A Derivation of the EM Updates for Finding the Maximum Likelihood Parameter Estimates of the Student's t Distribution

Carl Scheffler

First draft: 22 September 2008

Contents

1 The Student's t Distribution

Univariate

$$\begin{aligned}\text{Student}\left(x \,|\, \mu, \lambda, \nu\right) &\equiv \int_0^\infty \text{Normal}\left(x \,\middle|\, \mu, (\lambda\eta)^{-1}\right) \text{Gamma}\left(\eta \,|\, \nu/2, \nu/2\right) \,\mathrm{d}\eta \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}\end{aligned}$$

Multivariate

$$\begin{aligned}\text{Student}\left(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu\right) &\equiv \int_0^\infty \text{Normal}\left(\boldsymbol{x} \,\middle|\, \boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \text{Gamma}\left(\eta \,|\, \nu/2, \nu/2\right) \,\mathrm{d}\eta \\ &= \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{|\boldsymbol{\Lambda}|}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+D}{2}}\end{aligned}$$

where the vectors[1] $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are D-dimensional and $\boldsymbol{\Lambda}$ is $D \times D$.

2 The General EM Algorithm

We want to find the maximum likelihood estimate for a set of parameters Θ given a set of observed data X by maximizing

$$P\left(X \,|\, \Theta\right).$$

We assume that it is hard to solve this problem directly but that it is relatively easy to evaluate

$$P\left(X, Z \,|\, \Theta\right)$$

[1] They're all column vectors.

where Z is a set of latent variables such that

$$P(X \mid \Theta) = \int_Z P(X, Z \mid \Theta).$$

The EM method then involves the following steps.

1. Write down the complete data log likelihood, $\log P(X, Z \mid \Theta)$.
2. Write down the posterior latent distribution, $P(Z \mid X, \Theta)$.
3. E step: write down the expectations under the distribution $P(Z \mid X, \Theta_0)$ for all terms in the complete data log likelihood (step 1).
4. Write down the function to maximize,

$$Q(\Theta, \Theta_0) = \int_Z P(Z \mid X, \Theta_0) \log P(X, Z \mid \Theta),$$

   replacing integrals with the expectations from the E step.
5. M step: solve

$$\frac{\partial Q}{\partial \Theta} = 0$$

   to yield the update equations.

Once all of the expectation update equations (from step 3) and maximization update equations (from step 5) are known, we initialize our current estimate of the parameters Θ and update them by iterating the E and M updates until convergence. Note that a subscript 0 is used here and in the rest of the article to denote the old setting of the parameters. With each iteration the new parameter setting (found in step 5) will replace the old one.

3 Derivation of the EM Update Equations for the Univariate Student's t Distribution

To cast the Student's t distribution in the EM framework we write the likelihood for a single data point,

$$P(x_i \mid \Theta) = \text{Student}(x_i \mid \mu, \lambda, \nu).$$

By viewing this as an infinite mixture of Normal distributions,

$$P(x_i \mid \Theta) = \int_{\eta_i} \text{Normal}\left(x_i \,\middle|\, \mu, (\lambda\eta_i)^{-1}\right) \text{Gamma}(\eta_i \mid \nu/2, \nu/2),$$

we identify

$$X = \{x_i\}, \quad Z = \{\eta_i\}, \quad \Theta = \{\mu, \lambda, \nu\}$$

so that the complete likelihood function is

$$P(X, Z \mid \Theta) = \prod_{i=1}^{N} \text{Normal}\left(x_i \,\middle|\, \mu, (\lambda\eta_i)^{-1}\right) \text{Gamma}(\eta_i \mid \nu/2, \nu/2).$$

Step 1: Complete log likelihood function

$$\begin{aligned}
\log P(X, Z \mid \Theta) &= \sum_{i=1}^{N} \log \text{Normal}\left(x_i \,\middle|\, \mu, (\lambda\eta_i)^{-1}\right) + \log \text{Gamma}(\eta_i \mid \nu/2, \nu/2) \\
&= \sum_{i=1}^{N} -\frac{1}{2}\log 2\pi + \frac{1}{2}\log\lambda + \frac{1}{2}\log\eta_i - \frac{\lambda\eta_i}{2}(x_i - \mu)^2 \\
&\qquad - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log\frac{\nu}{2} + \left(\frac{\nu}{2} - 1\right)\log\eta_i - \frac{\nu}{2}\eta_i \qquad (1)
\end{aligned}$$

Step 2: Posterior latent distribution

$$\begin{aligned} P(Z \mid X, \Theta) &\propto P(X, Z \mid \Theta) \\ &= \prod_{i=1}^{N} \text{Normal}\left(x_i \mid \mu, (\lambda\eta_i)^{-1}\right) \text{Gamma}(\eta_i \mid \nu/2, \nu/2) \\ &\propto \prod_{i=1}^{N} \text{Gamma}(\eta_i \mid a_i, b_i) \end{aligned} \tag{2}$$

where the last step follows from the fact that the Gamma distribution is the conjugate prior to a Normal distribution with unknown precision. We find the parameters a_i and b_i by combining the factors from the Normal and Gamma distributions.[2]

$$\begin{aligned} P(X, Z \mid \Theta) &= \prod_{i=1}^{N} \text{Normal}\left(x_i \mid \mu, (\lambda\eta_i)^{-1}\right) \text{Gamma}(\eta_i \mid \nu/2, \nu/2) \\ &\propto \prod_{i=1}^{N} \left[\eta_i^{\frac{\nu-1}{2}} \exp\left(-\eta_i \left(\frac{\nu}{2} + \frac{\lambda}{2}(x_i - \mu)^2\right)\right)\right] \\ &\quad \text{(All factors independent of } \eta_i \text{ are taken up in the proportionality.)} \\ &\propto \prod_{i=1}^{N} \text{Gamma}\left(\eta_i \,\middle|\, \frac{\nu+1}{2}, \frac{\nu}{2} + \frac{\lambda}{2}(x_i - \mu)^2\right) \end{aligned}$$

and hence

$$\begin{aligned} a_i &= \frac{\nu+1}{2}, \\ b_i &= \frac{\nu}{2} + \frac{\lambda}{2}(x_i - \mu)^2. \end{aligned}$$

Step 3: Expectation By looking at (1) we find that we need to calculate the expectations of 1, η_i and $\log \eta_i$ under the posterior latent distribution (2).

$$\mathbb{E}[1] = 1$$

$$\begin{aligned} \mathbb{E}[\eta_i] &= \int_Z \eta_i \prod_{j=1}^{N} \text{Gamma}(\eta_j \mid a_j, b_j) \\ &= \int_{\eta_i} \eta_i \, \text{Gamma}(\eta_i \mid a_i, b_i) \\ &= a_i / b_i \\ &= \frac{\nu_0 + 1}{\nu_0 + \lambda_0 (x_i - \mu_0)^2} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\log \eta_i] &= \int_Z \log \eta_i \prod_{j=1}^{N} \text{Gamma}(\eta_j \mid a_j, b_j) \\ &= \int_{\eta_i} \log \eta_i \, \text{Gamma}(\eta_i \mid a_i, b_i) \\ &= \psi(a_i) - \log b_i \\ &= \psi\left(\frac{\nu_0 + 1}{2}\right) - \log \frac{\nu_0 + \lambda_0 (x_i - \mu_0)^2}{2} \end{aligned}$$

See Appendix A for the definition of the digamma function, $\psi(\cdot)$.

[2] See Appendix A for the algebraic expansions of the Normal and Gamma distributions.

Step 4: Function to optimize

$$\begin{aligned} Q(\Theta, \Theta_0) &= \int_Z P\left(Z \mid X, \Theta_0\right) \log P\left(X, Z \mid \Theta\right) \\ &= -\frac{N}{2} \log 2\pi + \frac{N}{2} \log \lambda + \frac{1}{2} \sum_{i=1}^{N} \mathbb{E}[\log \eta_i] - \frac{\lambda}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \mathbb{E}[\eta_i] \\ &\quad -N \log \Gamma\left(\frac{\nu}{2}\right) + \frac{N\nu}{2} \log \frac{\nu}{2} + \left(\frac{\nu}{2} - 1\right) \sum_{i=1}^{N} \mathbb{E}[\log \eta_i] - \frac{\nu}{2} \sum_{i=1}^{N} \mathbb{E}[\eta_i] \end{aligned}$$

Note that all the elements of Θ_0 are now implicit in the expectations.

Step 5: Maximization

$$\begin{aligned} \frac{\partial Q}{\partial \mu} = 0 \quad &\Rightarrow \quad \lambda \sum_{i=1}^{N} (x_i - \mu) \mathbb{E}[\eta_i] = 0 \\ &\Rightarrow \quad \mu = \frac{\sum_{i=1}^{N} x_i \mathbb{E}[\eta_i]}{\sum_{i=1}^{N} \mathbb{E}[\eta_i]} \end{aligned}$$

$$\begin{aligned} \frac{\partial Q}{\partial \lambda} = 0 \quad &\Rightarrow \quad \frac{N}{2\lambda} - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 \mathbb{E}[\eta_i] = 0 \\ &\Rightarrow \quad \lambda = \left(\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \mathbb{E}[\eta_i] \right)^{-1} \end{aligned}$$

Note that we require the updated μ value to find λ.

$$\begin{aligned} \frac{\partial Q}{\partial \nu} = 0 \quad &\Rightarrow \quad -\frac{N}{2} \psi\left(\frac{\nu}{2}\right) + \frac{N}{2} \log \frac{\nu}{2} + \frac{N}{2} + \frac{1}{2} \sum_{i=1}^{N} \mathbb{E}[\log \eta_i] - \frac{1}{2} \sum_{i=1}^{N} \mathbb{E}[\eta_i] = 0 \\ &\Rightarrow \quad \psi\left(\frac{\nu}{2}\right) - \log \frac{\nu}{2} = 1 + \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\log \eta_i] - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\eta_i] \end{aligned}$$

Note that there is no closed form solution for ν—it has to be found numerically.

4 Multivariate Case

The derivation of the multivariate case requires taking vector and matrix derivatives to find the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ but is otherwise very similar to the univariate case. The update equations turn out to be[3]

$$\mathbb{E}[\eta_i] = \frac{\nu_0 + D}{\nu_0 + (\boldsymbol{x}_i - \boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0 (\boldsymbol{x}_i - \boldsymbol{\mu}_0)}$$

$$\mathbb{E}[\log \eta_i] = \psi\left(\frac{\nu_0 + D}{2}\right) - \log \frac{\nu_0 + (\boldsymbol{x}_i - \boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0 (\boldsymbol{x}_i - \boldsymbol{\mu}_0)}{2}$$

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i \mathbb{E}[\eta_i]}{\sum_{i=1}^{N} \mathbb{E}[\eta_i]}$$

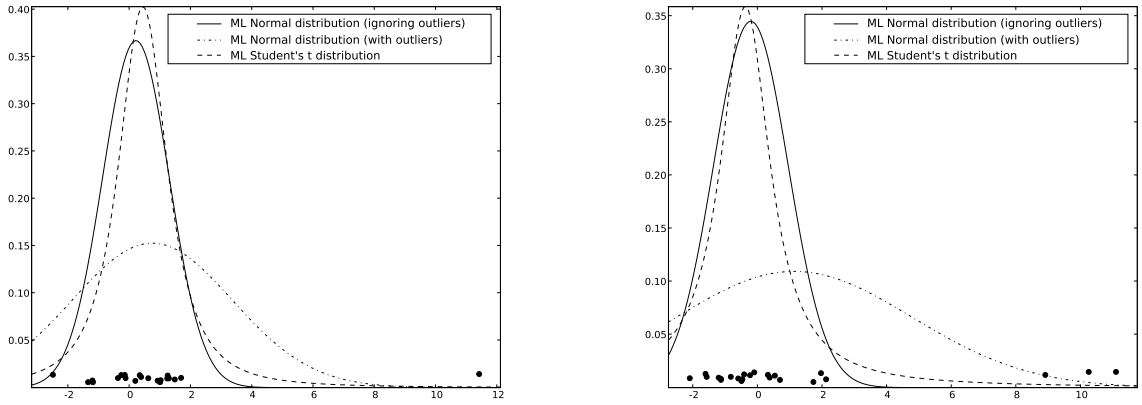
$$\boldsymbol{\Lambda} = \left(\frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T \mathbb{E}[\eta_i] \right)^{-1}$$

The update equation for ν remains unchanged.

[3] D is still the dimensionality.

5 Demonstration

The update equations for the univariate case were implemented in Python.[4] Figure 1 shows the results. Notice how the ML fit to the Normal distribution has to increase its variance (flatten out) more and more to accommodate the increasing number of outliers. The Student's t distribution is relatively robust to this.



1 outlier 3 outliers

Figure 1: Plots of data points drawn from a Normal distribution along with a small number of outliers. The ML fit of a Normal (with and without considering the outliers) and a Student's t to the data are also shown.

A Distributions and Relations

$$
\begin{aligned}
\text{Normal}\left(x \,\middle|\, \mu, \sigma^2\right) &= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\
\text{Gamma}\left(x \,|\, a, b\right) &= \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx} \\
\mathbb{E}_{\text{Gamma}}[x] &= a/b \\
\mathbb{E}_{\text{Gamma}}[\log x] &= \psi(a) - \log b \\
\psi(x) &= \frac{\Gamma'(x)}{\Gamma(x)}
\end{aligned}
$$

References

[1] C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, NY, USA, 2006.

[2] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley, New York, NY, USA, 1997.

[4] If you have a look at the code, you'll probably notice that the update equations are a bit different from those in the text. This is simply due to optimization—they are mathematically equivalent.