Approximating MLEs with Variational Inference and Monte Carlo EM

Ryan Halstater

Setting

Data: ($\mathbf{Y}_o$, $\mathbf{Y}_m$)

Latent Parameters: $\boldsymbol{\theta}$

Example: Poisson Generalized Linear Mixed Model

Latent Variables: β, σ^2, τ^2

$$Y_{it} \sim \text{Pois}(\lambda_{it})$$

$$\log \lambda_{it} = X_{it}\boldsymbol{\beta} + \gamma_i + \epsilon_{it}$$

$$\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad \gamma_i \overset{\text{iid}}{\sim} N(0, \tau^2) \quad \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\gamma}$$

Observed Data	Missing Data
X_{11}, Y_{11}	γ_1
X_{12}, Y_{12}	γ_1
X_{21}, Y_{21}	γ_2
X_{22}, Y_{22}	γ_2

Problem

Goal: Maximize (intractable) observed likelihood

$$\mathscr{L}(\boldsymbol{\theta}|\mathbf{Y}_o) = p(\mathbf{Y}_o|\boldsymbol{\theta}) = \int p(\mathbf{Y}_o, \mathbf{Y}_m|\boldsymbol{\theta})\mathrm{d}\mathbf{Y}_m$$

We explore two distinct ways to do this:

Monte Carlo Expectation Maximization: Iteratively converge to “viable” estimate of latent variables by approximating a surrogate function using MCMC, then maximizing it

Variational Inference (Frequentist): Simultaneously estimate the latent variables and approximate the missing data by picking a distribution from a family of “friendly” distributions that, together, maximize a lower bound on the observed likelihood

EM - A Review

Goal: Maximize

$$\mathscr{L}(\boldsymbol{\theta}|\mathbf{Y}_o) = p(\mathbf{Y}_o|\boldsymbol{\theta}) = \int p(\mathbf{Y}_o, \mathbf{Y}_m|\boldsymbol{\theta})\mathrm{d}\mathbf{Y}_m$$

Method: Iteratively

(E) Compute:

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = \mathrm{E}_{\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}_o, \mathbf{Y}_m; \boldsymbol{\theta} \mid \mathbf{Y}_o, \boldsymbol{\theta}^{(t)})]$$

(M) Maximize with respect to $\boldsymbol{\theta}$

Notes:

- $\boldsymbol{\theta}^{(t)}$ is only used to integrate out the missing data $\mathbf{Y}$□
- This method has ascent property

Issue: what if this is intractable too?

MCEM: Basic Version

Idea: Approximate the expectation using MCMC

(E): Markov Chain generates samples $\mathbf{Y}_m^{(t,j)} \sim p(\mathbf{Y}_m | \mathbf{Y}_o, \boldsymbol{\theta}^{(t)})$

Then, compute

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = \mathrm{E}_{\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}_o, \mathbf{Y}_m; \boldsymbol{\theta} \mid \mathbf{Y}_o, \boldsymbol{\theta}^{(t)})] \approx \frac{\sum_{j=1}^{m_t} \log p(\mathbf{Y}_o, \mathbf{Y}_m^{(t,j)}; \boldsymbol{\theta})}{m_t}$$

- Does not inherently preserve ascent property, more detailed methods have been developed (Ascent Based MCEM, Caffo et al, 2005)
    - Achieved by computing and controlling asymptotic lower bound for

$$Q\left(\boldsymbol{\theta}^{(t+1)} | \boldsymbol{\theta}^{(t)}\right) - Q\left(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t)}\right)$$

Variational Inference

- Uses **Importance sampling** and Jensen's inequality to maximize the bound on how badly the log likelihood can be approximated (Evidence Lower BOund)
- q(· ;**ω**) is from a family of distributions $\mathbb{D}$ meant to approximate the missing data

$$
\begin{aligned}
p(\mathbf{Y}_o;\boldsymbol{\theta}) &= \int p(\mathbf{Y}_o,\mathbf{Y}_m;\boldsymbol{\theta})d\mathbf{Y}_m \\
&= \int \frac{p(\mathbf{Y}_o,\mathbf{Y}_m;\boldsymbol{\theta})}{q(\mathbf{Y}_m;\boldsymbol{\omega})} q(\mathbf{Y}_m;\boldsymbol{\omega})d\mathbf{Y}_m \\
&= \mathbb{E}_{\mathbf{Y}_m\sim q(\cdot;\boldsymbol{\omega})}\left(\frac{p(\mathbf{Y}_o,\mathbf{Y}_m;\boldsymbol{\theta})}{q(\mathbf{Y}_m;\boldsymbol{\omega})}\right)
\end{aligned}
$$

Variational Inference

- Uses Importance sampling and **Jensen's inequality** to maximize the bound on how badly the log likelihood can be approximated (Evidence Lower BOund)
- q is a family of distributions meant to approximate missing data and to have computation-friendly densities

$$
\begin{aligned}
\ell(\boldsymbol{\theta} \mid \mathbf{Y}_o) &= \log p(\mathbf{Y}_o; \boldsymbol{\theta}) \\
&= \log \mathbb{E}_{\mathbf{Y}_m \sim q(\cdot; \boldsymbol{\omega})} \left(\frac{p(\mathbf{Y}_o, \mathbf{Y}_m; \boldsymbol{\theta})}{q(\mathbf{Y}_m; \boldsymbol{\omega})} \right) \\
&\geq \mathbb{E}_{\mathbf{Y}_m \sim q(\cdot; \boldsymbol{\omega})} \left(\log \frac{p(\mathbf{Y}_o, \mathbf{Y}_m; \boldsymbol{\theta})}{q(\mathbf{Y}_m; \boldsymbol{\omega})} \right) \\
&= \mathbb{E}_{\mathbf{Y}_m \sim q(\cdot; \boldsymbol{\omega})} \log p(\mathbf{Y}_o, \mathbf{Y}_m; \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{Y}_m \sim q(\cdot; \boldsymbol{\omega})} (\log q(\mathbf{Y}_m; \omega)) \\
&= \mathrm{ELBO}(\boldsymbol{\omega}, \boldsymbol{\theta} \mid \mathbf{Y}_o).
\end{aligned}
$$

Future Plans: Simulation Study

- Apply Variational Inference and basic MCEM to Poisson GLMM

$$Y_{it} \sim \text{Pois}(\lambda_{it})$$

$$\log \lambda_{it} = X_{it}\boldsymbol{\beta} + \gamma_i + \epsilon_{it}$$

$$\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad \gamma_i \overset{\text{iid}}{\sim} N(0, \tau^2) \quad \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\gamma}$$

- X being Gaussian noise
- Balanced design
- Testing under 2 random intercept variances * 2 numbers of random intercept groups
- Metrics: MSE of parameter estimates, runtime