The Calibration of t Distributions Using the EM Algorithm

Chun-Yuan Chiu

The EM algorithm can be thought of as an extension of the maximum likelihood method, so it is worth reviewing the maximum likelihood method first.

1 MLE for Normal Distributions

Suppose we are given a set of sample points $x_1, x_2, \ldots, x_n$ to calibrate a normal distribution with the probability density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Denote by θ the set of parameters $\theta \equiv (\mu, \sigma)$. The likelihood of θ given sample x_i is $f_X(x_i|\theta)$. So the likelihood product is $\prod_{i=1}^n f(x_i|\theta)$. The maximum likelihood estimator for the parameter θ is

$$\begin{aligned} \theta &= \operatorname*{argmax}_\theta \prod_{i=1}^n f(x_i;\theta) \\ &= \operatorname*{argmax}_\theta \sum_{i=1}^n \log f(x_i;\theta). \end{aligned}$$

We want to find the maximum of the log-likelihood

$$\begin{aligned} L(\mu,\sigma) &\equiv \sum_{i=1}^n \log f(x_i;\theta) \\ &= \sum_{i=1}^n \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) \\ &= \sum_{i=1}^n \left(-\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right). \end{aligned}$$

Setting the partial derivative w.r.t. μ to zero gives

$$\frac{\partial L}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^n \frac{\partial}{\partial \mu}\left(\frac{x_i-\mu}{\sigma}\right)^2 = 0,$$

which simplifies to

$$0 = -\frac{1}{2}\sum_{i=1}^n \frac{\partial}{\partial \mu}(x_i-\mu)^2 = \sum_{i=1}^n (x_i-\mu),$$

or equivalently $\sum_{i=1}^n x_i = n\mu$, or

$$\mu = \frac{1}{n}\sum_{i=1}^n x_i. \tag{1}$$

This is the maximum likelihood estimator of μ.

Similarly, set the partial derivative w.r.t. σ to zero to yield

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^n \frac{\partial}{\partial \sigma}\left(-\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right) = 0,$$

which simplifies to

$$\begin{aligned} 0 &= \sum_{i=1}^n \left(-\frac{\partial}{\partial\sigma}\log\sigma - \frac{(x_i-\mu)^2}{2}\frac{\partial}{\partial\sigma}\left(\frac{1}{\sigma^2}\right)\right) \\ &= \sum_{i=1}^n \left(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}\right). \end{aligned}$$

Multiplying σ^3 to both sides, we get

$$0 = \sum_{i=1}^n \left(-\sigma^2 + (x_i-\mu)^2\right),$$

or $n\sigma^2 = \sum_{i=1}^n (x_i-\mu)^2$. So the maximum likelihood estimator of σ is

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2},$$

where μ is given by the maximum likelihood estimator (1).

2 The Calibration of t Distributions

The normal variance-mean mixture representation of a t distribution with ν degrees of freedom is $X = \mu + \sqrt{W}Z$, where $Z \sim N(0,\sigma^2)$, and W is the mixing distribution following an inverse gamma distribution $IG(\nu/2, \nu/2)$. To calibrate the distribution, let us first derive the maximum likelihood estimators as if all the observations $\{x_i\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$ are available.

The joint probability density function of X and W is

$$f_{X,W} = f_{X|W}(x|W;\mu,\sigma) f_W(w;\nu),$$

so the log-likelihood is

$$\begin{aligned} L(\mu,\sigma,\nu) &\equiv \log\left(\prod_{i=1}^n f_{X,W}(x_i,w_i;\mu,\sigma,\nu)\right) \\ &= \sum_{i=1}^n \left(\log f_{X|W}(x_i|w_i;\mu,\sigma) + \log f_W(w_i;\nu)\right). \end{aligned}$$

Plugging in the density functions

$$\begin{aligned} f_{X|W}(x_i|W=w_i;\mu,\sigma) &= \frac{1}{\sigma\sqrt{2\pi w_i}} e^{-\frac{1}{2w_i}\left(\frac{x_i-\mu}{\sigma}\right)^2}, \\ f_W(w_i;\nu) &= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} w_i^{-\frac{\nu}{2}-1} e^{-\frac{\nu}{2w_i}}, \end{aligned}$$

the log-likelihood function can be rewritten as

$$\begin{aligned} L(\mu,\sigma,\nu) = \sum_{i=1}^{n} \Bigg(& -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\log w_i - \frac{1}{2w_i}\left(\frac{x_i-\mu}{\sigma}\right)^2 \\ & - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log\left(\frac{\nu}{2}\right) - \left(\frac{\nu}{2}+1\right)\log w_i - \frac{\nu}{2w_i}\Bigg). \end{aligned}$$

Setting all the partial derivatives $\partial L/\partial\mu, \partial L/\partial\sigma, \partial L/\partial\nu$ to zero, we get

$$\begin{aligned} \frac{\partial L}{\partial \mu} &= -\frac{1}{2}\sum_{i=1}^{n} \frac{1}{w_i}\frac{\partial}{\partial\mu}\left(\frac{x_i-\mu}{\sigma}\right)^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{x_i-\mu}{w_i} = 0, \\ \frac{\partial L}{\partial \sigma} &= \sum_{i=1}^{n}\frac{\partial}{\partial\sigma}\left(-\log(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2w_i}\frac{1}{\sigma^2}\right) \\ &= \sum_{i=1}^{n}\left(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{w_i}\frac{1}{\sigma^3}\right) \\ &= -\frac{1}{\sigma^3}\sum_{i=1}^{n}\left(\sigma^2 - \frac{(x_i-\mu)^2}{w_i}\right) = 0, \\ \frac{\partial L}{\partial \nu} &= \sum_{i=1}^{n}\frac{\partial}{\partial\nu}\left(-\log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\log w_i - \frac{\nu}{2w_i}\right) \\ &= \frac{1}{2}\sum_{i=1}^{n}\left(-\psi_0\left(\frac{\nu}{2}\right) + 1 + \log\left(\frac{\nu}{2}\right) - \log w_i - \frac{1}{w_i}\right) = 0, \end{aligned}$$

where $\psi_0(x) = d\log\Gamma(x)/dx$ is the digamma function. Thus, the maximum likelihood estimators of μ and σ are respectively

$$\mu = \frac{\sum_{i=1}^{n} x_i \frac{1}{w_i}}{\sum_{i=1}^{n}\frac{1}{w_i}}, \tag{2}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{w_i}(x_i-\mu)^2\right)}, \tag{3}$$

and the estimator of ν is the solution of

$$-\psi_0\left(\frac{\nu}{2}\right) + 1 + \log\left(\frac{\nu}{2}\right) - \frac{1}{n}\sum_{i=1}^{n}\log w_i - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_i} = 0. \tag{4}$$

If all the observations $\{x_i\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$ were available, we can use the above formulae and the calibration is done. However, in practice only $\{x_i\}_{i=1}^n$ can be observed, so we still need other estimators for $\{w_i\}_{i=1}^n$. The best guess we have about w_i is the conditional expectation $E[W|X = x_i]$. Thus we replace all the $g(w_i)$ by $E[g(W)|X = x_i]$ in Eqs. (5), (6), (7) to get the new estimators

$$\mu = \frac{\sum_{i=1}^n x_i E\left[\frac{1}{W}\middle| X = x_i\right]}{\sum_{i=1}^n E\left[\frac{1}{W}\middle| X = x_i\right]}, \tag{5}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^n \left((x_i - \mu)^2 E\left[\frac{1}{W}\middle| X = x_i\right]\right)}, \tag{6}$$

and ν the solution of

$$-\psi_0\left(\frac{\nu}{2}\right) + 1 + \log\left(\frac{\nu}{2}\right) - \frac{1}{n}\sum_{i=1}^n E\left[\log W|X = x_i\right] - \frac{1}{n}\sum_{i=1}^n E\left[\frac{1}{W}\middle| X = x_i\right] = 0. \tag{7}$$

This leads to the EM algorithm:

1. Get an initial guess $\mu^{[0]}, \sigma^{[0]}$ and $\nu^{[0]}$
2. Apply (5), (6) and (7) to evaluate $\mu^{[k+1]}, \sigma^{[k+1]}$ and $\nu^{[k+1]}$, where all the expectations $E[g(W)|X]$ are evaluated with parameters $\mu = \mu^{[k]}, \sigma = \sigma^{[k]}, \nu = \nu^{[k]}$
3. Repeat step 2 until convergence

In the common understanding of the EM algorithm, there are E step and M step. In E step, one evaluates the conditional expectation

$$\begin{aligned}
&E[L(X_1, \ldots, X_n, W_1, \ldots, W_n; \mu, \sigma, \nu)|X_1 = x_1, \ldots, X_n = x_n] \\
&= E\left[\sum_{i=1}^n \left(-\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\log W_i - \frac{1}{2W_i}\left(\frac{X_i - \mu}{\sigma}\right)^2\right.\right. \\
&\qquad \left.\left. - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\log W_i - \frac{\nu}{2W_i}\right)\middle| X_1 = x_1, \ldots, X_n = x_n\right].
\end{aligned}$$

The result will be a function of μ, σ and ν. Then one finds the parameters that maximize this conditional expectation in the M step. When evaluating the conditional expectation in the E step, everything can be pulled out of the expectation and finally the whole expectation can be rewritten in terms of $E[\log W_i|X_1, \ldots, X_n]$ and $E[1/W_i|X_1, \ldots, X_n]$, which we evaluate with the parameters we get from the last iteration. The resulting estimators and algorithm are the same as aforementioned.

A Closed Form Formulae of $E[g(W)|X = x_i]$

Note that the joint density function of X and W is

$$f_{X,W}(x, w) = f_{W|X}(w|X)f_X(x) = f_{X|W}(x|W)f_W(w).$$

Thus we know that

$$f_{W|X=x_i}(w|X=x_i) = f_W(w)\frac{f_{X|W}(x_i|W=w)}{f_X(x_i)}.$$

Plugging in the density functions

$$\begin{aligned} f_{X|W}(x_i|W=w) &= \frac{1}{\sigma\sqrt{2\pi w}} e^{-\frac{1}{2w}\left(\frac{x_i-\mu}{\sigma}\right)^2}, \\ f_W(w) &= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} w^{-\frac{\nu}{2}-1} e^{-\frac{\nu}{2w}}, \\ f_X(x_i) &= \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \end{aligned}$$

we get

$$\begin{aligned} f_{W|X=x_i}(w|X=x_i) &= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} w^{-\frac{\nu}{2}-1} e^{-\frac{\nu}{2w}} \frac{\frac{1}{\sigma\sqrt{2\pi w}} e^{-\frac{1}{2w}\left(\frac{x_i-\mu}{\sigma}\right)^2}}{\frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}} \\ &= \frac{1}{\Gamma\left(\frac{\nu+1}{2}\right)}\left(\frac{\nu}{2}\right)^{\frac{\nu+1}{2}} w^{-\frac{\nu}{2}-\frac{3}{2}} e^{-\frac{\nu}{2w}-\frac{1}{2w}\left(\frac{x_i-\mu}{\sigma}\right)^2}\left(1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right)^{\frac{\nu+1}{2}} \\ &= \frac{1}{\Gamma\left(\frac{\nu+1}{2}\right)}\left(\frac{\nu}{2}+\frac{(x_i-\mu)^2}{2\sigma^2}\right)^{\frac{\nu+1}{2}} w^{-\frac{\nu}{2}-\frac{3}{2}} e^{-\frac{1}{w}\left(\frac{\nu}{2}+\frac{(x_i-\mu)^2}{2\sigma^2}\right)}, \end{aligned}$$

equivalent to the density function of an inverse gamma distribution $IG(\alpha,\beta)$ with

$$\alpha = \frac{\nu+1}{2}, \qquad \beta = \frac{\nu}{2}+\frac{(x_i-\mu)^2}{2\sigma^2}.$$

Thus we introduce a new random variable $Y \sim IG(\alpha,\beta)$, and $E[g(W)|X=x_i] = E[g(Y)]$. Recall that a random variable Y follows $IG(\alpha,\beta)$ if and only if its reciprocal $Z=1/Y$ follows a gamma distribution $\Gamma(\alpha,\beta)$. Thus

$$\begin{aligned} E\left[\frac{1}{W}\middle|X=x_i\right] &= E\left[\frac{1}{Y}\right] = E[Z] = \frac{\alpha}{\beta}, \\ E\left[\log W|X=x_i\right] &= E\left[\log Y\right] = -E[\log Z] = \log\beta - \psi_0(\alpha), \end{aligned}$$

where $\psi_0(x) = d\log\Gamma(x)/dx$ is the digamma function.