GLMM Lasso for High-dimensional Data

Presenter: Junjie Liang

College of IST

jul672@ist.psu.edu

GLMM – problem definition

Likelihood for GLM is:

$$f(y_i|\theta_i, \phi) = \exp\left(\phi^{-1}\left(y_i \cdot \theta_i - \xi(\theta_i)\right) + c(y_i, \phi)\right)$$

where ϕ and θ_i are model parameters. For GLMM:

$h(\theta_i) = \mu_i = \mathbb{E}[y_i|b_i] = g^{-1}(x_i\beta + \color{red}{z_i b_i})$ where $b_i \sim N(0, Q^{-1})$

For GLMM, since we have unobserved random effect, the augmented likelihood is:

$$f(y_i; \theta_i, \phi, Q) = \int f(y_i|\theta_i, \phi, b_i) \cdot f(b_i|Q)\mathrm{d}b_i$$

This is intractable because we have no close form solution to the integral!

GLMM – solutions

Bayesian inference:

- Variational Bayesian Inference (a.k.a. Variational Bayes) (VBI)
- MCMC with auxiliary variables (MCMC)

ML estimation:

- Monte Carlo EM (MC-EM)
- Laplace approximation (LA)

Variational Bayesian Inference

- We want to know $f(\boldsymbol{\theta}|y) = \frac{f(y|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(y)}$, but:
  - don't know the normalizing constant $f(y)$
  - Computing $f(y|\boldsymbol{\theta})$ is also very difficult (e.g., GLMM)
- Let's use a proposal distribution $q(\boldsymbol{\theta})$ to approximate $f(\boldsymbol{\theta}|y)$.
- A good proposal distribution would be to minimize the KL divergence

$$q^*(\boldsymbol{\theta}) = \min_{q(\boldsymbol{\theta})} KL\big(q(\boldsymbol{\theta})||f(\boldsymbol{\theta}|y)\big)$$

- The best solution is when $q(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|y)$.
- Simplifying the KL divergence, we find that

$$q^*(\boldsymbol{\theta}) = \min_{q(\boldsymbol{\theta})} KL\big(q(\boldsymbol{\theta})||f(\boldsymbol{\theta}|y)\big) = \min_{q(\boldsymbol{\theta})} KL\big(q(\boldsymbol{\theta})||\color{red}{f(\boldsymbol{\theta},y)}\big)$$

- Therefore, the best solution is converted to when $q(\boldsymbol{\theta}) = f(y|\boldsymbol{\theta})f(\boldsymbol{\theta})$, but this is still intractable.

Variational Bayesian Inference (cont.)

- What if we assume $q(\boldsymbol{\theta}) = \prod_i q(\theta_i)$ (mean field theory)
- Then we have

$$q^*(\theta_i) = \min_{q(\theta_i)} KL\big(q(\boldsymbol{\theta})||f(\boldsymbol{\theta}, y)\big)$$

where:

$$\begin{aligned}
&\min_{q(\theta_i)} KL\big(q(\boldsymbol{\theta})||f(\boldsymbol{\theta}, y)\big) \\
&= \min_{q(\theta_i)} \int \prod_i q(\theta_i) \log \frac{\prod_i q(\theta_i)}{f(\boldsymbol{\theta}, y)} d\boldsymbol{\theta} \\
&= \min_{q(\theta_i)} \int \prod_i q(\theta_i) \log \prod_i q(\theta_i)\, d\boldsymbol{\theta} - \int \prod_i q(\theta_i) \log f(\boldsymbol{\theta}, y)\, d\boldsymbol{\theta} \\
&= \min_{q(\theta_i)} \int q(\theta_i) \log q(\theta_i)\, d\theta_i - \int q(\theta_i) \mathbb{E}_{q(\theta_{-i})}[\log f(\boldsymbol{\theta}, y)] d\theta_i \\
&= \min_{q(\theta_i)} KL\big(q(\theta_i)|| \exp\big(\mathbb{E}_{q(\theta_{-i})}[\log f(\boldsymbol{\theta}, y)]\big)\big)
\end{aligned}$$

Therefore, following the mean field theory, we want to find $\theta_i, i = 1, \dots, |\theta|$:

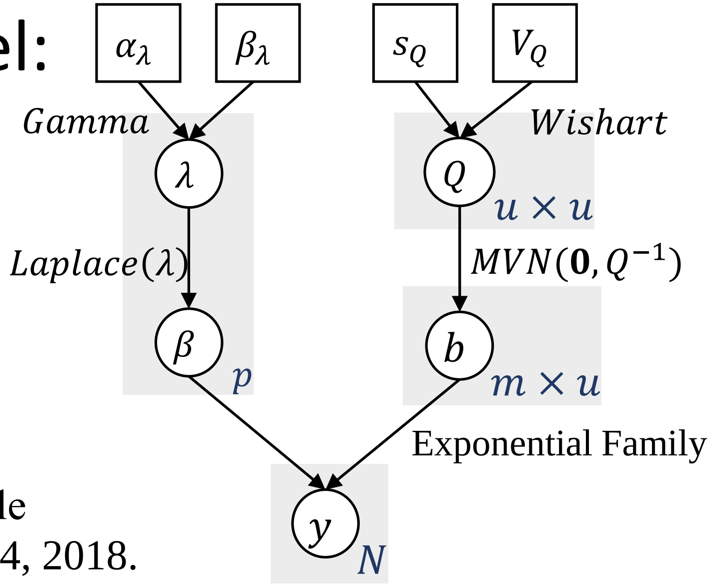
$$q(\theta_i) \propto \exp\big(\mathbb{E}_{q(\theta_{-i})}[\log f(\boldsymbol{\theta}, y)]\big)$$

where $q(\theta_i)$ should be a valid distribution.

Variational Bayesian Inference (cont.)

General guidance for finding $q(\theta_i)$:

- Following $q^*(\theta_i) = \min\limits_{q(\theta_i)} KL\big(q(\boldsymbol{\theta})||f(\boldsymbol{\theta}, y)\big)$. Find $q^*(\theta_i)$ directly using conjugate distribution.
- Following $q^*(\theta_i) \propto \exp\big(\mathbb{E}_{q(\theta_{-i})}[\log f(\boldsymbol{\theta}, y)]\big)$. Find $q^*(\theta_i)$ using some tricks.
- In BI framework for GLMM, [1] gives a graphical model:

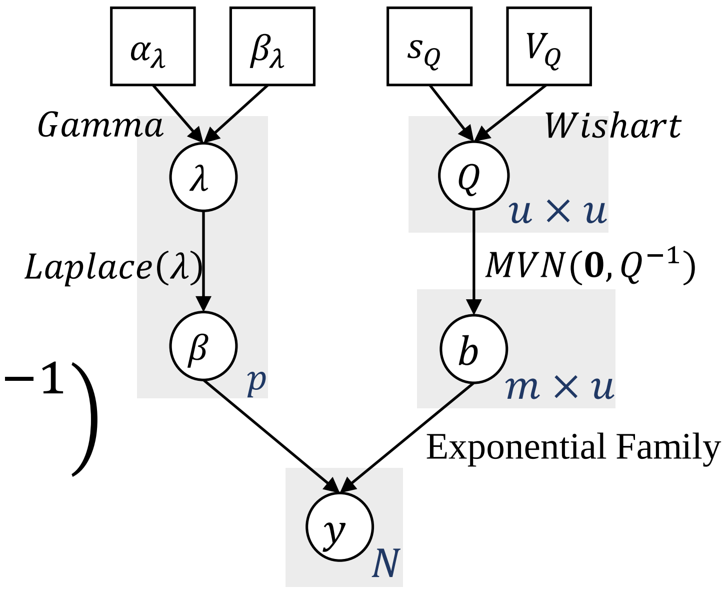


Therefore, our goal is:

$$q^*(\lambda, \beta, Q, b) = \min_q KL\big(q(\lambda, \beta, Q, b)||f(\lambda, \beta, Q, b, y)\big)$$

[1] D. T. Tung, M.-N. Tran, and T. M. Cuong, “Bayesian adaptive lasso with variational Bayes for variable selection in high-dimensional generalized linear mixed models,” *Commun. Stat.-Simul. Comput.*, pp. 1–14, 2018.

Solving GLMM with VBI



Following mean field theory, we have:

- $q^*(Q) \sim Wishart\left(s_Q + m, \left(V_Q^{-1} + \sum_{i=1}^{m} \mathbb{E}_{q(b)}\left[b_i \cdot b_i^T\right]\right)^{-1}\right)$
- $q^*(\lambda) \sim Gamma\left(\alpha_\lambda + 1, \beta_\lambda + \mathbb{E}_{q(\beta)}[|\beta|]\right)$
- $q^*(b) \propto \exp \mathbb{E}_{-q(b)}[\log f(y|b,\beta) f(b|Q)]$
- $q^*(\beta) \propto \exp \mathbb{E}_{-q(\beta)}[\log P(y|b,\beta) P(\beta|\lambda)]$

For $q(b)$ and $q(\beta)$, I use gaussian approximation, thus:

$$q(b_i) \sim MVN\left(b_i^*, -\left(Z_i^T \cdot \mathrm{Diag}\{\zeta''\left(\eta_i(b_i^*)\right) \times_e \eta_i''(b_i^*)\} \cdot Z_i - \mathbb{E}_{q(Q)}[Q]\right)^{-1}\right)$$

$$q(\beta) \sim MVN\left(\beta^*, -\left(X^T \cdot \mathrm{Diag}\{\zeta''\left(\eta_i(\beta^*)\right) \times_e \eta_i''(\beta^*)\} \cdot X\right)^{-1}\right)$$

Generalized EM algorithm

In EM algorithm, for iteration t, we do the following:

$$f\left(y|\theta^{(t)}\right) = \int f\left(y,z|\theta^{(t)}\right)dz = \int \frac{f\left(y,z|\theta^{(t)}\right)}{q^{(t)}(z)} q^{(t)}(z)dz$$

where $q^{(t)}(z) = f\left(z|y,\theta^{(t)}\right)$. Let's continue to use $q(z)$, then we have:

$$f\left(y|\theta^{(t)}\right) = \mathbb{E}_{q^{(t)}(z)}\left[\frac{f\left(y,z|\theta^{(t)}\right)}{q^{(t)}(z)}\right]$$

Assume that we want to maximize the log-likelihood, then:

$$l\left(\theta^{(t)};y\right) = \max_{\theta} \log f\left(y|\theta^{(t)}\right) \geq \mathbb{E}_{q^{(t)}(z)}\left[\log \frac{f\left(y,z|\theta^{(t)}\right)}{q^{(t)}(z)}\right]$$

$$= \int q^{(t)}(z) \log \frac{f\left(z|y,\theta^{(t)}\right) f\left(y|\theta^{(t)}\right)}{q^{(t)}(z)} dz$$

$$= \log f\left(y|\theta^{(t)}\right) - KL\left(q^{(t)}(z) || f\left(z|y,\theta^{(t)}\right)\right)$$

Therefore, when $q^{(t)}(z) = f\left(z|y,\theta^{(t)}\right)$, we have $KL\left(q^{(t)}(z) || f\left(z|y,\theta^{(t)}\right)\right) = 0$, which is the optimal solution.

Generalized EM algorithm

For the E-step, what we exactly want is to compute:

$$f\left(y|\theta^{(t)}\right) = \mathbb{E}_{q^{(t)}(z)}\left[\frac{f\left(y,z|\theta^{(t)}\right)}{q^{(t)}(z)}\right]$$

This can be achieved by two ways:

1. If we can draw samples from $q^{(t)}(z)$, then we don't need to know the form of $q^{(t)}(z)$. This results in MC-EM algorithm.
2. If we need the close form distribution of $q^{(t)}(z)$, then our goal is:

$$q^{(t)}(z) = \min_{q(z)} KL\left(q(z)||f\left(z|y,\theta^{(t)}\right)\right)$$

   This results in VBI.

Generalized EM algorithm:

E-step: solve $q^{(t)}(z) = \min_{q(z)} KL\left(q(z)||f\left(z|y,\theta^{(t)}\right)\right)$

M-step: solve $\theta^{(t+1)} = \max_{\theta} \mathbb{E}_{q^{(t)}(z)}\left[\frac{f\left(y,z|\theta^{(t)}\right)}{q^{(t)}(z)}\right]$

Solving GLMM with MCEM [2]

We want to optimize the following objective function:

$$\hat{\theta} = \max_{\theta} \sum_{i=1}^{m} \log \int_{\mathbb{R}^q} f(y_i|\beta, b_i, Q) f(b_i|Q) \mathrm{d}b_i - \lambda \left|\left|\beta\right|\right|_1$$

E-step: we want to solve $\mathbb{E}_{b_i}\left[\log \frac{f(y_i|\theta_i, b_i) f(b_i|Q)}{f\left(b_i|\theta_i^{(t)}, y_i\right)} |\theta_i^{(t)}, y_i\right]$

M-step: $\hat{\theta}^{(t+1)} = \max_{\theta} \sum_{i=1}^{m} \mathbb{E}_{b_i}\left[\log \frac{f(y_i|\theta_i, b_i) f(b_i|Q)}{f\left(b_i|\theta_i^{(t)}, y_i\right)} |\theta_i^{(t)}, y_i\right] - \lambda \left|\left|\beta\right|\right|_1$

For E step, I apply the Metropolis-Hasting algorithm, where

$$f\left(b_i|\beta^{(t)}, Q^{(t)}, y_i\right) = \frac{f\left(y_i|b_i, \beta^{(t)}\right) f\left(b_i|Q^{(t)}\right) f\left(\beta^{(t)}\right)}{f(y_i, \beta^{(t)}, Q^{(t)})}$$

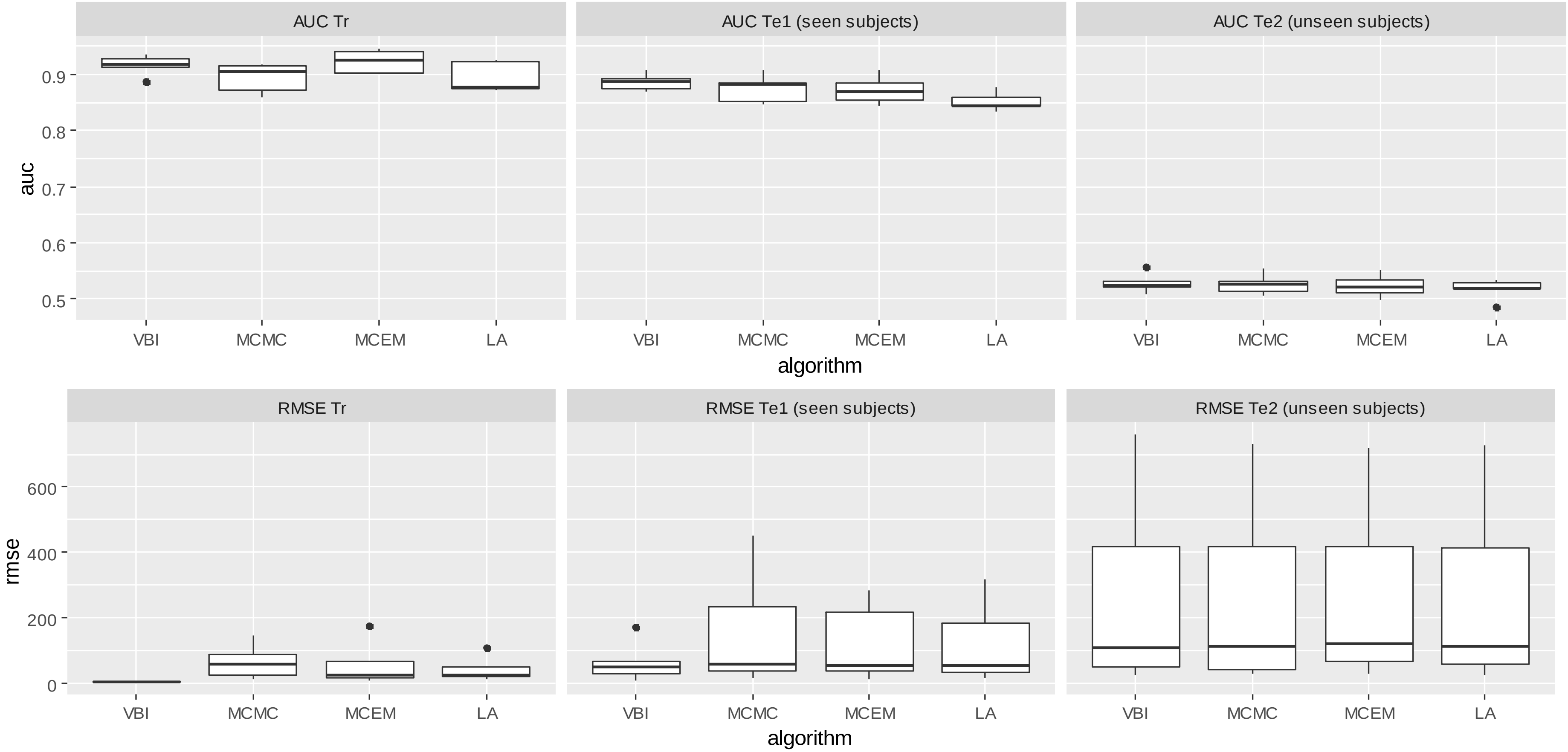
C. E. McCulloch, “Maximum likelihood algorithms for generalized linear mixed models,” *J. Am. Stat. Assoc.*, vol. 92, no. 437, pp. 162–170, 1997.

Simulation study

- Logistic regression and Poisson regression.
- BI: VBI and MCMC;   MLE: MC-EM and LA
- Simulation process:
  - Generate β, X, Z.
  - For each subject, generate $b_i \sim N(0, Q^{-1})$
  - Compute $\mu_{it} = g^{-1}(x_{it}\beta + z_{it}b_i)$
  - Simulate y_{it} with mean μ_{it}
  - Simulate two test sets. One is to reuse b_i, assuming predicting the outcome for existing subjects; one is only reusing β and re-simulate b_i, assuming predicting the outcome for new subjects.
- Evaluation metrics:
  - AUC for Logistic Reg, RMSE for Poisson Reg
  - Runtime
  - Overall coverage rate/coverage rate for sparsity

Results

Algorithm	Runtime/ iteration	AUC for Tr	AUC for Te_1	AUC for Te_2	Coverage	Coverage sparse
VBI	93.03	0.977	0.824	0.572	0.397	1.000
MCMC	224.74/100	1.000	0.879	0.573	0.064	0.089
MCEM	215.23	1.000	0.825	0.536	1.000	1.000
LA	932.46	0.962	0.801	0.550	1.000	1.000



Conclusion

- In all cases, algorithms based on BI model is faster than MLE
- For low-dimensional case, VBI > MC-EM > MCMC > LA
- For high-dimensional case, algorithms based on BI model outperform MLE
- MLE usually has larger variance.
- Difficulty of derivation and implementation:

VBI > MC-EM > LA > MCMC