Applying EM to Probit Regression

In this model, our observed data \( y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \) 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 \( X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \).

\( \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \) is unobserved, and will be thought of as our “missing data” in this problem (so this \( \theta \) plays the role of the \( \theta \) 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 \( i = 1, \ldots, n \), where \( X_i = \begin{bmatrix} 1 & x_i \end{bmatrix} \) is the \( i \)th row of \( X \), and \( \epsilon_i \sim N(0,1) \). \( \beta \) is the “unknown parameter” in this model, and is identified with \( \phi \) on the first handout. Given our data, \( y \), we have a posterior distribution \( f_{\beta|y}(\beta|y) \) over \( \beta \), and we want to find the value \( \hat{\beta} \) of \( \beta \) at which this density is highest. That’s what the EM algorithm does.

Assume we’ve chosen some initial value \( \beta^{(0)} \) for \( \beta \), 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 \( \hat{\beta} \).

1. **E Step:** Compute \( Q(\beta | \beta^{(t)}) \).

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

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,

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

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

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

where the first term on the right doesn’t involve \( \beta \). Thus, it suffices to maximize

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

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

$$
E_{\theta|\beta(t), \gamma}[\ln \text{const}] + E_{\theta|\beta(t), \gamma}[\ln f_{\theta|\beta}(\theta|\beta)].
$$

(3)

Our model specifies that $\theta \sim N_\theta(X\beta, \mathbf{I})$, 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”:

$$
E_{\theta|\beta(t), \gamma}[(\theta - X\beta)'(\theta - X\beta)] &= E_{\theta|\beta(t), \gamma}[(\theta - X\beta)'(\theta - X\beta)]
\quad = \text{const} - 2\beta'(\beta'X'X) + \beta'X'\beta.
$$

(4)

2. **M Step**: Set $\beta^{(t+1)} := \arg\max_\beta Q(\beta|\beta(t))$.

Setting the derivative of (4) with respect to $\beta$ to 0:

$$
-2(E_{\theta|\beta(t), \gamma}[X'\theta])' + 2\beta'X'X = 0
\quad \Rightarrow (E_{\theta|\beta(t), \gamma}[X'\theta])' = \beta'X'X
\quad \Rightarrow E_{\theta|\beta(t), \gamma}[X'\theta] = X'X\beta
\quad \Rightarrow (X'X)^{-1}E_{\theta|\beta(t), \gamma}[X'\theta] =: \beta^{(t+1)}.
$$

Note that $E_{\theta|\beta(t), \gamma}[X'\theta]$, a $2 \times 1$ vector, is the minimal sufficient statistic. However, in practice it’s easiest to find just $E_{\theta|\beta(t), \gamma}[\theta]$, an $n \times 1$ vector and use the fact that

$$
E_{\theta|\beta(t), \gamma}[X'\theta] = X' E_{\theta|\beta(t), \gamma}[\theta].
$$

---

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

---

2