Generalized Hyperbolic Skew t Parameter Estimation via EM and PXEM

ZHOU Rui

March 31, 2022

1 Generalized Hyperbolic Skew t Distribution

Suppose $\mathbf{x} \in \mathbb{R}^N$is a random vector following the generalized hyperbolic (GH) Skew t distribution $\mathsf{ST}\left(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}\right)$, its probability density function (pdf) can be written as:

$$f\left(\mathbf{x}\right) = \frac{2^{1-\frac{\nu+N}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\left(\pi\nu\right)^{\frac{N}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}} \frac{K_{\frac{\nu+N}{2}}\left(\left\{\left(\nu + Q\left(\mathbf{x}\right)\right)\boldsymbol{\gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\right\}^{\frac{1}{2}}\right)\exp\left(\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\right)}{\left\{\left(\nu + Q\left(\mathbf{x}\right)\right)\boldsymbol{\gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\right\}^{-\frac{\nu+N}{4}}\left(1 + Q\left(\mathbf{x}\right)/\nu\right)^{\frac{\nu+N}{2}}},$$

where $\boldsymbol{\mu} \in \mathbb{R}^N$ is the location vector, $\boldsymbol{\gamma} \in \mathbb{R}^N$ is the skewness vector, $\boldsymbol{\Sigma} \in \mathbb{S}_+^N$ is the scatter matrix, $\nu > 0$ is the degrees of freedom. K_α is the modified Bessel function, and $Q\left(\mathbf{x}\right) = \left(\mathbf{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)$. Some reference can be accessed via: https://arxiv.org/pdf/1703.02177.pdf.

Luckly, it can be represented in a hierarchical structure:

$$\begin{aligned} \mathbf{x}|\tau &\sim \mathcal{N}\left(\boldsymbol{\mu} + \frac{1}{\tau}\boldsymbol{\gamma}, \frac{1}{\tau}\boldsymbol{\Sigma}\right), \\ \tau &\sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned} \tag{1}$$

1.1 Comparison with Other Skew t Models

Apart from the Generalized Hyperbolic Skew t, there also exist some other skew t models, e.g.,

- The restricted multivariate skew t distribution (whose PX-EM model has been developed in [1])

$$\begin{aligned} \mathbf{x}|\boldsymbol{u}, w &\sim \mathcal{N}\left(\boldsymbol{\mu} + u\boldsymbol{\gamma}, \frac{1}{\tau}\boldsymbol{\Sigma}\right), \\ u|w &\sim HN\left(\mathbf{0}, \frac{1}{\tau}\right), \\ \tau &\sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \end{aligned}$$

- The unrestricted multivariate skew t distribution

$$\begin{aligned} \mathbf{x}|\boldsymbol{u}, w &\sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{u}\odot\boldsymbol{\gamma}, \frac{1}{\tau}\boldsymbol{\Sigma}\right), \\ \boldsymbol{u}|w &\sim HN\left(\mathbf{0}, \frac{1}{\tau}\boldsymbol{I}_N\right), \\ \tau &\sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \end{aligned}$$

2 Parameter Estimation via Expectation-maximization (EM) Algorithm

Considering we have observed T samples $\{\mathbf{x}_t\}_{t=1}^T$, each of them follows a i.i.d GH Skew t distribution with parameter $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\gamma}, \nu, \boldsymbol{\Sigma}\}$. Directly solving the maximum likelihood estimation (MLE) problem can be very intractable as the probability density function (pdf) of a GH Skew t is very complicated. The EM algorithm can be applied here for solving the MLE problem. Regarding the $\{\tau_t\}_{t=1}^T$ as the latent variables, we can write the complete data log-likelihood function as:

$$
\begin{aligned}
&\ell\left(\{\mathbf{x}_t\}_{t=1}^T, \{\tau_t\}_{t=1}^T | \boldsymbol{\theta}\right)\\
=&\sum_{t=1}^T \log\left(f_{\mathcal{N}}\left(\mathbf{x}_t|\boldsymbol{\mu}+\frac{1}{\tau_t}\boldsymbol{\gamma}, \frac{1}{\tau_t}\boldsymbol{\Sigma}\right)\cdot f_{\mathrm{GM}}\left(\tau_t|\frac{\nu}{2},\frac{\nu}{2}\right)\right)\\
=&\sum_{t=1}^T\left\{-\frac{\tau_t}{2}\left(\mathbf{x}_t-\boldsymbol{\mu}-\frac{1}{\tau_t}\boldsymbol{\gamma}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}-\frac{1}{\tau_t}\boldsymbol{\gamma}\right)-\frac{1}{2}\log\det\left(\boldsymbol{\Sigma}\right)+\frac{\nu}{2}\left(\log\tau_t-\tau_t\right)+\frac{\nu}{2}\log\frac{\nu}{2}-\log\Gamma\left(\frac{\nu}{2}\right)\right\}+\mathit{const.}\\
=&\sum_{t=1}^T\Bigg\{-\frac{1}{2}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\left(\tau_t(\mathbf{x}_t-\boldsymbol{\mu})(\mathbf{x}_t-\boldsymbol{\mu})^T-2(\mathbf{x}_t-\boldsymbol{\mu})\boldsymbol{\gamma}^T+\frac{1}{\tau_t}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right)\right\}-\frac{1}{2}\log\det\left(\boldsymbol{\Sigma}\right)\\
&\qquad+\frac{\nu}{2}\left(\log\tau_t-\tau_t\right)+\frac{\nu}{2}\log\frac{\nu}{2}-\log\Gamma\left(\frac{\nu}{2}\right)\Bigg\}+\mathit{const.}
\end{aligned}
\tag{2}
$$

2.1 Expectation (E) Step

The expectation step is to compute the following minimum sufficient statistics: $\{\mathrm{E}(\tau_t)\}_{t=1}^T$, $\{\mathrm{E}\left(\tau_t^{-1}\right)\}_{t=1}^T$, and $\{\mathrm{E}(\log\tau_t)\}_{t=1}^T$. The distribution of τ_t conditional on current estimates $\boldsymbol{\theta}^{(k)}$ and $\mathbf{x}_t$ is given in the following Lemma:

Lemma 1. *The conditional distribution of τ, given the estimates $\boldsymbol{\theta}$ and sample $\mathbf{x}$, is*

$$
\tau|\boldsymbol{\theta},\mathbf{x}\sim\mathsf{GIG}\left(\lambda,\delta^2,\kappa^2\right),
\tag{3}
$$

where $\mathsf{GIG}(\cdot)$ means the generalized inverse Gaussian distribution with $\lambda=(\nu+N)/2$, $\delta=\sqrt{\boldsymbol{\gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}^T}$, and $\kappa=\sqrt{\nu+(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$.

Thus, the minimum sufficient statistics can be computed as

$$
\begin{aligned}
\mathrm{E}(\tau)&=\frac{\delta}{\kappa}\frac{K_{\lambda+1}(\delta\kappa)}{K_\lambda(\delta\kappa)},\\
\mathrm{E}\left(\tau^{-1}\right)&=\frac{\kappa}{\delta}\frac{K_{\lambda-1}(\delta\kappa)}{K_\lambda(\delta\kappa)},\\
\mathrm{E}(\log\tau)&=\log\frac{\delta}{\kappa}+\frac{\partial}{\partial y}\log K_y(\delta\kappa)\Big|_{y=\lambda},
\end{aligned}
\tag{4}
$$

where $K_y(\cdot)$ is the modified Bessel function of the second kind with a real parameter y.

NOTE: expression might meet numerical issues in practice, i.e., denominator might extremely small so that regarded as zero by computer. We may approximate the corresponding expectation via the Monte Carlo method.

The statistics $\mathrm{E}(\log\tau)$ does not admit a closed form expression and has to be evaluated numerically. Finally, we can obtain the expected log-likelihood function as

$$
\begin{aligned}
&Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)\\
=&\sum_{t=1}^T\Bigg\{-\frac{1}{2}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\left(\mathrm{E}(\tau_t)(\mathbf{x}_t-\boldsymbol{\mu})(\mathbf{x}_t-\boldsymbol{\mu})^T-2(\mathbf{x}_t-\boldsymbol{\mu})\boldsymbol{\gamma}^T+\mathrm{E}\left(\tau_t^{-1}\right)\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right)\right\}-\frac{1}{2}\log\det\left(\boldsymbol{\Sigma}\right)\\
&\qquad+\frac{\nu}{2}\left(\mathrm{E}(\log\tau_t)-\mathrm{E}(\tau_t)\right)+\frac{\nu}{2}\log\frac{\nu}{2}-\log\Gamma\left(\frac{\nu}{2}\right)\Bigg\}+\mathit{const.}
\end{aligned}
\tag{5}
$$

2.2 Maximization (M) Step

The maximization step is implemented via the block maximization, a.k.a. expectation conditional maximization (ECM) algorithm. The degrees of freedom ν can be updated as

$$
\nu^{(k+1)}=\arg\max_\nu\left\{\frac{\nu}{2}\frac{1}{T}\sum_{t=1}^T\left(\mathrm{E}(\log\tau_t)-\mathrm{E}(\tau_t)\right)+\frac{\nu}{2}\log\frac{\nu}{2}-\log\Gamma\left(\frac{\nu}{2}\right)\right\}
\tag{6}
$$

Besides, by setting the corresponding derivative be zero, we have

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}} &= \sum_{t=1}^{T} \tau_t \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_t - \boldsymbol{\mu} - \frac{1}{\tau_t}\boldsymbol{\gamma}\right) = \mathbf{0}, \\
\frac{\partial}{\partial \boldsymbol{\gamma}} &= \sum_{t=1}^{T} \tau_t \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_t - \boldsymbol{\mu} - \frac{1}{\tau_t}\boldsymbol{\gamma}\right)\frac{1}{\tau_t} = \mathbf{0}, \\
\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} &= -\frac{1}{2}\sum_{t=1}^{T}\left\{\mathrm{E}\left(\tau_t\right)(\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})^T - 2(\mathbf{x}_t - \boldsymbol{\mu})\boldsymbol{\gamma}^T + \mathrm{E}\left(\tau_t^{-1}\right)\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right\} + \frac{T}{2}\boldsymbol{\Sigma} = \mathbf{0},
\end{aligned} \tag{7}
$$

Therefore, the update for parameters $\boldsymbol{\mu}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\Sigma}$ is

$$
\begin{aligned}
\boldsymbol{\mu}^{(k+1)} &= \frac{\sum_{t=1}^{T}\left(\mathrm{E}\left(\tau_t\right)\mathbf{x}_t - \boldsymbol{\gamma}^{(k)}\right)}{\sum_{t=1}^{T}\mathrm{E}\left(\tau_t\right)}, \\
\boldsymbol{\gamma}^{(k+1)} &= \frac{\sum_{t=1}^{T}\left(\mathbf{x}_t - \boldsymbol{\mu}^{(k+1)}\right)}{\sum_{t=1}^{T}\mathrm{E}\left(\tau_t^{-1}\right)}, \\
\boldsymbol{\Sigma}^{(k+1)} &= \frac{1}{T}\sum_{t=1}^{T}\left\{\mathrm{E}\left(\tau_t\right)(\mathbf{x}_t - \boldsymbol{\mu}^{(k+1)})(\mathbf{x}_t - \boldsymbol{\mu}^{(k+1)})^T - (\mathbf{x}_t - \boldsymbol{\mu}^{(k+1)})\left(\boldsymbol{\gamma}^{(k+1)}\right)^T\right. \\
&\qquad \left. - \boldsymbol{\gamma}^{(k+1)}(\mathbf{x}_t - \boldsymbol{\mu}^{(k+1)})^T + \mathrm{E}\left(\tau_t^{-1}\right)\boldsymbol{\gamma}^{(k+1)}\left(\boldsymbol{\gamma}^{(k+1)}\right)^T\right\},
\end{aligned} \tag{8}
$$

3 Acceleration: Parameter Expanded (PX) EM

The parameter expanded model for the classical GH Skew t model is

$$
\begin{aligned}
\mathbf{x}|\tau &\sim \mathcal{N}\left(\boldsymbol{\mu}_* + \frac{1}{\tau}\boldsymbol{\gamma}_*, \frac{1}{\tau}\boldsymbol{\Sigma}_*\right), \\
\tau &\sim \alpha \cdot \mathrm{Gamma}\left(\frac{\nu_*}{2}, \frac{\nu_*}{2}\right),
\end{aligned} \tag{9}
$$

where $\{\boldsymbol{\mu}_*, \boldsymbol{\gamma}_*, \nu_*, \boldsymbol{\Sigma}_*, \alpha\} = \boldsymbol{\theta}_*$ is the expanded parameter set. The transformation from $\boldsymbol{\theta}_*$ to $\boldsymbol{\theta}$ is

$$
\boldsymbol{\mu} = \boldsymbol{\mu}_*, \quad \boldsymbol{\gamma} = \boldsymbol{\gamma}_*/\alpha, \quad \nu = \nu_*, \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_*/\alpha. \tag{10}
$$

The corresponding E step and M step are almost the same with a little modification:

E Step Just replace the minimum sufficient statistics with the following forms:

$$
\begin{aligned}
\mathrm{E}\left(\tau\right)_* &= \alpha \cdot \mathrm{E}\left(\tau\right) \\
\mathrm{E}\left(\tau^{-1}\right)_* &= \alpha^{-1} \cdot \mathrm{E}\left(\tau^{-1}\right) \\
\mathrm{E}\left(\log\tau\right)_* &= \log\alpha + \mathrm{E}\left(\log\tau\right)
\end{aligned} \tag{11}
$$

M step The update for $\boldsymbol{\mu}_*, \boldsymbol{\gamma}_*, \nu_*, \boldsymbol{\Sigma}_*$ keeps the same, and update for α is:

$$
\alpha^{(k+1)} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left(\tau_t\right). \tag{12}
$$

4 Numerical Simulation

We generate the GH Skew t data with $N = 20$, $T = 200$, and compare the EM and PXEM algorithms. We starts $\boldsymbol{\Sigma}^{(0)}$ from sample covariance matrix and compare the convergence of objective value. It is clear that the PXEM algorithm has the benefit of faster convergence speed. However, due to the numerical issue in the E step, the convergence is not stable for both algorithms.

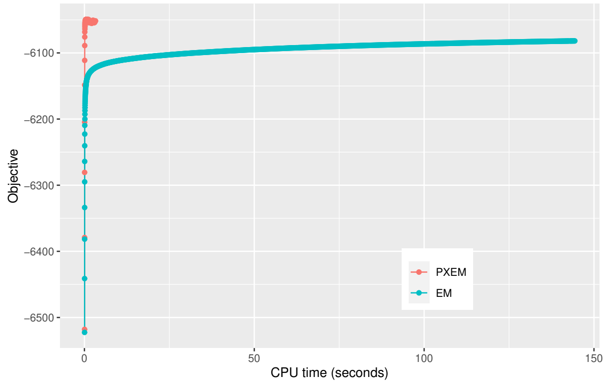

Figure 1: Objective changing of EM vs PXEM algorithm.

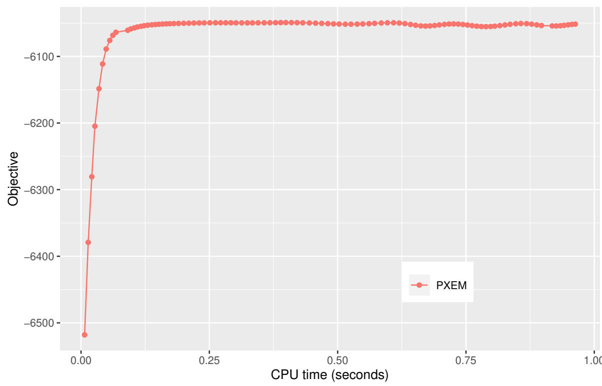

Figure 2: Objective change of PXEM algorithm.

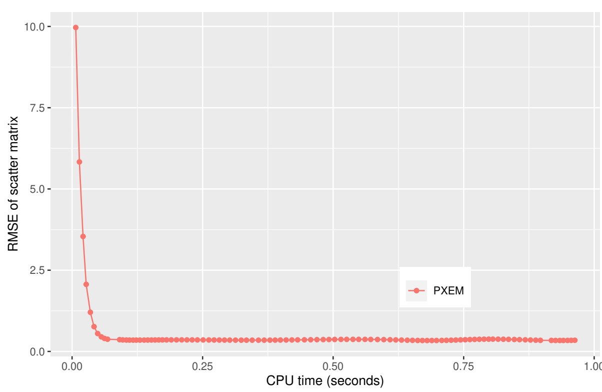

Figure 3: $\frac{\|\boldsymbol{\Sigma}^{(k)}-\boldsymbol{\Sigma}\|_F}{\|\boldsymbol{\Sigma}\|_F}$ of PXEM algorithm.

References

[1] R. Zhou and D. P. Palomar, "Accelerating the multivariate skew t parameter estimation," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 251–255.