Expectation Maximization [2] for Fitting T-distributions [4]

Qi Zhao

August 19, 2018

1. Introduction

Since the pdf takes the form of a marginalization of the joint distribution [3] with a hidden variable, it can use the EM algorithm to learn the parameters $\theta = \{\mu, \Sigma, \nu\}$ from a set of training data $\{x_i\}_{i=1}^{I}$.

In the E-step (Figure 1) it maximizes the bound with respect to the distributions $q_i\ (h_i)$ by finding the posterior $\Pr(h_i|x_i, \theta^{[t]})$ over each hidden variable h_i given associated observation x_i and the current parameter settings. By Bayes rule [1], it gets in Equation 1.

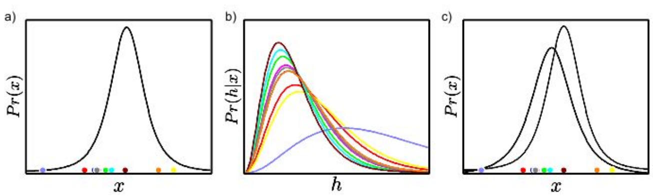

Figure 1. Expectation maximization for fitting t-distributions. a) Estimate of distribution before update. b) In the E-step we calculate the posterior distribution $\Pr(h_i\ |x_i\ )$ over the hidden variable h_i for each data point x_i. The color of each curve corresponds to that of the original data point in (a). c) In the M-step we use these distributions over h to update the estimate of the parameters $\theta = \mu$, σ^2, ν.

$$
\begin{aligned}
q_i(h_i) &= Pr(h_i|x_i, \theta^{[t]}) \\
&= \frac{Pr(x_i|h_i, \theta^{[t]})Pr(h_i)}{Pr(x_i|\theta^{[t]})} \\
&= \frac{Norm_{x_i}[\mu, \Sigma/h_i]Gam_{h_i}[\nu/2, \nu/2]}{Pr(x_i)} \\
&= Gam_{h_i}[\frac{\nu + D}{2}, \frac{(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}{2} + \frac{\nu}{2}]
\end{aligned} \tag{1}
$$

where it has used the fact that the gamma distribution is conjugate to the scaling factor for the normal variance. The E-step can be understood as follows: it is treating each data point x_i as if it were generated from one of the normals in the infinite mixture where the hidden variable h_i determines which normal. So, the E-step computes a distribution over h_i, which hence determines a distribution over which normal created the data.

2. Conclusions

In conclusion, the multivariate t-distribution provides an improved description of data with outliers (Figure 2). It has just one more parameter than the normal (the degrees of freedom, ν), and subsumes the normal as a special case (where ν becomes very large). However, this generality comes at a cost: there is no closed form solution for the maximum likelihood parameters and so it must resort to more complex approaches such as the EM algorithm [**?**] to fit the distribution.

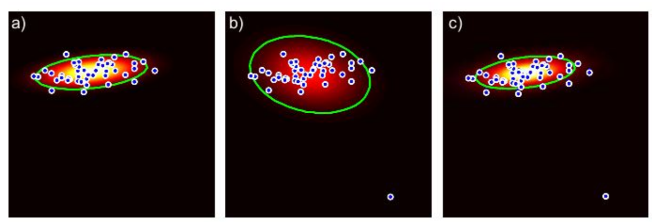

Figure 2. Motivation for t-distribution. a) The multivariate normal model fit to data. b) Adding a single outlier completely changes the fit. c) With the multivariate t-distribution the outlier does not have such a drastic effect.

References

[1] T. J. Anastasio, P. E. Patton, and K. Belkacem-Boussaid. Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12(5):1165–1187, 2000. 1

[2] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1-2):51–80, 1995. 1

[3] M. Longuet-Higgins. On the joint distribution of the periods and amplitudes of sea waves. *Journal of Geophysical Research*, 80(18):2688–2694, 1975. 1

[4] S. Shoham, M. R. Fellows, and R. A. Normann. Robust, automatic spike sorting using mixtures of multivariate t-distributions. *Journal of Neuroscience Methods*, 127(2):111–122, 2003. 1