EM for Student's $\boldsymbol{t}$ Mixture Models

Table of contents

1 Notation

You will see me being too careful about notation in this derivation. This is because this has not been an easy derivation and being explicit actually helps.

- Let y_j be the observed values of the j-th data vector (so y_j is a vector, and I'm not using vector notation to save up some typing).
- Let z_j be the *integer* value of the corresponding mixture component index.
- Let u_j be the *scalar* value of the corresponding latent variable in the Gaussian scale mixture.
- Let g denote the number of mixture components and n denote the number of training examples.
- Let θ denote all parameters, i.e., $\theta = \{\pi, \boldsymbol{\mu}_{1:g}, \boldsymbol{\Sigma}_{1:g}, v_{1:g}\}$.

Finally, the bulk of this derivation is self-contained and, for the most part, you do not need to refer to the original paper unless you point you there. The paper also contains some errors here and there so be careful. Please send me an email if you spot an error in this derivation.

2 Joint density

To prepare for Section 3, we want to express $p_{Z_j,U_j,Y_j}(z_j, u_j, y_j| \theta)$ in distributions that we already know (i.e., as defined by the forward model):

$$p_{Z_j,U_j,Y_j}(z_j, u_j, y_j; \theta) = p_{Y_j|Z_j,U_j}(y_j| z_j, u_j; \theta) p_{U_j|Z_j}(u_j| z_j; \theta) p_{Z_j}(z_j; \theta)$$

where:

$$\begin{aligned}
p_{Z_j}(z_j; \theta) &= \prod_{i=1}^{g} \pi_i^{\mathbb{I}(z_j=i)} \\
p_{U_j|Z_j}(u_j| z_j; \theta) &= \mathrm{Ga}(u_j| \frac{v_{z_j}}{2}, \frac{v_{z_j}}{2}) \\
&= \prod_{i=1}^{g} \mathrm{Ga}(u_j| \frac{v_i}{2}, \frac{v_i}{2})^{\mathbb{I}(z_j=i)} \\
p_{Y_j|U_j,Z_j}(y_j| u_j, z_j; \theta) &= \mathcal{N}(y_j| \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j}/u_j) \\
&= \prod_{i=1}^{g} \mathcal{N}(y_j| \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i/u_j)^{\mathbb{I}(z_j=i)}
\end{aligned}$$

The above notation of multiplying g terms of which $g-1$ terms is just 1 is a notation used to simplify later calculations.

3 Complete data log-likelihood

For the complete data log likelihood (we'll take expectation later in Section 4.1), this joint distribution factors and we get

$$\sum_{j=1}^{n} \log p_{Z_j,U_j,Y_j}(z_j, u_j, y_j; \theta) = \sum_{j=1}^{n} \log p_{Y_j|Z_j,U_j}(y_j| z_j, u_j; \theta) + \sum_{j=1}^{n} \log p_{U_j|Z_j}(u_j| z_j; \theta) + \sum_{j=1}^{n} \log p_{Z_j}(z_j; \theta)$$

where each term is concerned about different parameters (which is great isn't it?):

$$\begin{aligned}
\sum_{j=1}^{n} \log p_{Z_j}(z_j; \theta) &= \sum_{j=1}^{n} \log \prod_{i=1}^{g} \pi_i^{\mathbb{I}(z_j=i)} \\
&= \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j = i) \log \pi_i
\end{aligned}$$

$$
\begin{aligned}
\sum_{j=1}^{n} \log p_{U_j|Z_j}(u_j|\, z_j;\theta) \quad &= \quad \sum_{j=1}^{n} \log \prod_{i=1}^{g} \mathrm{Ga}(u_j|\, \frac{v_i}{2}, \frac{v_i}{2})^{\mathbb{I}(z_j=i)} \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \log \mathrm{Ga}(u_j|\, \frac{v_i}{2}, \frac{v_i}{2}) \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \log \left[\frac{(v_i/2)^{v_i/2}}{\Gamma(v_i/2)} \cdot u_j^{v_i/2-1} \cdot \exp\{-u_j \cdot (v_i/2)\} \right] \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \left(\frac{v_i}{2} \log \frac{v_i}{2} - \log \Gamma(v_i/2) + \left(\frac{v_i}{2} - 1 \right) \log u_j - \frac{v_i}{2} \cdot u_j \right) \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \left(\frac{v_i}{2} \log \frac{v_i}{2} - \log \Gamma(v_i/2) + \frac{v_i}{2} (\log u_j - u_j) - \log u_j \right) \\
&=_{\mathrm{grad}} \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \left(\frac{v_i}{2} \log \frac{v_i}{2} - \log \Gamma(v_i/2) + \frac{v_i}{2} (\log u_j - u_j) \right)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{j=1}^{n} \log p_{Y_j|Z_j,U_j}(y_j|\, z_j, u_j;\theta) \quad &= \quad \sum_{j=1}^{n} \log \prod_{i=1}^{g} \mathcal{N}(y_j|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i/u_j)^{\mathbb{I}(z_j=i)} \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \log \mathcal{N}(y_j|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i/u_j) \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \log \left[\frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_i/u_j|^{1/2}} \exp\left\{ -\frac{1}{2} (y_j - \boldsymbol{\mu}_j)^T (\boldsymbol{\Sigma}_i/u_j)^{-1} (y_j - \boldsymbol{\mu}_j) \right\} \right] \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \log \left[\frac{u_j^{1/2}}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{ -\frac{u_j}{2} (y_j - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_i^{-1} (y_j - \boldsymbol{\mu}_j) \right\} \right] \\
&= \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \left(-\frac{p}{2} \log (2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \frac{1}{2} \log u_j - \frac{u_j}{2} (y_j - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_i^{-1} (y_j - \boldsymbol{\mu}_j) \right) \\
&=_{\mathrm{grad}} \quad \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbb{I}(z_j=i) \left(-\frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{u_j}{2} (y_j - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_i^{-1} (y_j - \boldsymbol{\mu}_j) \right)
\end{aligned}
$$

4 Expected data log-likelihood

4.1 Objective for each parameter group

Now comes the time to compute the expected log likelihood over Z_j's and U_j's. We don't have to take this expectation over these random variables together in one expectation. Instead, we can break it into an inner and an outer expectation and take the two sequentially. Conveniently, the expectation will by linearity factor across the three terms so we can compute the expectation for one term at a time:

First term (objective for for π)

$$\begin{aligned}
&\mathbb{E}_{Z_j \sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{j=1}^{n} \log p_{Z_j}(Z_j;\theta)\right] \\
=\;& \mathbb{E}_{Z_j \sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{I}(Z_j = i)\log \pi_i\right] \\
=\;& \sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{E}_{Z_j \sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}[\mathbb{I}(Z_j = i)]\log \pi_i \\
=\;& \sum_{i=1}^{g}\sum_{j=1}^{n} \underbrace{p_{Z_j|Y_j}(i|\, y_j;\theta_{\text{old}})}_{\tau_{ij}}\log \pi_i
\end{aligned}$$

Second term (objective for v_i's)

$$\begin{aligned}
&\mathbb{E}_{Z_j, U_j \sim p_{Z_j,U_j|Y_j}(\cdot,\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{j=1}^{n} \log p_{U_j|Z_j}(U_j|\, Z_j,\theta)\right] \\
=\;& \mathbb{E}_{Z_j \sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\mathbb{E}_{U_j \sim p_{U_j|Z_j,Y_j}(\cdot|\, Z_j,y_j;\theta_{\text{old}})}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{I}(Z_j = i)\left(\frac{v_i}{2}\log\frac{v_i}{2} - \log\Gamma(v_i/2) + \frac{v_i}{2}(\log U_j - U_j) - \log U_j\right)\right]\right] \\
=\;& \mathbb{E}_{Z_j \sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{I}(Z_j = i)\left(\frac{v_i}{2}\log\frac{v_i}{2} - \log\Gamma(v_i/2) + \frac{v_i}{2}(\mathbb{E}_{U_j \sim p_{U_j|Z_j,Y_j}(\cdot|\, Z_j,y_j;\theta_{\text{old}})}[\log U_j] - \mathbb{E}_{U_j \sim p_{U_j|Z_j,Y_j}(\cdot|\, Z_j,y_j;\theta_{\text{old}})}[U_j]) - \mathbb{E}_{U_j \sim p_{U_j|Z_j,Y_j}(\cdot|\, Z_j,y_j;\theta_{\text{old}})}[\log U_j]\right)\right] \\
=\;& \mathbb{E}_{Z_j \sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{I}(Z_j = i)\left(\frac{v_i}{2}\log\frac{v_i}{2} - \log\Gamma(v_i/2) + \frac{v_i}{2}\left(\mathbb{E}_{U_j \sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[\log U_j] - \mathbb{E}_{u_j \sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[Y_j]\right) - \mathbb{E}_{U_j \sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[\log U_j]\right)\right] \quad (*)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{E}_{Z_j \sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}[\mathbb{I}(Z_j = i)] \left(\frac{v_i}{2}\log\frac{v_i}{2} - \log\Gamma(v_i/2) + \frac{v_i}{2}\left(\mathbb{E}_{U_j\sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[\log U_j] - \mathbb{E}_{U_j\sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[U_j]\right) - \mathbb{E}_{u_j\sim p_{U_j|Z_j,Y_j}(u_j|1,y_j;\theta_{\text{old}})}[\log U_j]\right) \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n} \underbrace{p_{Z_j|Y_j}(i|\, y_j;\theta_{\text{old}})}_{\tau_{ij}} \left(\frac{v_i}{2}\log\frac{v_i}{2} - \log\Gamma(v_i/2) + \frac{v_i}{2}\left(\underbrace{\mathbb{E}_{U_j\sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[\log U_j]}_{\text{see paper}} - \underbrace{\mathbb{E}_{U_j\sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[U_j]}_{\text{see paper}}\right) - \mathbb{E}_{U_j\sim p_{U_j|Z_j,Y_j}(\cdot|1,y_j;\theta_{\text{old}})}[\log U_j]\right)
\end{aligned}
$$

Third term (objective for $\boldsymbol{\mu}_i$'s and $\boldsymbol{\Sigma}_i$'s)

$$
\begin{aligned}
&\mathbb{E}_{Z_j,U_j\sim p_{Z_j,U_j|Y_j}(\cdot,\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{j=1}^{n} \log p_{Y_j|Z_j,U_j}(y_j|\, Z_j, U_j;\theta)\right] \\
&= \mathbb{E}_{Z_j\sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\mathbb{E}_{U_j\sim p_{U_j|Z_j,Y_j}(\cdot|\, Z_j,y_j;\theta_{\text{old}})}\left[\sum_{j=1}^{n} \log p_{Y_j|Z_j,U_j}(y_j|\, Z_j, U_j;\theta)\right]\right] \\
&= \mathbb{E}_{Z_j\sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\mathbb{E}_{U_j\sim p_{U_j|Z_j,Y_j}(\cdot|\, Z_j,y_j;\theta_{\text{old}})}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{I}(Z_j = i)\left(-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{U_j}{2}(y_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_i^{-1}(y_j - \boldsymbol{\mu}_j)\right)\right]\right] \\
&= \mathbb{E}_{Z_j\sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{I}(Z_j = i)\left(-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\mathbb{E}_{U_j\sim p_{U_j|Y_j,Z_j}(\cdot|\, y_j,Z_j;\theta)}[U_j](y_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_i^{-1}(y_j - \boldsymbol{\mu}_j)\right)\right] \\
&= \mathbb{E}_{Z_j\sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{I}(Z_j = i)\left(-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\mathbb{E}_{U_j\sim p_{U_j|Y_j,Z_j}(\cdot|\, y_j,1;\theta)}[U_j](y_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_i^{-1}(y_j - \boldsymbol{\mu}_j)\right)\right] \quad (*) \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n} \underbrace{\mathbb{E}_{Z_j\sim p_{Z_j|Y_j}(\cdot|\, y_j;\theta_{\text{old}})}[\mathbb{I}(Z_j = i)]}_{\tau_{ij}}\left(-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\underbrace{\mathbb{E}_{U_j\sim p_{U_j|Y_j,Z_j}(\cdot|\, y_j,1;\theta)}[U_j]}(y_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_i^{-1}(y_j - \boldsymbol{\mu}_j)\right)
\end{aligned}
$$

($*$) The term behind $\mathbb{I}(Z_j = 1)$ is taken into account (i.e., not zero) only when $Z_j = 1$, so we can safely set $Z_j = 1$ in that term. By doing so, that term turns into a constant rather a random variable. *I believe this is the most important step in the whole derivation; neither the original paper nor Murphy's book talk about this step, and I was stuck there for some time.*

4.2 E-step: evaluating expectations

In the E-step, we compute the stuff enclosed in curly brackets in last subsection. Please refer to Section 6 of the original paper for this stuff – that section is easy to follow and error-free.

4.3 M-step: solving the objectives

In the M-step, we solve for the maxima of the objectives in Section 4.1. Please refer to Section 7 of the original paper for this stuff – that section is easy to follow but contains an error in Equation 32, which I've commented about in my code.