EM Algorithm for Linear Mixed Model

Mingxuan Cai

August 2025

Consider a dataset $\{\mathbf{y}, \mathbf{X}, \mathbf{Z}\}$ with n samples, where $\mathbf{y} \in \mathbb{R}^n$ is the vector of response variable, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix of p independent variables, and $\mathbf{Z} \in \mathbb{R}^{n \times c}$ is another matrix of c variables. The linear mixed model builds upon a linear relationship from $\mathbf{y}$ to $\mathbf{X}$ and $\mathbf{Z}$ by

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\omega} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\boldsymbol{\omega} \in \mathbb{R}^c$ is the vector of fixed effects, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of random effects with $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 \mathbf{I}_p)$, and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$ is the independent noise term. Let $\boldsymbol{\Theta}$ denote the set of unknown parameters $\boldsymbol{\Theta} = \{\boldsymbol{\omega}, \sigma_\beta^2, \sigma_e^2\}$. We can treat $\boldsymbol{\beta}$ as a latent variable because is it unobserved.

1. Read Chapter 9 of PRML. Derive and implement an Expectation-Maximization (EM) algorithm for the above linear mixed model and return the marginal likelihood at each iteration, the estimate of $\boldsymbol{\Theta}$, the and posterior mean of $\boldsymbol{\beta}$.

2. Apply the implementation to the dataset here. The XYZ_MoM.txt file has the first column as $\mathbf{y}$. The columns labeled with Z.1∼Z.30 correspond to the matrix $\mathbf{Z}$ and the columns labeled with X.1∼X.6000 correspond to the matrix $\mathbf{X}$. Report the parameter estimates.

3. Read Chapter 10 of PRML. Suppose that we are using mean-field variational inference (MFVI) $q(\boldsymbol{\beta})$ to approximate the true posterior distribution $\Pr(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}; \boldsymbol{\Theta})$, where $q(\boldsymbol{\beta}) = \prod_{j=1}^{p} q(\boldsymbol{\beta}_j)$. Derive an algorithm to obtain optimal mean-field approximation $q^*(\boldsymbol{\beta})$ and estimate model parameters $\boldsymbol{\Theta}$ (hint: in the E-step, you optimize $q(\boldsymbol{\beta})$, and in the M-step, you optimize $\boldsymbol{\Theta}$). Track both the marginal log-likelihood and the evidence lower bound (ELBO). Note the gap between them. Compare the posterior mean and variance of $\boldsymbol{\beta}$ obtained through MFVI and EM.

4. Summarize your understanding of the model, derivation of the algorithms, and data application results, and present using PPT slides.