Fitting and diagnosing generalized linear mixed models using a partially noncentered parametrization

Linda Tan

(Joint work with David Nott)

National University of Singapore

Australian Statistical Conference in conjunction with the Institute of Mathematical Statistics Annual Meeting
8 July 2014
Generalized linear mixed models (GLMMs)

- Wide applications
- Estimation is challenging (integral over random effects intractable)
- Computationally intensive: numerical quadrature, MCMC
- Approximate methods:
  - Penalized quasi-likelihood (Breslow and Clayton 1993)
  - Laplace approximation and its extensions (Raudenbush et al. 2000)
  - Integrated nested Laplace approximations (Fong et al. 2010)
  - Gaussian variational approximation (Ormerod and Wand 2012)
We propose to

- fit GLMMs using nonconjugate variational message passing (NCVMP, Knowles and Minka 2011)
- derive prior-likelihood conflict diagnostics from NCVMP
- reparametrize GLMMs using partial noncentering
  - Used in hierarchical models to boost efficiency in MCMC algorithms (Papaspiliopoulos et al. 2003; 2007)
  - Accelerate convergence and produce more accurate posterior approximations in variational algorithms
Variational Approximation

- \( y \): observed data, \( \theta \): set of unknown parameters
- Approximate \( p(\theta|y) \) by more tractable density function \( q(\theta) \)
- Minimize Kullback-Leibler divergence between \( q(\theta) \) and \( p(\theta|y) \)

\[
\log p(y) = \int q(\theta) \log \frac{p(y, \theta)}{q(\theta)} \, d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} \, d\theta
\]

- Lower bound (\( \mathcal{L} \))
- Kullback-Leibler divergence \( \geq 0 \)

Maximizing \( \mathcal{L} \) \( \iff \) minimizing Kullback-Leibler divergence
Variational Bayes (VB)

- Assume \( q(\theta) = \prod_{i=1}^{m} q_i(\theta_i) \) for \( \theta = \{\theta_1, \ldots, \theta_m\} \)
  - Optimal densities maximizing lower bound \( \mathcal{L} \) satisfy
    \[
    q_i(\theta_i) \propto \exp\{E_{-\theta_i} \log p(y, \theta)\} \quad \text{for each } i
    \]
    \( E_{-\theta_i} \): expectation w.r.t. \( \prod_{j \neq i} q_j(\theta_j) \)

- Conjugate priors:
  - Density of optimal \( q_i \) recognizable
  - Suffices to update parameters of \( q_i \)

- Variational message passing (Winn and Bishop, 2005)
  - General algorithm for applying VB to conjugate-exponential models
Nonconjugate variational message passing (NCVMP)

- Assume
  1. $q(\theta) = \prod_{i=1}^{m} q_i(\theta_i)$ for $\theta = \{\theta_1, \ldots, \theta_m\}$ (VB)
  2. each $q_i$ belongs to some exponential family:
     
     $q_i(\theta_i) = \exp\{\lambda_i^T t_i(\theta_i) - h_i(\lambda_i)\}$

     $\lambda_i$: vector of natural parameters, $t_i(\cdot)$: sufficient statistics

- $\nabla_\lambda L = 0$ when $L$ is maximized w.r.t. $\lambda_i$

NCVMP Algorithm

Initialize $\lambda_i$ for $i = 1, \ldots, m$.

Cycle:

$\lambda_i \leftarrow \mathbf{V}_i(\lambda_i)^{-1} \nabla_\lambda E_q\{\log p(y, \theta)\}$ for $i = 1, \ldots, m$

until convergence

- $\mathbf{V}_i(\lambda_i)$: covariance matrix of $t_i(\theta_i)$
Nonconjugate variational message passing (NCVMP)

- Fixed-point iteration algorithm
- Convergence problems (fix using damping)
- When \( q_i(\theta_i) = N(\mu_i, \Sigma_i) \), update for \( \lambda_i \) (Wand, 2014):

\[
\Sigma_i \leftarrow -\frac{1}{2} \left[ \text{vec}^{-1} \left( \frac{\partial E_q\{\log p(y, \theta)\}}{\partial \text{vec}(\Sigma_i)} \right) \right]^{-1}
\]

\[
\mu_i \leftarrow \mu_i + \Sigma_i \frac{\partial E_q\{\log p(y, \theta)\}}{\partial \mu_i}
\]

- Explicit updates reduce computational cost significantly
Generalized linear mixed models

One-parameter exponential family (e.g. Bernoulli, Poisson):

- $y_{ij}$: $j$th response in cluster $i$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$.

- Conditional on $u_i (r \times 1) \sim iid \sim N(0, D)$,

$$y_{ij} | u_i \sim iid \sim \exp \{y_{ij} \zeta_{ij} - b(\zeta_{ij}) + c(y_{ij})\}$$

- $\zeta_{ij}$: canonical parameter
- $b(\cdot), c(\cdot)$: functions specific to exponential family
- $\mu_{ij} = E(y_{ij} | u_i)$ and $g(\mu_{ij}) = \eta_{ij}$ for link function $g(\cdot)$, where

$$\eta_{ij} = x_{ij}^T \beta + z_{ij}^T u_i$$

- $x_{ij} (p \times 1)$ and $z_{ij} (r \times 1)$: vectors of covariates
- $\beta (p \times 1)$: unknown fixed regression parameters.
Generalized linear mixed models

- For $i$th cluster:

\[
y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \eta_i = \begin{bmatrix} \eta_{i1} \\ \vdots \\ \eta_{in_i} \end{bmatrix}, \quad X_i = \begin{bmatrix} x_{i1}^T \\ \vdots \\ x_{in_i}^T \end{bmatrix}, \quad Z_i = \begin{bmatrix} z_{i1}^T \\ \vdots \\ z_{in_i}^T \end{bmatrix}.
\]

- Assume
  - first column of $Z_i$ is $1_{n_i}$ if $Z_i$ is not a zero matrix
  - columns of $Z_i$ are a subset of columns of $X_i$

- Priors:
  - Diffuse prior: $\beta \sim N(0, \Sigma_\beta)$
  - Default conjugate prior: $D \sim IW(\nu, S)$ (Kass and Natarajan, 2006)
Partial noncentering

Linear Mixed model:

\[ y_i = X_i \beta + Z_i u_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \text{ for } i = 1, \ldots, n \]

- Assume \( p(\beta) \propto 1 \) and \( \sigma^2, D \) are known
- Suppose \( X_i = Z_i \). Let \( \alpha_i = (\beta + u_i) \sim N(\beta, D) \)

A partially noncentered parametrization:

- Let \( \tilde{\alpha}_i = \alpha_i - W_i \beta \) where \( W_i_{(r \times r)} \) is a tuning matrix
- Centered (\( W_i = 0 \)) and noncentered (\( W_i = I_r \))
- Rewrite \( y_i = Z_i W_i \beta + Z_i \tilde{\alpha}_i + \epsilon_i \text{ where } \tilde{\alpha}_i \sim N((I - W_i) \beta, D) \)
Partial noncentering in VB

- Apply VB: \( q(\theta) = q(\beta) \prod_{i=1}^{n} q(\tilde{\alpha}_i) \)
- Optimal densities: \( q(\beta) = N(\mu^q_\beta, \Sigma^q_\beta), \ q(\tilde{\alpha}_i) = N(\mu^q_{\tilde{\alpha}_i}, \Sigma^q_{\tilde{\alpha}_i}) \)

**VB Algorithm: Linear mixed model**

Initialize \( \mu^q_{\tilde{\alpha}_i} \) and \( \Sigma^q_{\tilde{\alpha}_i} \) for \( i = 1, \ldots, n \).

Cycle:

1. \( \Sigma^q_\beta \leftarrow \left[ \sum_{i=1}^{n} \left\{ (I - W_i)^T D^{-1} (I - W_i) + \frac{1}{\sigma^2} W_i^T Z_i^T Z_i W_i \right\} \right]^{-1} \)
   \( \mu^q_\beta \leftarrow \Sigma^q_\beta \sum_{i=1}^{n} \left[ \frac{1}{\sigma^2} W_i^T Z_i^T y_i + \left\{ D^{-1} (I - W_i) - \frac{1}{\sigma^2} Z_i^T Z_i W_i \right\}^T \mu^q_{\tilde{\alpha}_i} \right] \)

2. For \( i = 1, \ldots, n, \)
   - \( \Sigma^q_{\tilde{\alpha}_i} \leftarrow \left( D^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i \right)^{-1} \)
   - \( \mu^q_{\tilde{\alpha}_i} \leftarrow \Sigma^q_{\tilde{\alpha}_i} \left[ \frac{1}{\sigma^2} Z_i^T y_i + \left\{ D^{-1} (I - W_i) - \frac{1}{\sigma^2} Z_i^T Z_i W_i \right\} \mu^q_\beta \right] \)

until convergence
Partial noncentering in variational Bayes

- VB algorithm converges instantly when
  \[ W_i = (Z_i^T Q_i Z_i + D^{-1})^{-1} D^{-1} \]
  where \( Q_i = \frac{1}{\sigma^2} I_r \)  \( (1) \)

- Partial noncentering yields more rapid convergence than centering or noncentering

- True posteriors recovered in (1) but not in centered or noncentered parametrizations

- Assumption of a factorized posterior in VB often results in underestimation of posterior variance

- Partial noncentering captures dependence between fixed and random effects via tuning parameters \( W_i \) so that true posterior can be recovered
Partially noncentered parametrization for GLMMs

- Partition $X_i$ as $[Z_i \quad X_{si} \quad X_{gi}]$ and $\beta$ as $[\beta_z^T, \beta_s^T, \beta_g^T]^T$
  - $X_{si} = 1_{n_i}x_{si}^T$ (subject specific covariates)

- Linear predictor
  $$\eta_i = Z_i(\beta_z + u_i) + 1_{n_i}x_{si}^T\beta_s + X_{gi}\beta_g$$
  $$= Z_i(C_i\beta_c + u_i) + X_{gi}\beta_g \quad \text{where} \quad \beta_c = \begin{bmatrix} \beta_z \\ \beta_s \end{bmatrix}, \quad C_i = \begin{bmatrix} I_r & x_{si}^T \\ 0 & 0 \end{bmatrix}$$

- Let $\alpha_i = C_i\beta_c + u_i \sim N(C_i\beta_c, D)$

- Let $\tilde{\alpha}_i = \alpha_i - W_iC_i\beta_c \sim N(\tilde{W}_i\beta, D)$ where $\tilde{W}_i = [(I_r - W_i)C_i \quad 0]$

- Partially noncentered parametrization:
  $$\eta_i = V_i\beta + Z_i\tilde{\alpha}_i \quad \text{where} \quad V_i = [Z_i W_i C_i \quad X_{gi}]$$
Specification of tuning parameters

- **Linear mixed model:**
  \[ W_i = (Z_i^T Q_i Z_i + D^{-1})^{-1} D^{-1} \]
  where \( Q_i = \frac{1}{\sigma^2} I_r \)
  \[ = (\mathcal{I}_f + D^{-1})^{-1} D^{-1} \]
  where \( \ell = \log p(y_i | \beta, \alpha_i) \) and \( \mathcal{I}_f = -\frac{\partial^2 \ell}{\partial \alpha_i \partial \alpha_i^T} \)

- **Extend partial non-centering to logistic and Poisson GLMMs:**
  - **Logistic:** \( Q_i = \text{diag}\left( \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \right)^2 \)
  - **Poisson:** \( Q_i \approx \text{diag}(y_i) \)

- **Initialize** \( W_i \) **using penalized quasi-likelihood**

- **Keep** \( W_i \) **fixed or update at the end of each cycle**
Variational approximation for GLMMs

\[ \theta = \{ \beta, D, \tilde{\alpha} \}. \text{ Consider} \]

\[ q(\theta) = q(\beta)q(D) \prod_{i=1}^{n} q(\tilde{\alpha}_i) \]

- Optimal \( q(D) \) is \( IW(\nu^q, S^q) \)
- Bernoulli or Poisson: \( p(y_i|\beta, \tilde{\alpha}_i) \) nonconjugate w.r.t. priors over \( \beta \) and \( \tilde{\alpha}_i \)
- Optimal densities for \( q(\beta) \) and \( q(\tilde{\alpha}_i) \) not recognizable
- NCVMP: assume \( q(\beta) \) is \( N(\mu^q_\beta, \Sigma^q_\beta) \), \( q(\tilde{\alpha}_i) \) is \( N(\mu^q_{\tilde{\alpha}_i}, \Sigma^q_{\tilde{\alpha}_i}) \)
  - Poisson: lower bound and all updates in closed form
  - Logistic: evaluate expectations using Gauss-Hermite quadrature
NCVMP algorithm for GLMMs

Initialize $\mu^q_{\beta}$, $\Sigma^q_{\beta}$, $S^q$ and $\mu^q_{\tilde{\alpha}_i}$, $\Sigma^q_{\tilde{\alpha}_i}$, $W_i$ for $i = 1, \ldots, n$. Set $\nu^q = n + \nu$.

Cycle:

1. Update $W_i$ and hence $V_i$ for $i = 1, \ldots, n$. (Optional)

2. Update local variational parameters for $i = 1, \ldots, n$
   
   - $\Sigma^q_{\tilde{\alpha}_i} \leftarrow \left(\nu^q S^q^{-1} + Z_i^T F_i Z_i\right)^{-1}$
   
   - $\mu^q_{\tilde{\alpha}_i} \leftarrow \mu^q_{\tilde{\alpha}_i} + \Sigma^q_{\tilde{\alpha}_i} \left\{ Z_i^T (y_i - G_i) - \nu^q S^{-1} (\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_{\beta}) \right\}$

3. Update global variational parameters

   - $\Sigma^q_{\beta} \leftarrow \left\{ \Sigma^{-1}_{\beta} + \sum_{i=1}^n (\tilde{W}_i^T \nu^q S^q^{-1} \tilde{W}_i + V_i^T F_i V_i) \right\}^{-1}$
   
   - $\mu^q_{\beta} \leftarrow \mu^q_{\beta} + \Sigma^q_{\beta} \left[ \sum_{i=1}^n \left\{ V_i^T (y_i - g_i) + \tilde{W}_i^T \nu^q S^{-1} (\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_{\beta}) \right\} - \Sigma^{-1}_{\beta} \mu^q_{\beta} \right]$ 
   
   - $S^q \leftarrow S + \sum_{i=1}^n \left\{ (\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_{\beta})(\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_{\beta})^T + \Sigma^q_{\tilde{\alpha}_i} + \tilde{W}_i \Sigma^q_{\beta} \tilde{W}_i^T \right\}$

   until absolute relative change in lower bound $\mathcal{L}$ is negligible
Prior-likelihood conflict diagnostics

- Diagnostic tests for identifying **divergent units**
  - e.g. identify hospitals divergent in quality of care provided

- **By-product of NCVMP**
  - Separate messages from above and below a node in a hierarchical model. “Mixed messages” indicate conflict
  - “Mixed messages” diagnostics approximate conflict diagnostics of Marshall and Spiegelhalter (2007)

\[
p(y_i | \beta, \tilde{\alpha}_i) \quad p(\tilde{\alpha}_i | \beta, D) \quad p(\tilde{\alpha}_i | \beta, D) \quad p(D | \nu, S)
\]

\[
p(\alpha_i | \beta, D) ~ p(y_i | \beta, \alpha_i) ~ p(\beta | \Sigma_{\beta})
\]

**Figure**: Factor graph for GLMM
Simulation based approach (Marshall and Spiegelhalter, 2007)

- Identify units that do not appear to be drawn from assumed random effects distributions
  - Compare replicates of $\tilde{\alpha}_i$ from its likelihood and predictive prior
  - Independent evidence about $\tilde{\alpha}_i$: conflict suggests model discrepancies

1. Generate **predictive prior replicate** $\tilde{\alpha}_i^{\text{rep}}$ from

   \[
   p_r(\tilde{\alpha}_i|y_{-i}) = \int p(\tilde{\alpha}_i|\beta, D) p(\beta, D|y_{-i}) \ d\beta \ dD
   \]

   - generate $\beta^{\text{rep}}$ and $D^{\text{rep}}$ from $p(\beta, D|y_{-i})$ using MCMC
   - simulate $\tilde{\alpha}_i^{\text{rep}}|\beta^{\text{rep}}, D^{\text{rep}}$

2. Generate **likelihood replicate** $\tilde{\alpha}_i^{\text{lik}} \sim p(\tilde{\alpha}_i|y_i)$ using only $y_i$ and a non-informative prior $p(\tilde{\alpha}_i)$ for $\tilde{\alpha}_i$ (Jeffreys’s prior)
Cross-validatory conflict p-values

- **Nuisance parameter** $\beta$: $p(\tilde{\alpha}_i|y_i) \propto p(\tilde{\alpha}_i) \int p(y_i|\beta, \tilde{\alpha}_i) p(\beta|\tilde{\alpha}_i) d\beta$ and $\beta$ is not estimable from unit $i$. Generate $\tilde{\alpha}_i^{\text{lik}}$ from

  $$p_i(\alpha_i|y) \propto p(\tilde{\alpha}_i) \int p(y_i|\tilde{\alpha}_i, \beta)p(\beta|y_{-i}) d\beta$$

- **Conflict p-values**: $\tilde{\alpha}_i^{\text{diff}} = \tilde{\alpha}_i^{\text{rep}} - \tilde{\alpha}_i^{\text{lik}}$
  - $\tilde{\alpha}_i^{\text{diff}}$ is scalar: lower tail: $p_{i,\text{con}}^l = P(\tilde{\alpha}_i^{\text{diff}} \leq 0|y)$,
    upper tail: $p_{i,\text{con}}^u = 1 - p_{i,\text{con}}^l$, two-sided: $2 \times \min(p_{i,\text{con}}^l, p_{i,\text{con}}^u)$
  - $\tilde{\alpha}_i^{\text{diff}}$ is $r \times 1$ vector: $\Delta = \mathbb{E}(\tilde{\alpha}_i^{\text{diff}}|y)^T \text{Cov}(\tilde{\alpha}_i^{\text{diff}}|y)^{-1} \mathbb{E}(\tilde{\alpha}_i^{\text{diff}}|y)$
    Assume $\tilde{\alpha}_i^{\text{diff}}$ is multivariate normal: $P(\chi_r^2 > \Delta)$ for testing $\tilde{\alpha}_i^{\text{diff}} = 0$

- **Simulation based full-data approach**: Simulate $\tilde{\alpha}_i^{\text{rep}}|\beta^{\text{rep}}, D^{\text{rep}}$ using $\beta^{\text{rep}}, D^{\text{rep}}$ generated from $p(\beta, D|y)$, without leaving out $y_i$
Conflict $p$-values from NCVMP

Update for $\lambda\tilde{\alpha}_i$

$$\mathcal{V}\tilde{\alpha}_i(\lambda\tilde{\alpha}_i)^{-1}\nabla\lambda\tilde{\alpha}_i E_q\{\log p(\tilde{\alpha}_i|\beta, D)\} + \mathcal{V}\tilde{\alpha}_i(\lambda\tilde{\alpha}_i)^{-1}\nabla\lambda\tilde{\alpha}_i E_q\{\log p(y_i|\tilde{\alpha}_i, \beta)\}$$

Message from prior:
Natural parameter of Gaussian approximation to $p_r(\tilde{\alpha}_i|y_{-i})$
$$\Rightarrow \tilde{\alpha}_i^{\text{rep}} \sim N(\mu_{\text{rep}}, \Sigma_{\text{rep}})$$
$$(\Sigma_{\text{rep}} = S^q/\nu^q, \mu_{\text{rep}} = \tilde{W}_i\mu^q_{\beta})$$

Message from likelihood:
Natural parameter of Gaussian approximation to $p_l(\tilde{\alpha}_i|y)$
$$\Rightarrow \tilde{\alpha}_i^{\text{lik}} \sim N(\mu_{\text{lik}}, \Sigma_{\text{lik}})$$
$$(\Sigma_{\text{lik}}^{-1} = Z_i^TF_iZ_i, \mu_{\text{lik}} = \mu^q_{\tilde{\alpha}_i} + \Sigma_{\text{lik}}Z_i^T(y_i - g_i))$$

- Assume $\tilde{\alpha}_i^{\text{rep}}$ and $\tilde{\alpha}_i^{\text{lik}}$ are independent,
  $$\tilde{\alpha}_i^{\text{diff}} \sim N(\mu_{\text{rep}} - \mu_{\text{lik}}, \Sigma_{\text{rep}} + \Sigma_{\text{lik}})$$
  (dependence increasingly weak as number of clusters increases)

- Compute conflict $p$-values after NCVMP algorithm has converged
NCVMP conflict diagnostics

Compare accuracy of different approaches

\[
\text{Mean abs difference in } z\text{-scores} = \frac{1}{n} \sum_{i=1}^{n} \left| \Phi^{-1}(p_{i,\text{con}}^{\text{CV}}) - \Phi^{-1}(p_{i,\text{con}}^{\text{method}}) \right|
\]

- Reflect importance of good agreement at the extremes
- \( p_{i,\text{con}}^{\text{CV}} \): cross-validatory conflict \( p \)-values
- \( p_{i,\text{con}}^{\text{method}} \): conflict \( p \)-values from method being assessed

- Large data sets: NCVMP conflict diagnostics an attractive alternative to simulation based MCMC approach
- Screening tool: study clusters flagged as divergent more closely, recompute conflict \( p \)-values by Monte Carlo
Bristol Inquiry data


- \( y_{ij} \sim \text{Bernoulli}(\pi_i) \) where \( y_{ij} = 1 \) if patient \( j \) at hospital \( i \) died and 0 otherwise. \( Y_i = \sum_{j=1}^{n_i} y_{ij} \) (number of deaths at hospital \( i \))

\[
\text{logit}(\pi_i) = \beta + u_i \quad \text{where} \quad u_i \sim N(0, D) \quad i = 1, \ldots, 12
\]

- Cross-validatory approach (remove each hospital in turn):
  - Excess mortality of concern. Upper-tail: \( p_{i,\text{con}} = P(\pi_i^{\text{rep}} \geq \pi_i^{\text{lik}}) \)
  - MCMC (OpenBUGS): \( p \)-values based on 100,000 simulations. Total time = 5 s \( \times \) 12 = 60 s
Bristol Inquiry data

<table>
<thead>
<tr>
<th></th>
<th>noncentered</th>
<th>centered</th>
<th>partially noncentered</th>
<th>MCMC (full-data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower bound ($\mathcal{L}$)</td>
<td>-1213.7</td>
<td>-1213.0</td>
<td>-1212.9</td>
<td>–</td>
</tr>
<tr>
<td>Time (fit model)</td>
<td>7.6</td>
<td>3.7</td>
<td>3.8</td>
<td>5</td>
</tr>
<tr>
<td>Time (compute conflict $p$-values)</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>14.4</td>
</tr>
<tr>
<td>Mean abs difference in $z$-scores</td>
<td>0.087</td>
<td>0.086</td>
<td>0.083</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Partial noncentering attained highest lower bound, was quick to converge and gave posterior approximations very close to that of MCMC.

NCVMP does better than simulation based full-data approach in terms of $z$-scores and computation time.

NCVMP: order of magnitude faster than cross-validatory approach.

Figure: Marginal posteriors estimated by MCMC (black) and NCVMP via centering (green), noncentering (blue) and partial noncentering (red)
Bristol Inquiry data

<table>
<thead>
<tr>
<th>hospital</th>
<th>$p_{i,\text{con}}^{\text{CV}}$</th>
<th>$p_{i,\text{con}}^{\text{NCVMP}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>2</td>
<td>0.436</td>
<td>0.450</td>
</tr>
<tr>
<td>3</td>
<td>0.935</td>
<td>0.928</td>
</tr>
<tr>
<td>4</td>
<td>0.125</td>
<td>0.138</td>
</tr>
<tr>
<td>5</td>
<td>0.298</td>
<td>0.311</td>
</tr>
<tr>
<td>6</td>
<td>0.720</td>
<td>0.725</td>
</tr>
<tr>
<td>7</td>
<td>0.737</td>
<td>0.745</td>
</tr>
<tr>
<td>8</td>
<td>0.661</td>
<td>0.667</td>
</tr>
<tr>
<td>9</td>
<td>0.440</td>
<td>0.453</td>
</tr>
<tr>
<td>10</td>
<td>0.380</td>
<td>0.390</td>
</tr>
<tr>
<td>11</td>
<td>0.763</td>
<td>0.764</td>
</tr>
<tr>
<td>12</td>
<td>0.721</td>
<td>0.727</td>
</tr>
</tbody>
</table>

**Figure:** Cross-validatory conflict $p$-values ($p_{i,\text{con}}^{\text{CV}}$) and conflict $p$-values from NCVMP ($p_{i,\text{con}}^{\text{NCVMP}}$) via partial noncentering

- Very good agreement between two sets of $p$-values. Both approaches suggest hospital 1 (Bristol) is discrepant.
Epilepsy data (Thall and Vail, 1990)

Clinical trial of 59 patients with epilepsy. Response: Number of seizures during two weeks before each of four clinic visits

Covariates:
- Visit ($\text{Visit}_1 = -0.3$, $\text{Visit}_2 = -0.1$, $\text{Visit}_3 = 0.1$ and $\text{Visit}_4 = 0.3$)
- Trt=1 (new anti-epileptic drug) or Trt=0 (placebo)
- Base (log of $\frac{1}{4}$ the number of baseline seizures)
- Age (log of patient age, centered to improve mixing in MCMC)

Model II (random intercept and slope model):

\[
\log \mu_{ij} = \beta_0 + \beta_1 \text{Base}_i + \beta_2 \text{Trt}_i + \beta_3 \text{Base}_i \times \text{Trt}_i + \beta_4 \text{Age}_i; \\
+ \beta_5 \text{Visit}_{ij} + u_{1i} + u_{2i} \text{Visit}_{ij}, [u_{1i} u_{2i}] \sim N \left(0, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix} \right)
\]

Model I: random slope dropped from Model II
Epilepsy data

- Examine suitability of random effects distribution (two-sided p-values)
- Cross-validatory approach (each patient removed in turn):
  - MCMC: p-values based on 50,000 simulations. Total time = $61 \times 59 = 3599$ s (Model I), $54 \times 59 = 3186$ s (Model II)

![Graphs showing marginal posteriors estimated by MCMC (black) and NCVMP via centering (green), noncentering (blue) and partial noncentering (red) for different variables.]

**Figure**: Model I: Marginal posteriors estimated by MCMC (black) and NCVMP via centering (green), noncentering (blue) and partial noncentering (red)
### Epilepsy data

<table>
<thead>
<tr>
<th></th>
<th>noncentered</th>
<th>centered</th>
<th>partially noncentered</th>
<th>MCMC (full-data)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model I</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower bounds ($\mathcal{L}$)</td>
<td>-707.0</td>
<td>-701.5</td>
<td>-701.1</td>
<td>–</td>
</tr>
<tr>
<td>Time (model fitting)</td>
<td>1.4</td>
<td>0.2</td>
<td>0.2</td>
<td>62</td>
</tr>
<tr>
<td>Time (compute conflict $p$-values)</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>4278.2</td>
</tr>
<tr>
<td>Mean abs difference in $z$-scores</td>
<td>0.167</td>
<td>0.159</td>
<td>0.155</td>
<td>0.103</td>
</tr>
<tr>
<td><strong>Model II</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower bounds ($\mathcal{L}$)</td>
<td>-701.4</td>
<td>-696.1</td>
<td>-695.3</td>
<td>–</td>
</tr>
<tr>
<td>Time (model fitting)</td>
<td>1.3</td>
<td>0.5</td>
<td>0.5</td>
<td>55</td>
</tr>
<tr>
<td>Time (compute conflict $p$-values)</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td>3109.6</td>
</tr>
<tr>
<td>Mean abs difference in $z$-scores</td>
<td>0.105</td>
<td>0.107</td>
<td>0.101</td>
<td>0.116</td>
</tr>
</tbody>
</table>

- Good agreement between cross-validatory and NCVMP
- NCVMP compares well in terms of $z$-scores and is faster than both simulation based approaches by an order of magnitude

*Figure: NCVMP via partial noncentering*
Conclusion

- Demonstrated how GLMMs can be fitted using NCVMP\(^1\)
- Showed that prior-likelihood conflict diagnostics can be obtained as a by-product of NCVMP\(^2\)
- Showed that partial noncentering can accelerate convergence of VB and produce more accurate posterior approximations

---


Justification (Message from prior)

- Large data sets: \( p(\beta, D|y_{-i}) \) is close to \( p(\beta, D|y) \)

\[
pr(\tilde{\alpha}_i|y_{-i}) = \int p(\tilde{\alpha}_i|\beta, D) p(\beta, D|y_{-i}) \, d\beta \, dD \\
\approx \int p(\tilde{\alpha}_i|\beta, D) q(\beta|\lambda_\beta)q(D|\lambda_D) \, d\beta \, dD \\
\geq \exp[E_{\tilde{\alpha}_i}\{\log p(\tilde{\alpha}_i|\beta, D)\}] \quad \text{(Jensen’s inequality)}
\]

\[
pr(\tilde{\alpha}_i|y_{-i}) \propto \exp[E_{\tilde{\alpha}_i}\{\log p(\tilde{\alpha}_i|\beta, D)\}] \\
\Rightarrow \tilde{\alpha}_i^{\text{rep}} \sim N(\tilde{W}_i\mu_{q(\beta)}, S_{q(D)}/\nu_{q(D)})
\]

- This is what we get if we interpret the first message as being natural parameter of a Gaussian approximation to \( p_r(\tilde{\alpha}_i|y_{-i}) \)
Justification (Message from likelihood)

- Think of \( p(\tilde{\alpha}_i|y_{-i}) \) as ‘prior’ to be updated when \( y_i \) is available:

\[
P(\tilde{\alpha}_i|y) \propto p(\tilde{\alpha}_i|y_{-i})p(y_i|\tilde{\alpha}_i, y_{-i})
\]

\[
\Rightarrow p(y_i|\tilde{\alpha}_i, y_{-i}) \propto \frac{p(\tilde{\alpha}_i|y)}{p(\tilde{\alpha}_i|y_{-i})}
\]

\[
\Rightarrow p(y_i|\tilde{\alpha}_i, y_{-i}) \propto \exp\left\{-\frac{1}{2}(\tilde{\alpha}_i - \mu^{q \tilde{\alpha}_i})^T \Sigma_{\tilde{\alpha}_i}^{-1}(\tilde{\alpha}_i - \mu^{q \tilde{\alpha}_i})\right\}
\]

\[
\exp\left\{-\frac{1}{2}(\tilde{\alpha}_i - \mu_{rep})^T \Sigma_{rep}^{-1}(\tilde{\alpha}_i - \mu_{rep})\right\}
\]

\[
= N(\mu_{lik}, \Sigma_{lik})
\]

- As

\[
p(y_i|\tilde{\alpha}_i, y_{-i}) = \int p(y_i|\beta, \tilde{\alpha}_i)p(\beta|\tilde{\alpha}_i, y_{-i}) \, d\beta
\]

and \( p(\beta|\tilde{\alpha}_i, y_{-i}) \) is close to \( p(\beta|y_{-i}) \) when number of clusters is large, second message can be considered as natural parameter of a Gaussian approximation to \( p_l(\tilde{\alpha}_i|y) \) if we assume uniform prior for \( p(\tilde{\alpha}_i) \)