by double loop, we can run only one iteration when solve the problem (3).

Algorithm 2 PX-EM algorithm

```
Require: $\mathbf{B}^{(0)}$,$\mathbf{\Psi}^{(0)}$ and $\nu^{(0)}$, $t = 0$
repeat
  E-step: calculate $\hat{\mathbf{S}}_{\tau YY}^{(t)}$ and $\hat{S}_{\tau*\tau}^{(t)}$
  M-step: update $\boldsymbol{\mu}^{(t+1)}$ and $\alpha^{(t+1)}$
          update $\nu^{(t+1)}$ by solving problem (2)
          update $\mathbf{B}^{(t+1)}$,$\mathbf{\Psi}^{(t+1)}$ by solving problem (3)
  $t \leftarrow t + 1$
until convergence
```

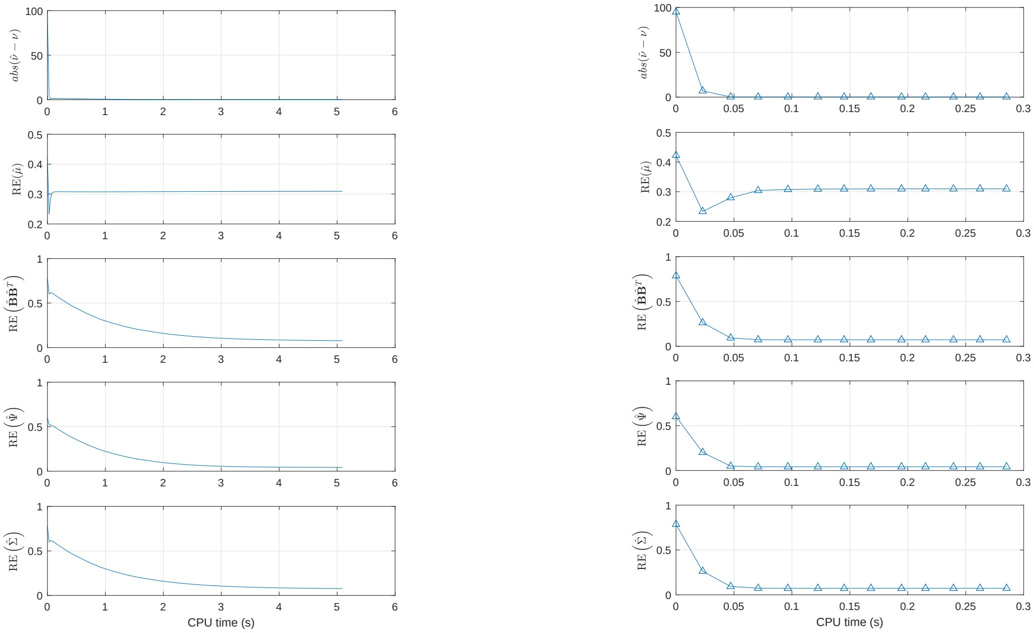
4 Numerical simulation

In this section, we compare the performance of classical EM algorithm and PX-EM algorithm. We generate data following student-t distribution $\mathbf{t}_p\left(\boldsymbol{\mu}, \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}, \nu\right)$ and compare the estimation results $\left\{\hat{\mathbf{B}}, \hat{\mathbf{\Psi}}\right\}$ in relative error as follows:

$$\text{RE}\left(\hat{\mathbf{B}}\hat{\mathbf{B}}^T\right) = \frac{\|\mathbf{B}\mathbf{B}^T - \hat{\mathbf{B}}\hat{\mathbf{B}}^T\|_F}{\|\mathbf{B}\mathbf{B}^T\|_F}, \quad \text{RE}\left(\hat{\mathbf{\Psi}}\right) = \frac{\|\mathbf{\Psi} - \hat{\mathbf{\Psi}}\|_F}{\|\mathbf{\Psi}\|_F}$$

$$\text{RE}\left(\hat{\mathbf{\Sigma}}\right) = \frac{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F}{\|\mathbf{\Sigma}\|_F}, \quad \text{RE}\left(\hat{\boldsymbol{\mu}}\right) = \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2}{\|\boldsymbol{\mu}\|_2}$$

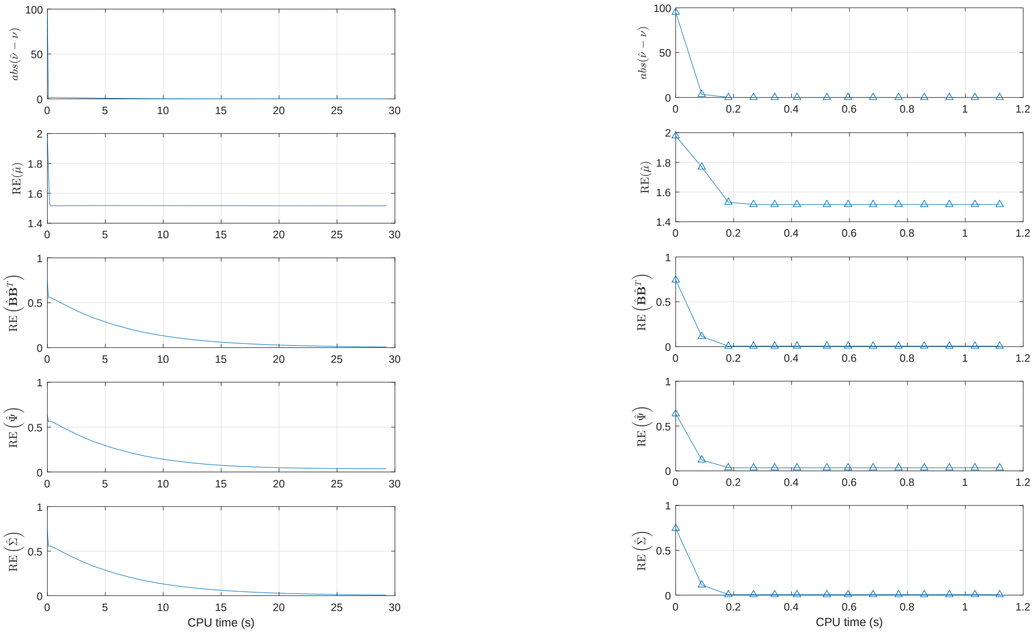
In Figure 1, we set $p = 200$, $n = 5p$ and $\nu = 5$. Here $\mathbf{B}$ and $\mathbf{\Psi}$ are generated randomly with all conditions satisfied. The initial value of estimated parameters are given by $\mathbf{B}^{(0)} = \mathbf{U}_K\mathbf{\Lambda}_K^{\frac{1}{2}}$, $\mathbf{\Psi}^{(0)} = \text{diag}\left(\mathbf{S} - \mathbf{B}^{(0)}\mathbf{B}^{(0)T}\right)$ with $\mathbf{U\Lambda U}$ be the eigen value decomposition of sample covariance matrix $\mathbf{S}$. In another word, the initial point is determined by PCA estimation. It is obvious that the proposed algorithms can both improve a lot compared with simple PCA. Surprisingly, the PX-EM can be one order of magnitude faster than the PX-EM. In Figure 2, we set $p = 400$, $n = 5p$ and $\nu = 5$. The same properties are still held.



(a) EM

(b) PX-EM

Figure 1: Comparison between EM and PX-EM, $p = 200$, $n = 1000$, $\nu = 5$



(a) EM

(b) PX-EM

Figure 2: Comparison between EM and PX-EM, $p = 400$, $n = 2000$, $\nu = 5$

References

[1] C. Liu, D. B. Rubin, and Y. N. Wu, “Parameter expansion to accelerate em: the px-em algorithm,” *Biometrika*, vol. 85, no. 4, pp. 755–770, 1998.

[2] K. Khamaru and R. Mazumder, “Computation of the maximum likelihood estimator in low-rank factor analysis,” *arXiv preprint arXiv:1801.05935*, 2018.