EM Algorithm for Multinomial Logit[*]

Soichiro Yamauchi[†]

March 14, 2020

1 Motivation

The multinomial regression is a useful model for analyzing relationships between categorical outcomes and predictors, and is routinely used in many fields. The model is also useful as a building block for more complex models. In fact, we encounter the multinomial logit every time we want to incorporate covariates in latent variable models (i.e., finite mixture models) such as stochastic blockmodels for network data or topic models for text data analysis.

In this note, I derive an EM algorithm for the standard multinomial logistic regression model as a useful reference. An open-source software package (emlogit) is available for implementing the proposed algorithm.

2 Model

Let $\mathbf{Y}_i$ denote the multinomial response with J categories. As in the standard setup, I assume that $\sum_{j=1}^{J} Y_{ij} = 1$, that is,

$$\mathbf{Y}_i \sim \text{Multinomial}(1, \boldsymbol{\psi}_i)$$

where

$$\psi_{ij} = \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j)}{\sum_{j'=1}^{J} \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_{j'})}.$$

To identify coefficients, I fix $\boldsymbol{\beta}_1 = \mathbf{0}$. The model is completed by placing a normal prior on coefficients,

$$\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

In this note, we are interested in estimating the posterior mode of $\boldsymbol{\beta}$.

[*] R package `emlogit` is available to implement the proposed method at https://github.com/soichiroy/emlogit.

[†] Graduate student, Department of Government, Harvard University. Email: syamauchi@g.harvard.edu.

3 Estimation

3.1 Setup

The joint density of $\mathbf{Y}$ and $\boldsymbol{\beta}$ is given by

$$p(\mathbf{Y}, \boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{j=1}^{J} \left[p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j)}{\sum_{j'=1}^{J} \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_{j'})} \right\}^{Y_{ij}} \right] \tag{3.1}$$

By taking log, we have

$$\begin{aligned} \log p(\mathbf{Y}, \boldsymbol{\beta}) &= \sum_{j=1}^{J} \log p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{i=1}^{n} \sum_{j=1}^{J} Y_{ij} \log \left\{ \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j)}{\sum_{j'=1}^{J} \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_{j'})} \right\} \\ &= \sum_{j=1}^{J} \log p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ &\quad + \sum_{i=1}^{n} \left[Y_{ij} \log \left\{ \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j)}{c_{ij} + \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j)} \right\} + \sum_{j' \neq j} Y_{ij'} \log \left\{ \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_{j'})}{c_{ij} + \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j)} \right\} \right] \end{aligned}$$

where $c_{ij} = \sum_{h \neq j} \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_h)$.

Then conditioning on $\boldsymbol{\beta}_{j'}$ for $j' \neq j$, we have

$$\begin{aligned} \log p(\mathbf{Y}, \boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}) &= \log p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{i=1}^{n} \log \left\{ \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j)^{Y_{ij}} / c_{ij}}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j) / c_{ij}} \right\} \\ &= \log p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{i=1}^{n} \log \left\{ \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j - \log c_{ij})^{Y_{ij}}}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta}_j - \log c_{ij})} \right\} \end{aligned}$$

Suppose now, we augment w_{ij} drawn from Polya-Gamma distribution and consider the joint density $p(\mathbf{Y}, \boldsymbol{\omega}_j, \boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j})$. Then, we have that

$$\begin{aligned} \log p(\mathbf{Y}, \boldsymbol{\omega}_j, \boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}) &= \log p(\mathbf{Y}, \boldsymbol{\omega}_j \mid \boldsymbol{\beta}_j, \boldsymbol{\beta}_{-j}) + \log p(\boldsymbol{\beta}_j) \\ &\propto \sum_{i=1}^{n} \Big\{ -\frac{1}{2} \omega_{ij} \psi_{ij}^2 + (Y_{ij} - 1/2) \psi_{ij} \Big\} + \log p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$

where $\psi_{ij} = \mathbf{X}_i^\top \boldsymbol{\beta}_j - \log(c_{ij})$.

3.2 EM-algorithm

- M-step: We cyclically update $\boldsymbol{\beta}_j$ for $j = 2, \dots, J$ by maximizing the following criteria with respect to $\boldsymbol{\beta}_j$ by treating $\boldsymbol{\beta}_{-j}$ as fixed.

$$\begin{aligned} Q_j(\boldsymbol{\beta}_j) &= \mathbb{E}_\omega[\log p(\mathbf{Y}, \boldsymbol{\omega}_j \mid \boldsymbol{\beta}_j, \boldsymbol{\beta}_{-j})] + \log p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ &= \sum_{i=1}^{n} \Big\{ -\frac{1}{2}\mathbb{E}[\omega_{ij}]\psi_{ij}^2 + (Y_{ij} - 1/2)\psi_{ij} \Big\} + \log p(\boldsymbol{\beta}_j \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned}$$

  Let $\mathbf{S} = \mathbf{X}^\top \mathrm{diag}(\{\omega_{ij}\}_{i=1}^n)\mathbf{X}$ and $d_i = \mathbb{E}[\omega_{ij}]\log(c_{ij}) + (Y_{ij} - 1/2)$. Then, the first order condition is

$$\begin{aligned} 0 &= \frac{\partial}{\partial \boldsymbol{\beta}_j} Q_j(\boldsymbol{\beta}_j) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}_j} \Big\{ -\frac{1}{2}\boldsymbol{\beta}_j^\top(\mathbf{S} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\beta}_j + \boldsymbol{\beta}_j^\top(\mathbf{X}^\top \boldsymbol{d} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0) \Big\} \\ &= -(\mathbf{S} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\beta}_j + (\mathbf{X}^\top \boldsymbol{d} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0) \end{aligned}$$

  which implies the closed form update:

$$\widehat{\boldsymbol{\beta}}_j \leftarrow (\mathbf{S} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\mathbf{X}^\top \boldsymbol{d} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0). \tag{3.2}$$

- E-step: We update $\mathbb{E}[\omega_{ij}]$ for $i = 1, \dots, n$ and $j = 1, \dots, J$. This expectation is over the posterior distribution of ω_{ij} evaluating $\boldsymbol{\beta}$ at the current value. Since $\omega_{ij} \mid \boldsymbol{\beta}, \mathbf{X}_i \sim \mathrm{PG}(1, \widehat{\boldsymbol{\psi}}_{ij})$, we can evaluate the expectation by

$$\mathbb{E}[\omega_{ij}] \leftarrow \frac{1}{2\widehat{\psi}_{ij}}\tanh(\widehat{\psi}_{ij}/2) \tag{3.3}$$

  where $\widehat{\psi}_{ij} = \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_j - \log \sum_{j' \neq j} \exp(\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_{j'})$. This step is a direct application of the following fact: If $\omega_i \sim \mathrm{PG}(b, c)$, then

$$\mathbb{E}[\omega_i] = \frac{b}{2c}\tanh(c/2).$$