Maximum Likelihood Estimation for a Student-t Distribution

September 19, 2017

1 Univariate Student-t

1.1 Problem Formulation

Suppose a univariate time series $\{y_t\}$ follows a Student-t distribution: $y_t \overset{i.i.d.}{\sim} t\big(\mu, \sigma^2, \nu\big)$, then the pdf of y_t is

$$f_t\left(y_t|\mu, \sigma^2, \nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{\left(y_t-\mu\right)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}. \tag{1}$$

The log-likelihood of $\{y_t\}$ is

$$l\left(\{y_t\}|\mu, \sigma^2, \nu\right) = \sum_{t=1}^{T}\left\{-\frac{\nu+1}{2}\log\left(\nu + \frac{\left(y_t-\mu\right)^2}{\sigma^2}\right) - \log\left(\sigma\right) + \log\left(\Gamma\left(\frac{\nu+1}{2}\right)\right) + \frac{\nu}{2}\log\left(\nu\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{1}{2}\log\left(\pi\right)\right\}. \tag{2}$$

which is very complicated and does not have closed-form maximizer.

1.2 EM Algorithm

Interestingly, the Student-t distribution $y_t \sim t\left(\mu, \sigma^2, \nu\right)$ can be regarded as a Gaussian mixture:

$$y_t|\mu, \sigma^2, \tau_t \sim \mathcal{N}\left(\mu, \sigma^2/\tau_t\right), \tag{3}$$

$$\tau_t \sim Gamma\left(\nu/2, \nu/2\right), \tag{4}$$

where the density function of the above Gamma distribution is

$$f_g\left(\tau_t|\frac{\nu}{2}, \frac{\nu}{2}\right) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\tau_t^{\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2}\tau_t\right). \tag{5}$$

Thus, we can use the EM algorithm to solve this optimization problem by regarding $\{\tau_t\}$ as latent data. The complete data log-likelihood is

$$
\begin{aligned}
l\left(\{y_t\},\{\tau_t\}|\mu,\sigma^2,\nu\right) =& \sum_{t=1}^{T}\log\left(f_N\left(y_t|\mu,\frac{\sigma^2}{\tau_t}\right)f_g\left(\tau_t|\frac{\nu}{2},\frac{\nu}{2}\right)\right)\\
=& \sum_{t=1}^{T}\left\{\log\left(f_N\left(y_t|\mu,\frac{\sigma^2}{\tau_t}\right)\right)+\log\left(f_g\left(\tau_t|\frac{\nu}{2},\frac{\nu}{2}\right)\right)\right\}.\\
=& \sum_{t=1}^{T}\left\{\log\left(\frac{1}{\sqrt{2\pi\sigma^2/\tau_t}}\exp\left(-\frac{1}{2\sigma^2/\tau_t}\left(y_t-\mu\right)^2\right)\right)+\log\left(\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\tau_t^{\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2}\tau_t\right)\right)\right\}\\
=& \sum_{t=1}^{T}\left\{-\frac{\tau_t}{2\sigma^2}\left(y_t-\mu\right)^2-\log\left(\sigma\right)\right.\\
&\left.-\frac{\nu}{2}\tau_t+\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)+\frac{\nu+1}{2}\log\left(\tau_t\right)-\log\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right\}+const.
\end{aligned}
\tag{6}
$$

The conditional density function of $\{\tau_t\}$ under the current estimates and $\{y_t\}$ is

$$
\begin{aligned}
f\left(\tau_t|\mu^k,\left(\sigma^k\right)^2,\nu^k,\{y_t\}\right) =& \frac{f_N\left(y_t|\mu^k,\frac{\left(\sigma^k\right)^2}{\tau_t}\right)f_g\left(\tau_t|\frac{\nu^k}{2},\frac{\nu^k}{2}\right)}{f_t\left(y_t|\mu^k,\left(\sigma^k\right)^2,\nu^k\right)}\\
=& \frac{\frac{1}{\sqrt{2\pi(\sigma^k)^2/\tau_t}}\exp\left(-\frac{1}{2(\sigma^k)^2/\tau_t}\left(y_t-\mu^k\right)^2\right)\frac{\left(\frac{\nu^k}{2}\right)^{\frac{\nu^k}{2}}}{\Gamma\left(\frac{\nu^k}{2}\right)}\tau_t^{\frac{\nu^k}{2}-1}\exp\left(-\frac{\nu^k}{2}\tau_t\right)}{\frac{\Gamma\left(\frac{\nu^k+1}{2}\right)}{\sqrt{\nu^k\pi}\sigma\Gamma\left(\frac{\nu^k}{2}\right)}\left(1+\frac{(y_t-\mu^k)^2}{\nu^k\sigma^2}\right)^{-\frac{\nu^k+1}{2}}}\\
\propto& \tau_t^{\frac{\nu^k-1}{2}}\exp\left(-\left(\frac{\left(y_t-\mu^k\right)^2}{2\left(\sigma^k\right)^2}+\frac{\nu^k}{2}\right)\tau_t\right).
\end{aligned}
\tag{7}
$$

Therefore,

$$
\tau_t|\mu^k,\left(\sigma^k\right)^2,\nu^k,\{y_t\}\sim Gamma\left(\frac{\nu^k+1}{2},\frac{\left(\sigma^k\right)^{-2}\left(y_t-\mu^k\right)^2+\nu^k}{2}\right).
\tag{8}
$$

The expected complete data log-likelihood, given the conditional distribution of $\{\tau_t\}$, is

$$
\begin{aligned}
Q\left(\mu,\sigma^2,\nu|\mu^k,\left(\sigma^k\right)^2,\nu^k\right) =& \mathsf{E}_{\tau_t|\mu^k,(\sigma^k)^2,\nu^k,\{y_t\}}\left(l\left(\{y_t\},\{\tau_t\}|\mu,\sigma^2,\nu\right)\right)\\
=& \sum_{t=1}^{T}\left\{-\frac{\mathsf{E}\left(\tau_t\right)}{2\sigma^2}\left(y_t-\mu\right)^2-\log\left(\sigma\right)\right.\\
&\left.-\frac{\nu}{2}\mathsf{E}\left(\tau_t\right)+\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\mathsf{E}\left(\log\left(\tau_t\right)\right)-\log\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right\}+const.
\end{aligned}
\tag{9}
$$

The optimization of μ and σ^2 are decoupled with the optimization of ν. There exist the closed-form solutions for μ^{k+1} and $\left(\sigma^{k+1}\right)^2$:

$$
\mu^{k+1}=\frac{\sum_{t=1}^{T}\mathsf{E}\left(\tau_t\right)y_t}{\sum_{t=1}^{T}\mathsf{E}\left(\tau_t\right)},
\tag{10}
$$

and

$$
\left(\sigma^{k+1}\right)^2=\frac{1}{T}\sum_{t=1}^{T}\mathsf{E}\left(\tau_t\right)\left(y_t-\mu^{k+1}\right)^2,
\tag{11}
$$

where

$$\mathsf{E}\left(\tau_t\right) = \frac{\nu^k + 1}{\nu^k + \left(\sigma^k\right)^{-2}\left(y_t - \mu^k\right)^2}. \tag{12}$$

And ν^{k+1} can be found by one-dimensional search.

In addition, the convergence speed of the EM algorithm can be accelerated by the ECME algorithm, which obtains ν^{k+1} by directly optimizing the objective function $l\left(\{y_t\}\,|\mu^{k+1}, \left(\sigma^{k+1}\right)^2, \nu\right)$ instead in every iteration [1]. We can use sample mean and variance as the initial point, and it usually take just several iterations to converge.

1.3 Block MM

It is difficult to conduct MM for all variables jointly, therefore, we partition the three variables into two blocks: μ and σ^2 as a block, and ν as a block. Then we apply the block MM to solve the optimization problem iteratively. In the k-th iteration, we first conduct the minorization and maximization for μ and σ^2 , with ν fixed as ν^k, and then optimize ν, with μ and σ^2 fixed as μ^{k+1} and $\left(\sigma^{k+1}\right)^2$.

For the minorization and maximization for μ and σ^2, if we apply

$$\log\left(x\right) \leq \frac{1}{x_k}\left(x - x_k\right) + \log\left(x_k\right), \tag{13}$$

then, at $\left(\mu^k, \left(\sigma^k\right)^2, \nu^k\right)$, $l\left(\{y_t\}\,|\mu, \sigma^2, \nu^k\right)$ is minorized by

$$\begin{aligned} l\left(\{y_t\}\,|\mu, \sigma^2, \nu^k\right) &= \sum_{t=1}^{T}\left\{-\frac{\nu^k+1}{2}\log\left(\nu^k + \frac{\left(y_t-\mu\right)^2}{\sigma^2}\right) - \log\left(\sigma\right) + const.\right\} \\ &\geq \sum_{t=1}^{T}\left\{-\frac{\nu^k+1}{2\left(\nu^k + \left(\sigma^k\right)^{-2}\left(y_t-\mu^k\right)^2\right)}\frac{\left(y_t-\mu\right)^2}{\sigma^2} - \log\left(\sigma\right) + const.\right\} \end{aligned} \tag{14}$$

which is very similar to (9) and will result in the same update formulas for μ and σ^2.

The optimization of ν is easy to solve. Since $l\left(\{y_t\}\,|\mu^{k+1}, \left(\sigma^{k+1}\right)^2, \nu\right)$is a scalar function of a scalar variable, we can find ν^{k+1} by one-dimensional search.

In short, the algorithm derived based on the block MM framework is the same with above ECME algorithm.

2 Multivariate Student-t

2.1 Problem Formulation

Suppose a mutivariate time series $\{\mathbf{y}_t\}$ follows a p-dimensional Student-t distribution: $\mathbf{y}_t \overset{i.i.d.}{\sim} t_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right)$, then the pdf of $\mathbf{y}_t$ is

$$f_t\left(\mathbf{y}_t|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right) = \frac{\Gamma\left(\left(\nu+p\right)/2\right)}{\Gamma\left(\nu/2\right)\left(\nu\pi\right)^{p/2}\det\left(\boldsymbol{\Sigma}\right)^{1/2}}\{1 + \nu^{-1}(\mathbf{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\}^{-(\nu+p)/2}. \tag{15}$$

The log-likelihood of $\{\mathbf{y}_t\}$ is

$$\begin{aligned} l\left(\{\mathbf{y}_t\}\,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right) = \sum_{t=1}^{T}\Bigg\{&-\frac{\nu+p}{2}\log\left(\nu + (\mathbf{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right) - \frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right) \\ &+ \log\left(\Gamma\left(\frac{\nu+p}{2}\right)\right) + \frac{\nu}{2}\log\left(\nu\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{p}{2}\log\left(\pi\right)\Bigg\}, \end{aligned} \tag{16}$$

which is very complicated and does not have closed-form maximizer.

2.2 EM Algorithm

Interestingly, the Student-t distribution $\mathbf{y}_t \overset{i.i.d.}{\sim} t_p\left(\boldsymbol{\mu}, \mathbf{\Sigma}, \nu\right)$ can be regarded as a Gaussian mixture:

$$\mathbf{y}_t|\boldsymbol{\mu}, \mathbf{\Sigma}, \tau_t \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{\Sigma}/\tau_t\right), \tag{17}$$

$$\tau_t \sim Gamma\left(\nu/2, \nu/2\right). \tag{18}$$

Thus, we can use the EM algorithm to solve this optimization problem by regarding $\{\tau_t\}$ as latent data. The complete data log-likelihood is

$$\begin{aligned} l\left(\{\mathbf{y}_t\}, \{\tau_t\}|\boldsymbol{\mu}, \mathbf{\Sigma}, \nu\right) &= \sum_{t=1}^{T} \log\left(f_N\left(\mathbf{y}_t|\boldsymbol{\mu}, \frac{\mathbf{\Sigma}}{\tau_t}\right) f_g\left(\tau_t|\frac{\nu}{2}, \frac{\nu}{2}\right)\right) \\ &= \sum_{t=1}^{T} \left\{\log\left(f_N\left(\mathbf{y}_t|\boldsymbol{\mu}, \frac{\mathbf{\Sigma}}{\tau_t}\right)\right) + \log\left(f_g\left(\tau_t|\frac{\nu}{2}, \frac{\nu}{2}\right)\right)\right\}. \\ &= \sum_{t=1}^{T} \left\{\log\left(\frac{1}{\sqrt{2\pi \det\left(\mathbf{\Sigma}^k/\tau_t\right)}} \exp\left(-\frac{\tau_t}{2}(\mathbf{y}_i - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right)\right) + \log\left(\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} \tau_t^{\frac{\nu}{2}-1} \exp\left(-\frac{\nu}{2}\tau_t\right)\right)\right\} \\ &= \sum_{t=1}^{T} \left\{-\frac{\tau_t}{2}(\mathbf{y}_i - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) - \frac{1}{2}\log\left(\det\left(\mathbf{\Sigma}^k\right)\right)\right. \\ &\qquad \left. -\frac{\nu}{2}\tau_t + \frac{\nu}{2}\log\left(\frac{\nu}{2}\right) + \left(\frac{\nu+p}{2} - 1\right)\log\left(\tau_t\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right\} + const. \end{aligned} \tag{19}$$

The conditional density function of $\{\tau_t\}$ under the current estimates and $\{\mathbf{y}_t\}$ is

$$\begin{aligned} f\left(\tau_t|\boldsymbol{\mu}^k, \mathbf{\Sigma}^k, \nu^k, \{\mathbf{y}_t\}\right) &= \frac{f_N\left(\mathbf{y}_t|\boldsymbol{\mu}^k, \frac{\mathbf{\Sigma}^k}{\tau_t}\right) f_g\left(\tau_t|\frac{\nu^k}{2}, \frac{\nu^k}{2}\right)}{f_t\left(\mathbf{y}_t|\boldsymbol{\mu}^k, \mathbf{\Sigma}^k, \tau_t\right)} \\ &= \frac{\frac{1}{\sqrt{2\pi \det\left(\mathbf{\Sigma}^k/\tau_t\right)}} \exp\left(-\frac{\tau_t}{2}(\mathbf{y}_i - \boldsymbol{\mu}^k)^T \left(\mathbf{\Sigma}^k\right)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}^k)\right) \frac{\left(\frac{\nu^k}{2}\right)^{\frac{\nu^k}{2}}}{\Gamma\left(\frac{\nu^k}{2}\right)} \tau_t^{\frac{\nu^k}{2}-1} \exp\left(-\frac{\nu^k}{2}\tau_t\right)}{\frac{\Gamma((\nu^k+p)/2)}{\Gamma(\nu^k/2)(\nu^k\pi)^{p/2}\det\left(\mathbf{\Sigma}^k\right)^{1/2}}\{1 + (\nu^k)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}^k)^T \left(\mathbf{\Sigma}^k\right)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}^k)\}^{-(\nu^k+p)/2}} \\ &\propto \tau_t^{\frac{\nu^k+p}{2}-1} \exp\left(-\frac{(\mathbf{y}_i - \boldsymbol{\mu}^k)^T \left(\mathbf{\Sigma}^k\right)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}^k) + \nu^k}{2}\tau_t\right). \end{aligned} \tag{20}$$

Therefore,

$$\tau_t|\boldsymbol{\mu}^k, \mathbf{\Sigma}^k, \nu^k, \{\mathbf{y}_t\} \sim Gamma\left(\frac{\nu^k+p}{2}, \frac{(\mathbf{y}_i - \boldsymbol{\mu}^k)^T \left(\mathbf{\Sigma}^k\right)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}^k) + \nu^k}{2}\right). \tag{21}$$

The expected complete data log-likelihood, given the conditional distribution of $\{\tau_t\}$, is

$$
\begin{aligned}
Q\left(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu|\boldsymbol{\mu}^k,\boldsymbol{\Sigma}^k,\nu^k\right) =& \mathsf{E}_{\tau_t|\boldsymbol{\mu}^k,\boldsymbol{\Sigma}^k,\nu^k,\{\mathbf{y}_t\}}\left(l\left(\{\mathbf{y}_t\},\{\tau_t\}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right)\right)\\
=&\sum_{t=1}^{T}\Bigg\{-\frac{\mathsf{E}\left(\tau_t\right)}{2}(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})-\frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}^k\right)\right)\\
&-\frac{\nu}{2}\mathsf{E}\left(\tau_t\right)+\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\mathsf{E}\left(\log\left(\tau_t\right)\right)-\log\left(\Gamma\left(\frac{\nu}{2}\right)\right)\Bigg\}+const.
\end{aligned}
\tag{22}
$$

The optimization of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are decoupled with the optimization of ν. There exist the closed-form solutions for $\boldsymbol{\mu}^{k+1}$ and $\boldsymbol{\Sigma}^{k+1}$:

$$
\boldsymbol{\mu}^{k+1}=\frac{\sum_{t=1}^{T}\mathsf{E}\left(\tau_t\right)\mathbf{y}_t}{\sum_{t=1}^{T}\mathsf{E}\left(\tau_t\right)}, \tag{23}
$$

and

$$
\boldsymbol{\Sigma}^{k+1}=\frac{1}{T}\sum_{t=1}^{T}\mathsf{E}\left(\tau_t\right)\left(\mathbf{y}_t-\boldsymbol{\mu}^{k+1}\right)\left(\mathbf{y}_t-\boldsymbol{\mu}^{k+1}\right)^T, \tag{24}
$$

where

$$
\mathsf{E}\left(\tau_t\right)=\frac{\nu^k+p}{\nu^k+(\mathbf{y}_i-\boldsymbol{\mu}^k)^T\left(\boldsymbol{\Sigma}^k\right)^{-1}(\mathbf{y}_i-\boldsymbol{\mu}^k)}. \tag{25}
$$

And ν^{k+1} can be found by one-dimensional search.

Similarly, the convergence speed of the EM algorithm can be accelerated by the ECME algorithm, which obtains ν^{k+1} by directly optimizing the objective function $l\left(\{\mathbf{y}_t\}|\boldsymbol{\mu}^{k+1},\boldsymbol{\Sigma}^{k+1},\nu\right)$ instead in every iteration [1]. We can use sample mean and variance as the initial point, and it usually take just several iterations to converge.

2.3 Block MM

Similarly, we can partition the three variables into two blocks: $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as a block, and ν as a block. Then we apply the block MM to solve the optimization problem iteratively. In the k-th iteration, we first conduct the minorization and maximization for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, with ν fixed as ν^k, and then optimize ν, with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ fixed as $\boldsymbol{\mu}^{k+1}$ and $\boldsymbol{\Sigma}^{k+1}$.

For the minorization and maximization for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, if we apply

$$
\log\left(x\right)\leq\frac{1}{x_k}\left(x-x_k\right)+\log\left(x_k\right), \tag{26}
$$

then, at $\left(\boldsymbol{\mu}^k,\boldsymbol{\Sigma}^k,\nu^k\right)$, $l\left(\{\mathbf{y}_t\}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu^k\right)$ is minorized by

$$
\begin{aligned}
l\left(\{\mathbf{y}_t\}|\boldsymbol{\mu},\,\boldsymbol{\Sigma},\,\nu^k\right)=&\sum_{t=1}^{T}\left\{-\frac{\nu^k+p}{2}\log\left(\nu^k+(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})\right)-\frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right)+const.\right\}\\
\geq&\sum_{t=1}^{T}\left\{-\frac{\nu^k+p}{2\left(\nu^k+(\mathbf{y}_i-\boldsymbol{\mu}^k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}^k)\right)}(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})-\frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right)+const.\right\},
\end{aligned}
\tag{27}
$$

which is very similar to (9) and will result in the same update formulas for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The optimization of ν is easy to solve. Since $l\left(\{y_t\}|\boldsymbol{\mu}^{k+1},\boldsymbol{\Sigma}^{k+1},\nu\right)$is a scalar function of a scalar variable, we can find ν^{k+1} by one-dimensional search.

In short, the algorithm derived based on the block MM framework is the same with above ECME algorithm.

References

[1] C. Liu and D. B. Rubin, "ML estimation of the t-distribution using EM and its extensions, ECM and ECME," *Stat. Sin.*, vol. 5, no. 1, pp. 19–39, 1995.