Applying EM to Probit Regression

In this model, our observed data hi$y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$ is a vector of our n data points (0's and 1's). Each y_i is associated with a scalar covariate x_i, from which we construct a design matrix $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix}$.

$\theta = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_n \end{bmatrix}$ is unobserved, and will be thought of as our "missing data" in this problem (so this θ plays the role of the θ on the first handout). There exists some vector $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ such that $\theta_i = X_i\beta + \epsilon_i$ for $i = 1, \dots, n$, where $X_i = \begin{bmatrix} 1 & x_i \end{bmatrix}$ is the ith row of X, and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$. β is the "unknown parameter" in this model, and is identified with ϕ on the first handout. Given our data, y, we have a posterior distribution $f_{\beta \mid y}(\beta \mid y)$ over β, and we want to find the value $\widehat{\beta}$ of β at which this density is highest. That's what the EM algorithm does.

Assume we've chosen some initial value $\beta^{(0)}$ for β, say $\beta^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. For $t = 0$ to N—where N is our number of iterations—we apply the following steps. At the end, $\beta^{(N)}$ will be our approximation to $\widehat{\beta}$.

1. E Step: Compute $Q(\beta \mid \beta^{(t)})$.

Since we only do this computation so that we can choose β to maximize $Q(\beta \mid \beta^{(t)})$ in the M step, it suffices to find only those parts of $Q(\beta \mid \beta^{(t)})$ that depend on β—the "sufficient statistics" of $Q(\beta \mid \beta^{(t)})$.

I'll use the notation $\mathbb{E}_X[g(X)]$ to refer to the expectation of some function $g(X)$ with respect to the random variable X, i.e.,

$$\mathbb{E}_X[g(X)] := \int g(x) f_X(x)\, dx.$$

By definition,

$$Q(\beta \mid \beta^{(t)}) = \mathbb{E}_{\theta \mid \beta^{(t)},\, y}\left[\ln f_{\theta,\, \beta \mid y}(\theta,\, \beta \mid y)\right]. \tag{1}$$

Since $f_{\theta,\, \beta \mid y}(\theta,\, \beta \mid y) = f_{\theta \mid y}(\theta \mid y) f_{\beta \mid \theta,\, y}(\beta \mid \theta,\, y)$, (1) becomes

$$Q(\beta \mid \beta^{(t)}) = \mathbb{E}_{\theta \mid \beta^{(t)},\, y}\left[\ln f_{\theta \mid y}(\theta \mid y)\right] + \mathbb{E}_{\theta \mid \beta^{(t)},\, y}\left[\ln f_{\beta \mid \theta,\, y}(\beta \mid \theta,\, y)\right],$$

where the first term on the right doesn't involve β. Thus, it suffices to maximize

$$\mathbb{E}_{\theta \mid \beta^{(t)},\, y}\left[\ln f_{\beta \mid \theta,\, y}(\beta \mid \theta,\, y)\right]. \tag{2}$$

Note that $f_{\beta \mid \theta,\, y}(\beta \mid \theta,\, y) = f_{\beta \mid \theta}(\beta \mid \theta)$ because for each i, if we know θ_i, we know its sign, and hence we automatically know y_i. Moreover, $f_{\beta \mid \theta}(\beta \mid \theta) \propto f_{\theta \mid \beta}(\theta \mid \beta) f_\beta(\beta)$, and taking a uniform prior $f_\beta(\beta) \propto$ const, we have $f_{\beta\,|\,\theta}(\beta\,|\,\theta) \propto f_{\theta\,|\,\beta}(\theta\,|\,\beta)$. Thus (2) becomes

$$\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[\ln\,(\text{const})\right] + \mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[\ln f_{\theta\,|\,\beta}(\theta\,|\,\beta)\right]. \tag{3}$$

Our model specifies that $\theta \sim N_n(X\beta, \mathbf{I})$, so

$$\ln f_{\theta\,|\,\beta}(\theta\,|\,\beta) \propto -\frac{1}{2}(\theta - X\beta)'(\theta - X\beta),$$

and maximizing (3) is equivalent to minimizing an "expected sum of squares":

$$\begin{aligned}\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[(\theta - X\beta)'(\theta - X\beta)\right] &= \mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[\theta'\theta\right] - 2\,\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[\beta' X'\theta\right] + \mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[\beta' X' X\beta\right] \\ &= \text{const} - 2\beta'\ \mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[X'\theta\right] + \beta' X' X\beta.\end{aligned} \tag{4}$$

2. M Step: Set $\beta^{(t+1)} := \underset{\beta}{\operatorname{argmax}}\ Q(\beta\,|\,\beta^{(t)})$.

Setting the derivative of (4) with respect to β to 0:[1]

$$\begin{aligned}-2\left(\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[X'\theta\right]\right)' + 2\beta' X' X &= \mathbf{0} \\ \left(\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[X'\theta\right]\right)' &= \beta' X' X \\ \mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[X'\theta\right] &= X' X\beta \\ (X'X)^{-1}\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[X'\theta\right] &=: \beta^{(t+1)}.\end{aligned}$$

Note that $\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[X'\theta\right]$, a 2×1 vector, is the minimal sufficient statistic. However, in practice it's easiest to find just $\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[\theta\right]$, an $n \times 1$ vector and use the fact that

$$\mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[X'\theta\right] = X'\ \mathbb{E}_{\theta\,|\,\beta^{(t)},\,y}\left[\theta\right].$$

[1] I'm using the following two rules of matrix calculus: For $\boldsymbol{a}, \boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m \times m}$, $\frac{\partial\, \boldsymbol{x}'\boldsymbol{a}}{\partial\, \boldsymbol{x}} = \boldsymbol{a}'$, and $\frac{\partial\, \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial\, \boldsymbol{x}} = \boldsymbol{x}'\left(\boldsymbol{A}' + \boldsymbol{A}\right)$.