A stochastic variational framework for fitting and diagnosing generalized linear mixed models

Linda Tan
(Joint work with David Nott)
National University of Singapore

Fourth Singapore Conference on Statistical Science (7 Feb 2014)
Outline

1. Introduction
   - Variational approximation methods

2. Generalized linear mixed models

3. Stochastic variational inference

4. Prior-likelihood conflict diagnostics

5. Examples

6. Conclusion
Introduction

- **Generalized linear mixed models (GLMMs)**
  - Extend generalized linear models by including random effects to account for correlated observations in grouped data
  - Have wide applications

- **Estimation of GLMMs is challenging**
  - Integral over random effects is intractable
  - Numerical quadrature and MCMC are computationally intensive
  - Approximate methods: penalized quasi-likelihood, Laplace approximation, Gaussian variational approximation, ...
Proposed methods

- **Fit GLMMs: nonconjugate variational message passing (NCVMP)**
  - An algorithm developed in machine learning extending variational Bayes to nonconjugate models (Knowles and Minka, 2011)

- **Accelerate convergence in NCVMP for moderately large data sets using stochastic variational inference**
  - Extend stochastic variational inference for conjugate-exponential models (Hoffman et al., 2013) to nonconjugate models

- **Derive prior-likelihood conflict diagnostics from NCVMP**
  - Alternative to simulation-based MCMC methods

- **Consider partially noncentered parametrization for GLMMs**
  - Partial centering has been used in hierarchical models to boost efficiency in MCMC algorithms (Papaspiliopoulos et al., 2003)
Variational Approximation Methods

- $y$: observed data, $\theta$: set of unknown parameters
- Bayesian inference based on posterior distribution $p(\theta|y)$ (often intractable)
- Approximate $p(\theta|y)$ by more tractable density function $q(\theta)$
- Restrictions on $q(\theta)$:
  - $q(\theta)$ belongs to some parametric distribution
  - $q(\theta) = \prod_{i=1}^{m} q_{i}(\theta_{i})$ for $\theta = \{\theta_{1}, \ldots, \theta_{m}\}$
- 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 ($L$)
- Kullback-Leibler divergence $\geq 0$

- Maximizing $L \iff$ minimizing Kullback-Leibler divergence
Variational Bayes (VB)

- Assume \( q(\theta) = \prod_{i=1}^{m} q_i(\theta_i) \) for \( \theta = \{\theta_1, \ldots, \theta_m\} \)
- Maximize lower bound \( \mathcal{L} \) w.r.t \( q_1, \ldots, q_m \):
  
  \[ \text{Optimal densities: } q_i(\theta_i) \propto \exp\{E_{-\theta_i} \log p(y, \theta)\} \quad \text{for each } i \]

  \[ E_{-\theta_i} : \text{ expectation w.r.t. } \prod_{j \neq i} q_j(\theta_j) \]

- Conjugate priors \( \Rightarrow \) density of optimal \( q_i \) is recognizable, suffices to update parameters of \( q_i \)

- **Variational message passing** (Winn and Bishop 2005)
  
  - general-purpose algorithm for applying VB to conjugate-exponential models
  
  - implemented in \textit{Infer.NET} (Microsoft Research)
Nonconjugate variational message passing (NCVMP)

- NCVMP (Knowles and Minka, 2011) assumes
  \[ q(\theta) = \prod_{i=1}^{m} q_i(\theta_i) \text{ for } \theta = \{\theta_1, \ldots, \theta_m\} \]
  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

- Updates: \( \nabla_{\lambda_i} \mathcal{L} = 0 \) when \( \mathcal{L} \) is maximized w.r.t. \( \lambda_i \)
  \[ \nabla_{\lambda_i} \mathcal{L} = \nabla_{\lambda_i} E_q\{\log p(y, \theta)\} - \mathcal{V}_i(\lambda_i)\lambda_i \]
  where \( \mathcal{V}_i(\lambda_i) \) denotes covariance matrix of \( t_i(\theta_i) \)
Nonconjugate variational message passing (NCVMP)

NCVMP Algorithm

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

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

until convergence

- NCVMP is a fixed-point iteration algorithm
- Lower bound not guaranteed to increase at each step
- Convergence problems may be encountered (fix using damping)
- NCVMP provides flexibility in how expectations are evaluated (use bounds or quadrature)
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 \( u_i \sim_{\text{iid}} N(0, D) \),
  \[
  y_{ij} | u_i \sim_{\text{iid}} \exp \{ y_{ij} \zeta_{ij} - b(\zeta_{ij}) + c(y_{ij}) \}
  \]
- \( \zeta_{ij} \): canonical parameter, \( b(\cdot) \), \( c(\cdot) \): functions specific to the exponential family
- \( \mu_{ij} = \mathbb{E}(y_{ij} | u_i) \) and \( g(\mu_{ij}) = \eta_{ij} \) for some link function \( g(\cdot) \), where
  \[
  \eta_{ij} = x_{ij}^T \beta + z_{ij}^T u_i
  \]
- \( x_{ij} \) and \( z_{ij} \): vectors of covariates
- \( \beta \): unknown fixed regression parameters.
Generalized linear mixed models

- For the $i$th cluster, let
  
  \[
  y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \eta_i = \begin{bmatrix} \eta_{i1} \\ \vdots \\ \eta_{in_i} \end{bmatrix}, \quad X_i = \begin{bmatrix} x_{i1}^T \\ \vdots \\ x_{in_i}^T \end{bmatrix}, \quad Z_i = \begin{bmatrix} z_{i1}^T \\ \vdots \\ z_{in_i}^T \end{bmatrix}.
  \]

- 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: $D \sim IW(\nu, S)$ based on a “minimally informative” prior guess of $D$ determined from first-stage data variability (Kass and Natarajan, 2006)
Idea of partial noncentering

- Introduce a tuning parameter via reparametrization of model and seek its optimal value for fastest convergence

- Linear Mixed model:

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

  - 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
  - Proportion of \( \beta \) subtracted from \( \alpha_i \) varies with \( i \) as each \( y_i \) carries different amount of information about underlying \( \alpha_i \)
  - Centered \((W_i = 0)\) and noncentered \((W_i = I_r)\)
  - Rewrite \( y_i = Z_iW_i\beta + Z_i\tilde{\alpha}_i + \epsilon_i \) where \( \tilde{\alpha}_i \sim N((I - W_i)\beta, D) \)
Partial noncentering in variational Bayes

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

**VB Algorithm: Linear mixed model**

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

**Cycle:**

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

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

**until convergence**
Partial noncentering in variational Bayes

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

- Partial noncentering yields more rapid convergence than centering or noncentering

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

- Assumption of a factorized posterior in VB tends to result 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
A partially noncentered parametrization for the GLMM

- Partition $X_i$ as $[Z_i \ X_{si} \ X_{gi}]$ and $\beta$ as $[\beta_z^T, \beta_s^T, \beta_g^T]^T$

- $X_{si} = 1_{n_i}x_{si}^T$ consist of 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 $\tilde{\alpha}_i = \alpha_i - W_i C_i \beta_c \sim N(\tilde{W}_i \beta, D)$ where $\tilde{W}_i = [(I_r - W_i) C_i \ 0]$

- Partially noncentered parametrization:

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

- Linear mixed model:

\[ W_i = (Z_i^T Q_i Z_i + D^{-1})^{-1} D^{-1} \quad \text{where} \quad Q_i = \frac{1}{\sigma^2} I_r \]

\[ = (I_f + D^{-1})^{-1} D^{-1} \]

where \( \ell = \log p(y_i | \beta, \alpha_i) \) and \( I_f = -\frac{\partial^2 \ell}{\partial \alpha_i \partial \alpha_i^T} \)

- Extend partial noncentering to logistic and Poisson GLMMs:

  Logistic: \( Q_i = \text{diag} \left( \frac{\exp(\eta_i)}{\{1+\exp(\eta_i)\}^2} \right) \)

  Poisson: \( Q_i \approx \text{diag}(y_i) \)

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

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

- \( \theta = \{ \beta, D, \tilde{\alpha} \} \). Consider

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

- Optimal \( q(D) \) is \( IW(\nu^q, S^q) \)

- For Bernoulli or Poisson responses, \( p(y_i|\beta, \tilde{\alpha}_i) \) is nonconjugate w.r.t. priors over \( \beta \) and \( \tilde{\alpha}_i \)

- Optimal densities for \( q(\beta) \) and \( q(\tilde{\alpha}_i) \) are not recognizable

- Apply NCVMP: assume \( q(\beta) \) is \( N(\mu_\beta^q, \Sigma_\beta^q) \), \( q(\tilde{\alpha}_i) \) is \( N(\mu_{\tilde{\alpha}_i}^q, \Sigma_{\tilde{\alpha}_i}^q) \)
  - 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 $i = 1, \ldots, n$
   - $\Sigma^q_{\tilde{\alpha}_i} \leftarrow (\nu^q S^q^{-1} + Z_i^T F_i Z_i)^{-1}$
   - $\mu^q_{\tilde{\alpha}_i} \leftarrow \mu^q_{\tilde{\alpha}_i} + \Sigma^q_{\tilde{\alpha}_i} \{ Z_i^T (y_i - G_i) - \nu^q S^q^{-1} (\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_\beta) \}$

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

until the absolute relative change in lower bound $L$ is negligible
Stochastic variational inference

- NCVMP algorithm for GLMMs iterates between
  - updating local variational parameters \((\mu^q_{\alpha_i}, \Sigma^q_{\alpha_i})\) for \(i = 1, \ldots, n\)
  - re-estimating global variational parameters \((\lambda_\beta \text{ and } \lambda_D): \text{natural parameters of } q(\beta) \text{ and } q(D)\)

- Inefficient for large data sets and impossible for streaming data

Stochastic variational inference (Hoffman et al. 2013)

At each iteration,

1. Randomly select mini-batch \(B\) of \(|B| \geq 1\) units from whole data
2. Optimize local variational parameters of units in mini-batch \(B\) (as a function of global variational parameters at their current setting)
3. Update global variational parameters using stochastic natural gradient ascent. Noisy gradients computed based on optimized local variational parameters of units in mini-batch \(B\)
Natural gradient of variational lower bound

- **Stochastic gradient ascent (global variational parameters):**
  \[ \lambda^{(t+1)} = \lambda^{(t)} + a_t \nabla_{\lambda} \mathcal{L}(\lambda^{(t)}) \]
  - Euclidean metric: direction of steepest ascent given by regular gradient \( \nabla_{\lambda} \mathcal{L}(\lambda^{(t)}) \) (noisy estimate used in stochastic version)

- **Motivation for natural gradients:**
  - Euclidean distance between \( \lambda \) and \( \lambda' \) often a poor measure of dissimilarity between \( q(\theta|\lambda) \) and \( q(\theta|\lambda') \)
  - More intuitive measure: symmetrized Kullback-Leibler divergence (invariant to parameter transformations)
  - Under this measure, direction of steepest ascent: natural gradient

- **Natural gradient** (Amari, 1998)
  \[ \nabla_{\lambda_i} \mathcal{L} = \mathcal{V}_i(\lambda_i)^{-1} \nabla_{\lambda_i} E_q\{\log p(y, \theta)\} - \lambda_i \]
  - NCVMP update
Unbiased estimates of natural gradients

Let $\lambda_{\tilde{\alpha}_i}^{\text{opt}}$ denote $\lambda_{\tilde{\alpha}_i}$ optimized as a function of $\lambda_\beta$ and $\lambda_D$

$$\tilde{\nabla}_{\lambda_\beta} L = \nabla_\beta (\lambda_\beta)^{-1} \nabla_\beta \left\{ E_q \{ \log p(\beta | \Sigma_\beta) \} \right\}$$

$$+ \sum_{i=1}^{n} \left[ E_q \{ \log p(y_i | \beta, \tilde{\alpha}_i) \} + E_q \{ \log p(\tilde{\alpha}_i | \beta, D) \} \right]_{\lambda_{\tilde{\alpha}_i}^{\text{opt}}} - \lambda_\beta$$

An unbiased estimate of $\tilde{\nabla}_{\lambda_\beta} L$ is $\hat{\lambda}_\beta - \lambda_\beta$ where

$$\hat{\lambda}_\beta = \nabla_\beta (\lambda_\beta)^{-1} \nabla_\beta \left\{ E_q \{ \log p(\beta | \Sigma_\beta) \} \right\}$$

$$+ \frac{n}{|B|} \sum_{i \in B} \left[ E_q \{ \log p(y_i | \beta, \tilde{\alpha}_i) \} + E_q \{ \log p(\tilde{\alpha}_i | \beta, D) \} \right]_{\lambda_{\tilde{\alpha}_i}^{\text{opt}}}$$

Similarly, an unbiased estimate of $\tilde{\nabla}_{\lambda_D} L$ is $\hat{\lambda}_D - \lambda_D$, where

$$\hat{\lambda}_D = \nabla_D (\lambda_D)^{-1} \nabla_D \left[ \frac{n}{|B|} \sum_{i \in B} E_q \{ \log p(\tilde{\alpha}_i | \beta, D) \} \right]_{\lambda_{\tilde{\alpha}_i}^{\text{opt}}} + E_q \{ \log p(D | \nu, B) \}$$
Stochastic natural gradient approximation

\[
\lambda^{(t+1)}_\beta = \lambda^{(t)}_\beta + a_t (\hat{\lambda}_\beta - \lambda^{(t)}_\beta) = (1 - a_t) \lambda^{(t)}_\beta + a_t \hat{\lambda}_\beta
\]

\[
\lambda^{(t+1)}_D = \lambda^{(t)}_D + a_t (\hat{\lambda}_D - \lambda^{(t)}_D) = (1 - a_t) \lambda^{(t)}_D + a_t \hat{\lambda}_D
\]

where \(a_t\) should satisfy

\[
a_t \to 0, \quad \sum_{t=0}^{\infty} a_t = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} a^2_t < \infty
\]

- When \(a_t = 1\) and \(B\) is whole data set, \(\lambda^{(t+1)}_\beta\) is NCVMP update
- NCVMP is a natural gradient method with step size 1
Stochastic NCVMP algorithm for GLMMs

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

For $t = 0, 1, 2, \ldots$

1. Randomly select mini-batch $B$ of $|B| \geq 1$ units from whole data

2. Update local variational parameters for $i \in B$ repeatedly using:
   \[
   \begin{align*}
   \Sigma^q_{\tilde{\alpha}_i} &\leftarrow (\nu^q S^q^{-1} + Z_i^T F_i Z_i)^{-1} \\
   \mu^q_{\tilde{\alpha}_i} &\leftarrow \mu^q_{\tilde{\alpha}_i} + \Sigma^q_{\tilde{\alpha}_i} \{Z_i^T(y_i - g_i) - \nu^q S^q^{-1}(\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_\beta)\}
   \end{align*}
   \]
   until convergence is reached

3. Update global variational parameters:
   \[
   \begin{align*}
   \Sigma^q_\beta &\leftarrow \left[(1 - a_t) \Sigma^q_\beta^{-1} + a_t \left\{ \Sigma^{-1}_\beta + \frac{n}{|B|} \sum_{i \in B} (\nu^q \tilde{W}_i^T S^q^{-1} \tilde{W}_i + V_i^T F_i V_i) \right\} \right]^{-1} \\
   \mu^q_\beta &\leftarrow \mu^q_\beta + a_t \Sigma^q_\beta \left[ \frac{n}{|B|} \sum_{i \in B} \{\nu^q \tilde{W}_i^T S^q^{-1}(\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_\beta) + V_i^T(y_i - g_i)\} - \Sigma^{-1}_\beta \mu^q_\beta \right] \\
   S^q &\leftarrow (1 - a_t) S^q + a_t \left[ \frac{n}{|B|} \sum_{i \in B} \{ (\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_\beta)(\mu^q_{\tilde{\alpha}_i} - \tilde{W}_i \mu^q_\beta)^T + \Sigma^q_{\tilde{\alpha}_i} + \tilde{W}_i \Sigma^q_\beta \tilde{W}_i^T \} + S \right]
   \end{align*}
   \]
Implementation

- **Initialization**: Penalized quasi-likelihood or fit from generalized linear model (pool all data and set random effects to zero)

- **Mini-batches**: Random partitioning of data set

- **Convergence of local variational parameters**: 
  \[ \frac{\|\mu^B_{q(\tilde{\alpha})}(t) - \mu^B_{q(\tilde{\alpha})}(t-1)\|}{\|\mu^B_{q(\tilde{\alpha})}(t)\|} < 0.05 \text{ where } \mu^B_{q(\tilde{\alpha})} = \begin{bmatrix} \mu_{\tilde{\alpha}_j}^T, \ldots, \mu_{\tilde{\alpha}_{j|B|}}^T \end{bmatrix} \]

- **Moderately large data sets**: 
  - Switch from stochastic to standard NCVMP when relative increase in \(\mathcal{L}\) after a sweep through the data is less than \(10^{-3}\)
  - Terminate standard NCVMP when relative increase in \(\mathcal{L}\) < \(10^{-6}\)

- **Gain sequence**: 
  - \(M\): number of partitions of data set. Step size: \(a_t = M/(t + A)\)
  - Investigate performance of stability constants \(A\) for various mini-batch sizes
Prior-likelihood conflict diagnostics

Consider diagnostic tests for identification of divergent units
  e.g. identify hospitals which are divergent in terms of quality of care provided or choice of surgical procedure for treating a cancer

Prior-likelihood conflict diagnostics as by-product of NCVMP

Idea: Separate messages coming from above and below a node in a hierarchical model. “Mixed messages” indicate conflict

Our “mixed messages” diagnostics approximate the conflict diagnostics of Marshall and Spiegelhalter (2007)
GLMM: $\eta_i = V_i \beta + Z_i \tilde{\alpha}_i$, $\tilde{\alpha}_i \sim N(\tilde{W}_i \beta, D)$ for $i = 1, \ldots, n$

To 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

Prior and likelihood replications represent independent sources of evidence about $\tilde{\alpha}_i$ and conflict suggests discrepancies in model

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

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

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

Generate **likelihood replicate** $\tilde{\alpha}_i^{lik} \sim p(\tilde{\alpha}_i | y_i)$ 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: let $\tilde{\alpha}_i^{\text{diff}} = \tilde{\alpha}_i^{\text{rep}} - \tilde{\alpha}_i^{\text{lik}}$

  - $\tilde{\alpha}_i^{\text{diff}}$ is scalar: lower tail: $p_i^{l,\text{con}} = P(\tilde{\alpha}_i^{\text{diff}} \leq 0|y)$, upper tail: $p_i^{u,\text{con}} = 1 - p_i^{l,\text{con}}$, two-sided: $2 \times \min(p_i^{l,\text{con}}, p_i^{u,\text{con}})$

  - $\tilde{\alpha}_i^{\text{diff}}$ is $r \times 1$ vector:

    $$\Delta = E(\tilde{\alpha}_i^{\text{diff}}|y)^T \text{Cov}(\tilde{\alpha}_i^{\text{diff}}|y)^{-1} E(\tilde{\alpha}_i^{\text{diff}}|y)$$

    Assume $\tilde{\alpha}_i^{\text{diff}}$ is multivariate normal, conflict $p$-value: $P(\chi_r^2 > \Delta)$
    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 $\lambda\tilde{\alpha}_i$:

$\nabla_{\lambda\tilde{\alpha}_i} E_q\{\log p(\tilde{\alpha}_i|\beta, D)\} + \nabla_{\lambda\tilde{\alpha}_i} E_q\{\log p(y_i|\tilde{\alpha}_i, \beta)\}$

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

2. **Second term: message from likelihood** $p(y_i|\tilde{\alpha}_i, \beta)$
   - Natural parameter of Gaussian approximation to $p_l(\tilde{\alpha}_i|y) \Rightarrow \tilde{\alpha}_{i}^{\text{lik}} \sim N(\mu_{\text{lik}}, \Sigma_{\text{lik}})$ ($\Sigma_{\text{lik}}^{-1} = Z_i^TF_iZ_i$, $\mu_{\text{lik}} = \mu^q_{\tilde{\alpha}_i} + \Sigma_{\text{lik}}Z_i^T(y_i - g_i)$)

- Assuming $\tilde{\alpha}_{i}^{\text{rep}}$ 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)

- Conflict $p$-values can be calculated when NCVMP algorithm converges to identify divergent units
Remarks

Compare accuracy of different approaches

Mean abs difference in z-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 (gold-standard)
- \( p_{i,\text{con}}^{\text{method}} \): conflict \( p \)-values computed from method being assessed

- Large data sets: computing conflict diagnostics within NCVMP an attractive alternative to simulation based MCMC approach
- NCVMP as a screening tool: clusters flagged as divergent can be studied more closely, conflict \( p \)-values recomputed by Monte Carlo
Bristol Inquiry data

- Consider mortality rates in open surgeries for 12 hospitals (Bristol: hospital 1) for children under 1 year old, from 1991 to 1995

\[ y_{ij} \sim \text{Bernoulli}(\pi_i) \text{ where } y_{ij} = 1 \text{ if patient } j \text{ at hospital } i \text{ died and } 0 \text{ otherwise. } \]

\[ Y_i = \sum_{j=1}^{n_i} y_{ij} \text{ (number of deaths at hospital } i) \]

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

- Cross-validatory approach (remove each hospital in turn):
  - Generate \( \beta^\text{rep}, D^\text{rep} | y_{-i} \) using MCMC, simulate \( \pi_i^\text{rep} | \beta^\text{rep}, D^\text{rep} \)
  - \( \pi_i^\text{lik} \sim p(\pi_i | y_i) = \text{Beta}(Y_i + 0.5, n_i - Y_i + 0.5) \) (\( \pi_i \): Jeffreys’s prior)
  - Excess mortality is of concern. Upper-tail: \( p_{i,\text{con}} = P(\pi_i^\text{rep} \geq \pi_i^\text{lik}) \)
  - MCMC (OpenBUGS): two chains (51,000 iterations each, first 1000 discarded as burn-in), 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 (model fitting)</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 is of an 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

- Plot indicates 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 \( (\text{Visit}_1 = -0.3, \text{Visit}_2 = -0.1, \text{Visit}_3 = 0.1 \text{ and } \text{Visit}_4 = 0.3) \)
  - 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):**
  \[
  \log \mu_{ij} = \beta_0 + \beta_1 \text{Base}_i + \beta_2 \text{Trt}_i + \beta_3 \text{Base}_i \times \text{Trt}_i + \beta_4 \text{Age}_i \\
  + \beta_5 \text{Visit}_{ij} + u_{1i} + u_{2i} \text{Visit}_{ij}, [u_{1i}, u_{2i}] \sim N \left(0, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix} \right)
  \]

- **Model I:** random slope dropped from Model II

- Examine suitability of random effects distribution (two-sided conflict \( p \)-values)
Epilepsy data

- Simulation based approaches: nuisance parameter $\beta_5$ in Model I
- Cross-validatory approach (each patient removed in turn):
  - MCMC: two chains (26,000 iterations each, first 1000 discarded as burn-in), $p$-values based on 50,000 simulations. Total time = $61 \times 59 = 3599$ s (Model I), $54 \times 59 = 3186$ s (Model II)
  - Likelihood replicates: assume Jeffreys’s prior for $\alpha_i$, adaptive rejection metropolis sampling via HI (Petris and Tardella, 2003)

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>Model</th>
<th>Lower bounds ($\mathcal{L}$)</th>
<th>noncentered</th>
<th>centered</th>
<th>partially noncentered</th>
<th>MCMC (full-data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>Time (model fitting)</td>
<td>1.4</td>
<td>0.2</td>
<td>0.2</td>
<td>62</td>
</tr>
<tr>
<td></td>
<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></td>
<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>Model II</td>
<td>Lower bounds ($\mathcal{L}$)</td>
<td>-701.4</td>
<td>-696.1</td>
<td>-695.3</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Time (model fitting)</td>
<td>1.3</td>
<td>0.5</td>
<td>0.5</td>
<td>55</td>
</tr>
<tr>
<td></td>
<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></td>
<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
Epilepsy data

<table>
<thead>
<tr>
<th>Patient</th>
<th>$p_{i,con}^{CV}$</th>
<th>$p_{i,con}^{NCVMP}$</th>
<th>Patient</th>
<th>$p_{i,con}^{CV}$</th>
<th>$p_{i,con}^{NCVMP}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.047</td>
<td>0.056</td>
<td>10</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>25</td>
<td>0.048</td>
<td>0.062</td>
<td>25</td>
<td>0.024</td>
<td>0.049</td>
</tr>
<tr>
<td>35</td>
<td>0.038</td>
<td>0.044</td>
<td>56</td>
<td>0.038</td>
<td>0.051</td>
</tr>
<tr>
<td>56</td>
<td>0.023</td>
<td>0.028</td>
<td>58</td>
<td>0.002</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table: Conflict $p$-values for outliers at the 0.05 level in models I and II using cross-validatory approach and NCVMP via partial noncentering

- $p$-values from two approaches are close
- Some outliers identified by cross-validatory approach were not detected using NCVMP
- NCVMP as screening tool (very useful for large data sets): flag all patients with conflict $p$-values $< 0.1$ as possible outliers. Recompute conflict $p$-values for this smaller group using cross-validatory approach
Skin cancer study (Greenberg et al., 1989)

- Consider \( n = 1683 \) subjects
- \( y_{ij} \): number of new skin cancers in year \( j \) for \( i \)th subject
- **Covariates for \( i \)th subject**
  - Age\(_i\), age in years at the beginning of the study
  - Gender\(_i\) = 1 if male and 0 if female
  - Skin\(_i\) = 1 if skin has burns and 0 otherwise
  - Exposure\(_i\), a count of the number of previous skin cancers
  - Year\(_{ij}\), year of follow-up

- **Random intercept and slope model**

\[
\log(\mu_{ij}) = \beta_0 + \beta_1 \text{Year}_{ij} + \beta_2 \text{Age}_i + \beta_3 \text{Gender}_i + \beta_4 \text{Skin}_i + \beta_5 \text{Exposure}_i \\
+ u_1i + u_2i \text{Year}_{ij}, \begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N \left( 0, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix} \right), 1 \leq j \leq 5
\]

- MCMC: two chains (11,000 iterations each, first 1000 discarded as burn-in)
## Skin cancer study

<table>
<thead>
<tr>
<th></th>
<th>noncentered</th>
<th>centered</th>
<th>partially noncentered</th>
<th>MCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower bound ($\mathcal{L}$)</td>
<td>-4054.1</td>
<td>-4054.1</td>
<td>-4051.7</td>
<td>-</td>
</tr>
<tr>
<td>Time (model fitting)</td>
<td>46.6</td>
<td>42.6</td>
<td>42.0</td>
<td>11113</td>
</tr>
</tbody>
</table>

**Figure**: Marginal posteriors estimated by MCMC (black) and NCVMP via centering (green), noncentering (blue) and partial noncentering (red)
Skin cancer study

- Simulate larger data set with \( n = 1683 \times 15 = 25245 \)
- Replicate design matrices for each cluster 15 times
- Generate responses by using variational posterior means from model fitted using NCVMP (via partial noncentering) as parameters
- Simulated data: NCVMP (partial noncentering) converged in 1230.9 s
- For each setting, ten runs of stochastic NCVMP were performed
- Computing times reduced by factor of \( \geq 2 \) across wide range of settings

Figure: Computation times (s) corresponding to different stability constants A
Skin cancer study

| $|B|$ | 63 | 126 | 252 | 504 |
|-----|----|----|----|-----|
| $A$ | 8  | 4  | 2  | 0   |
| time| 266.3 | 224.4 | 205.3 | 200.8 |

Table: Shortest average time to convergence (s) for different mini-batch sizes together with corresponding stability constant $A$

Figure: Plot of $\log(-57957 - \mathcal{L})$ against time for mini-batch size 504 ($A = 0$) fitted using stochastic NCVMP and whole data set fitted using standard NCVMP

- Larger stability constants $A$ preferred for smaller mini-batch-sizes
- Shortest average time to convergence of 200.8 s represents reduction in computation time by factor of 6.
- Similar results can be achieved by smaller mini-batch sizes with appropriate step sizes
- Stochastic NCVMP able to make much bigger gains than standard NCVMP in first few sweeps
Conclusion

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

---
