Estimating Parametric Continuous Latent Distributions with the EM Algorithm

Brad Hanson

October 1, 1996

The observed data are $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, the missing data are $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$, and the complete data are $[(\mathbf{y}_1, \boldsymbol{\theta}_1), \ldots, (\mathbf{y}_N, \boldsymbol{\theta}_N)]$, where $\mathbf{y}_i$ is the vector of item responses and $\boldsymbol{\theta}_i$ is the vector of unobserved latent variables (possibly multivariate) for randomly sampled examinee i. Let the item parameters for item j be given by $\boldsymbol{\delta}_j$, and the collection of item parameters for all K items by $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_j, \ldots, \boldsymbol{\delta}_K]$. It is assumed the item parameters for all items are known. The goal is to estimate the parameters $(\boldsymbol{\pi})$ of the distribution of the latent variables in the population of examinees $[g(\boldsymbol{\theta} \mid \boldsymbol{\pi})]$ assuming the the item parameters $(\boldsymbol{\Delta})$ are known. Let $s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$, $k = 1, \ldots, m$, be sufficient statistics for the parameters $\boldsymbol{\pi}$.

The EM algorithm consists of the E step in which the expected values of the missing sufficient statistics $[s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)]$ over the distribution of the missing data given the observed data and provisional values of the parameters $\boldsymbol{\pi}$ are computed, and the M step in which the expected values of the sufficient statistics computed in the E step are used to compute complete data maximum likelihood estimates of the parameters (Dempster, Laird, and Rubin, 1977). For computing estimates of $\boldsymbol{\pi}$ the E step at iteration $s = 0, 1, \ldots$ consists of computing the k quantities

$$E_{\boldsymbol{\Theta}}[s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \mid \mathbf{Y}, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}]\,, \tag{1}$$

where $\boldsymbol{\Theta}$ is the vector of latent random variables for the N examinees (these random variables are independent and identically distributed), and the expectation is over the conditional distribution of these random variables given the observed data and fixed known values of the parameters ($\boldsymbol{\pi}^{(s)}$ and $\boldsymbol{\Delta}$). Equation 1 can be written as

$$\begin{aligned}
&E_{\boldsymbol{\Theta}}[s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \mid \mathbf{Y}, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}] \\
&\quad = \int_{\boldsymbol{\theta}_1} \cdots \int_{\boldsymbol{\theta}_N} s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) p[(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \mid \mathbf{Y}, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}] d\boldsymbol{\theta}_1 \ldots d\boldsymbol{\theta}_N \\
&\quad = \int_{\boldsymbol{\theta}_1} \cdots \int_{\boldsymbol{\theta}_N} s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) p_1[\boldsymbol{\theta}_1 \mid \mathbf{y}_1, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}] \ldots p_N[\boldsymbol{\theta}_N \mid \mathbf{y}_N, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}]\, d\boldsymbol{\theta}_1 \ldots d\boldsymbol{\theta}_N\,. \quad (2)
\end{aligned}$$

Going from the second to the third line of Equation 2 follows from the fact that examinees are independently sampled so that the latent random variables for the individual examinees are mutually independent. Since $(\mathbf{y}_i, \boldsymbol{\theta}_i)$ for $i = 1, \ldots N$ are identically distributed

$$p_i[\boldsymbol{\theta}_i \mid \mathbf{y}_i, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}] = p[\boldsymbol{\theta}_i \mid \mathbf{y}_i, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}]\,. \tag{3}$$

If $s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ can be written as

$$s_k(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) = \sum_{i=1}^{N} t_k(\boldsymbol{\theta}_i)\,, \tag{4}$$

then Equation 2 can be written as

$$\sum_{i=1}^{N} \int_{\boldsymbol{\theta}} t_k(\boldsymbol{\theta}) p[\boldsymbol{\theta} \mid \mathbf{y}_i, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}] d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} t_k(\boldsymbol{\theta}) \left\{ \sum_{i=1}^{N} p[\boldsymbol{\theta} \mid \mathbf{y}_i, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}] \right\} d\boldsymbol{\theta} \tag{5}$$

Using Bayes Theorem, the distribution $p[\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}]$ is given by

$$p[\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}] = \frac{f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\Delta}) g(\boldsymbol{\theta} \mid \boldsymbol{\pi}^{(s)})}{\int_{\boldsymbol{\theta}'} f(\mathbf{y} \mid \boldsymbol{\theta}', \boldsymbol{\Delta}) g(\boldsymbol{\theta}' \mid \boldsymbol{\pi}^{(s)}) d\boldsymbol{\theta}'}\,. \tag{6}$$

If it is assumed that the vector of latent random variables is multivariate normal then the sufficient statistics are the sum of each latent variable, the sum of squares of each latent variable, and the sum of cross products of pairs of the latent variables. Consequently, the sufficient statistics can be expressed in the form of Equation 4, and Equation 5 can be used for the E-step calculations of the sufficient statistics. The maximum likelihood estimates of the parameters of the multivariate normal distribution are simple functions of the sufficient statistics, so the M-step consists of simply calculating maximum likelihood parameter estimates from the sufficient statistics calculated in the E-step. The solution given by Mislevy (1984) when the latent distribution is multivariate normal is the same as that given by Equation 5.

References

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, *39*, 1-38.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.