EM Algorithm

Applied to Probit Regression

Gabriel E. Cabrera-Guzmán

The University of Manchester
Alliance Manchester Business School

October 3, 2023

Introduction

1. $y_i = \begin{bmatrix} y_i \\ ... \\ y_n \end{bmatrix}$ is a vector of n data points (0's and 1's).
2. Each y is associated with a scalar covariates x_i, from which we construct a design matrix:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ ... & ... \\ 1 & x_n \end{bmatrix}$$

3. $\theta = \begin{bmatrix} \theta_1 \\ ... \\ \theta_n \end{bmatrix}$ is unobserved $\rightarrow$ "missing data".

Introduction (Cont'd)

4. Exists some vector $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ such that:

$$\theta_i = X_i \beta + \epsilon_i \quad \text{for} \quad i = 1, ..., n.$$

   Where $X_i = [1 \ x_i]$ is the ith row of X, and $\epsilon_i \sim N(0, 1)$

5. Given y, we have a posterior distribution $f_{\beta|y}(\beta|y)$ over β. Then, we need to:

   - Find the value $\hat{\beta}$ of β at which this density is highest
   - Assume initial value $\beta^{(0)}$ for β, say $\beta^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. For $t = 0$ to N (iterations) we apply the E Step and M Step

E Step: Compute $Q(\beta|\beta^{(t)})$

It suffices to find only those parts of $Q(\beta|\beta^{(t)})$ that depend on β:

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

by definition:

$$\begin{aligned} Q(\beta|\beta^{(t)}) &= \mathbb{E}_{\theta|\beta^{(t)},y}[\ln f_{\theta,\beta|y}(\theta,\beta|y)] \\ &= \mathbb{E}_{\theta|\beta^{(t)},y}[\ln f_{\theta|y}(\theta|y)] + \mathbb{E}_{\theta|\beta^{(t)},y}[\ln f_{\beta|\theta,y}(\beta|\theta,y)] \\ &= \mathbb{E}_{\theta|\beta^{(t)},y}[\ln f_{\beta|\theta,y}(\beta|\theta,y)]. \end{aligned}$$

Why?

E Step: Compute $Q(\beta|\beta^{(t)})$ (Cont'd)

$$
\begin{aligned}
f_{\beta|\theta,y}(\beta|\theta,y) &= f_{\theta|y}(\theta|y) f_{\beta|\theta,y}(\beta|\theta,y) \\
&= f_{\beta|\theta,y}(\beta|\theta,y) \\
&= \ \ldots \\
&= f_{\beta|\theta}(\beta|\theta) \propto f_{\theta|\beta}(\theta|\beta) f_{\beta}(\beta)
\end{aligned}
$$

Taking a uniform prior $f_\beta(\beta) \propto$ const:

$$
f_{\beta|\theta}(\beta|\theta) \propto f_{\theta|\beta}(\theta|\beta)
$$

Therefore becomes:

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

E Step: Compute $Q(\beta|\beta^{(t)})$ (Cont'd)

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

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

Maximizing is equivalent to minimizing an "expected sum of squares":

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

Where $(\cdot) = \theta|\beta^{(t)}, y$ to save some space.

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:

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

Two rules of matrix calculus

For α, $x \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times m}$:

$$\frac{\partial x'}{\partial x} = \alpha' \quad \text{and} \quad \frac{\partial x' A x}{\partial x} = x'(\mathbf{A}' + \mathbf{A})$$

EM Algorithm From Scratch