Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Fast and Accurate Binary Response Mixed Model Analysis via Expectation Propagation

P. Hall, I.M. Johnstone, J.T. Ormerod, M.P. Wand & J.C.F. Yu

To cite this article: P. Hall, I.M. Johnstone, J.T. Ormerod, M.P. Wand & J.C.F. Yu (2019): Fast and Accurate Binary Response Mixed Model Analysis via Expectation Propagation, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1665529

To link to this article: https://doi.org/10.1080/01621459.2019.1665529

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–15: Theory and Methods
https://doi.org/10.1080/01621459.2019.1665529

Taylor & Francis
Taylor & Francis Group

Fast and Accurate Binary Response Mixed Model Analysis Via Expectation Propagation

P. Hall[a], I. M. Johnstone[b], J. T. Ormerod[c], M. P. Wand[d], and J. C. F. Yu[d]

[a]School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia; [b]Department of Statistics, Stanford University, Stanford, CA; [c]School of Mathematics and Statistics, University of Sydney, Sydney, Australia; [d]School of Mathematical and Physical Sciences, University of Technology Sydney, Ultimo, Australia

ABSTRACT

Expectation propagation is a general prescription for approximation of integrals in statistical inference problems. Its literature is mainly concerned with Bayesian inference scenarios. However, expectation propagation can also be used to approximate integrals arising in *frequentist* statistical inference. We focus on likelihood-based inference for binary response mixed models and show that fast and accurate quadrature-free inference can be realized for the probit link case with multivariate random effects and higher levels of nesting. The approach is supported by asymptotic calculations in which expectation propagation is seen to provide consistent estimation of the exact likelihood surface. Numerical studies reveal the availability of fast, highly accurate and scalable methodology for binary mixed model analysis. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2018
Revised August 2019

KEYWORDS
Best prediction; Generalized linear mixed models; Kullback–Leibler projection; Maximum likelihood; Message passing; Quasi-Newton methods; Scalable statistical methodology.

1. Introduction

Binary response mixed model-based data analysis is ubiquitous in many areas of application, with examples such as analysis of biomedical longitudinal data (e.g., Diggle et al. 2002), social science multilevel data (e.g., Goldstein 2010), small area survey data (e.g., Rao and Molina 2015), and economic panel data (e.g., Baltagi 2013). The standard approach for likelihood-based inference in the presence of multivariate random effects is Laplace approximation, which is well known to be inconsistent and prone to inferential inaccuracy. Our main contribution is to overcome this problem using expectation propagation. The new approach possesses speed and scalability on par with that of Laplace approximation, but is provably consistent and demonstrably very accurate. Bayesian approaches and Monte Carlo methods offer another route to accurate inference for binary response mixed models (e.g., Gelman and Hill 2007). However, speed and scalability issues aside, frequentist inference is the dominant approach in many areas in which mixed models are used. Henceforth, we focus on frequentist binary mixed model analysis.

The main obstacle for likelihood-based inference for binary mixed models is the presence of irreducible integrals. For grouped data with one level of nesting, the dimension of the integrals matches the number of random effects. The two most common approaches to dealing with these integrals are (1) quadrature and (2) Laplace approximation. For example, in the R computing environment (R Core Team 2019) the function `glmer()` in the package `lme4` (Bates et al. 2015) supports both adaptive Gauss–Hermite quadrature and Laplace approximation for univariate random effects. For multivariate random effects only Laplace approximation is supported by `glmer()`, presumably because of the inherent difficulties of higher dimensional quadrature. Laplace approximation eschews multivariate integration via quadratic approximation of the log-integrand. However, the resultant approximate inference is well known to be inaccurate, often to an unacceptable degree, in binary mixed models (e.g., McCulloch et al., sec. 14.4). An embellishment of Laplace approximation, known as integrated nested Laplace approximation (Rue, Martino, and Chopin 2009), has been successful in various Bayesian inference contexts.

Expectation propagation (e.g., Minka 2001) is a general prescription for approximation of integrals that arise in statistical inference problems. Most of its literature is within the realm of Computer Science and, in particular, geared toward approximate inference for Bayesian graphical models (e.g., Bishop 2006, chap. 10). A major contribution of this article is transferral of expectation propagation methodology to frequentist statistical inference. In principle, our approach applies to any generalized linear mixed model situation. However, expectation propagation for binary response mixed model analysis has some especially attractive features and therefore we focus on this class of models. In the special case of probit mixed models, the expectation propagation approximation to the log-likelihood is exact regardless of the dimension of the random effects. This leads to a new practical alternative to multivariate quadrature. Moreover, asymptotic theory reveals that expectation propagation provides consistent approximation of the exact likelihood surface. This implies very good inferential accuracy of expectation propagation, and is supported by our simulation results. We are not aware of any other quadrature-free approaches to generalized mixed model analysis that has such a strong theoretical underpinning.

CONTACT Matt P. Wand matt.wand@uts.edu.au University of Technology Sydney, School of Mathematical and Physical Sciences, P.O. Box 123, Broadway, 2007 Australia.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

To facilitate widespread use of the new approach, a new package in the R language (R Core Team 2019) has been launched. The package, `glmmEP` (Wand and Yu 2019), uses a low-level language implementation of expectation propagation for speedy approximate likelihood-based inference and scales well to large sample sizes.

Binary response mixed models and their inherent computational challenges are summarized in Section 2. The expectation propagation approach to fitting and approximate inference, with special attention given to the quadrature-free probit link situation, is given in Section 3. Section 4 presents the results of numerical studies for both simulated and real data, and shows expectation propagation to be of great practical value as a fast, high-quality approximation that scales well to big data and big model situations. Theoretical considerations are summarized in Section 5. Higher level and random effects extensions are touched upon in Section 6. Lastly, we briefly discuss transferral of our new approach to other generalized linear mixed model settings in Section 7.

2. Binary Response Mixed Models

Binary mixed models for grouped data with one level of nesting and Gaussian random effects has the general form

$$y_{ij}|\boldsymbol{u}_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\big(F\big(\boldsymbol{\beta}^T\boldsymbol{x}^{\text{F}}_{ij} + \boldsymbol{u}_i^T\boldsymbol{x}^{\text{R}}_{ij}\big)\big), \quad \boldsymbol{u}_i \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, \Sigma), \\ 1 \le i \le m, \quad 1 \le j \le n_i \tag{1}$$

where F, the *inverse link*, is a prespecified cumulative distribution function and y_{ij} is the jth response for the ith group, where the number of groups is m and the number of response measurements within the ith group is n_i. Also, $\boldsymbol{x}^{\text{F}}_{ij}$ is a $d^{\text{F}} \times 1$ vector of predictors corresponding to y_{ij}, modeled as having fixed effects with coefficient vector $\boldsymbol{\beta}$. Similarly, $\boldsymbol{x}^{\text{R}}_{ij}$ is a $d^{\text{R}} \times 1$ vector of predictors modeled as having random effects with coefficient vectors $\boldsymbol{u}_i$, $1 \le i \le m$. Typically, $\boldsymbol{x}^{\text{R}}_{ij}$ is a sub-vector of $\boldsymbol{x}^{\text{F}}_{ij}$. It is also very common for each of $\boldsymbol{x}^{\text{R}}_{ij}$ and $\boldsymbol{x}^{\text{F}}_{ij}$ to have first entry equal to 1, corresponding to fixed and random intercepts. The random effects covariance matrix Σ has dimension $d^{\text{R}} \times d^{\text{R}}$.

By far, the most common choices for F are

$$F = \begin{cases} \text{expit} & \text{for logistic mixed models} \\ \Phi & \text{for probit mixed models} \end{cases}$$

where $\text{expit}(x) \equiv 1/(1+e^{-x})$ and Φ is the cumulative distribution function of the $N(0,1)$ distribution.

Despite the simple form of (1), likelihood-based inference for the parameters $\boldsymbol{\beta}$ and Σ and best prediction of the random effects $\boldsymbol{u}_i$ is very numerically challenging. Assuming that $F(x) + F(-x) = 1$, as is the case for the logistic and probit cases, the log-likelihood is

$$\ell(\boldsymbol{\beta}, \Sigma) = \sum_{i=1}^{m} \log \int_{\mathbb{R}^{d^{\text{R}}}} \left\{ \prod_{j=1}^{n_i} F\big((2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}^{\text{F}}_{ij} + \boldsymbol{u}^T\boldsymbol{x}^{\text{R}}_{ij})\big) \right\} \\ \times |2\pi\,\Sigma|^{-1/2} \exp(-\tfrac{1}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u})\, d\boldsymbol{u} \tag{2}$$

and the best predictor of $\boldsymbol{u}_i$ is

$$\text{BP}(\boldsymbol{u}_i) = \frac{\displaystyle\int_{\mathbb{R}^{d^{\text{R}}}} \boldsymbol{u} \left\{ \prod_{j=1}^{n_i} F\big((2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}^{\text{F}}_{ij} + \boldsymbol{u}^T\boldsymbol{x}^{\text{R}}_{ij})\big) \right\} \times \exp(-\tfrac{1}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u})\, d\boldsymbol{u}}{\displaystyle\int_{\mathbb{R}^{d^{\text{R}}}} \left\{ \prod_{j=1}^{n_i} F\big((2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}^{\text{F}}_{ij} + \boldsymbol{u}^T\boldsymbol{x}^{\text{R}}_{ij})\big) \right\} \times \exp(-\tfrac{1}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u})\, d\boldsymbol{u}}, \\ 1 \le i \le m.$$

The d^{R}-dimensional integrals in the $\ell(\boldsymbol{\beta}, \Sigma)$ and $\text{BP}(\boldsymbol{u}_i)$ expressions cannot be reduced further and multivariate numerical integration must be called upon for their evaluation. In addition, $\ell(\boldsymbol{\beta}, \Sigma)$ has to be maximized over $\{d^{\text{F}} + \frac{1}{2}\,d^{\text{R}}(d^{\text{R}}+1)\}$-dimensional space to obtain maximum likelihood estimates. Last, there is the problem of obtaining approximate confidence intervals for the entries of $\boldsymbol{\beta}$ and Σ and approximate prediction intervals for the entries of $\boldsymbol{u}_i$.

Starting around the early 1990s there have been several proposals for likelihood-based estimation and inference for binary response mixed models and their generalized linear mixed model extensions. Section 14.3 of McCulloch, Searle, and Neuhaus (2008) provides a summary of the main approaches up until the mid-2000s. Some more recent contributions include Jeon, Rijmen, and Rabe-Hesketh (2017), Lele, Nadeem, and Schmuland (2010), Ogden (2015), and Wand and Ormerod (2012). Section 3.3.1 of Jiang (2017) provides a more recent historical overview. The relative strengths and weakness of the various proposals depend on attributes such as accuracy, ease of implementation, computational speed and theoretical tractability and properties.

3. Expectation Propagation Likelihood Approximation

We will first explain expectation propagation for approximation of the log-likelihood $\ell(\boldsymbol{\beta}, \Sigma)$. Approximation of $\text{BP}(\boldsymbol{u}_i)$ follows relatively quickly. First note that $\ell(\boldsymbol{\beta}, \Sigma) = \sum_{i=1}^{m} \ell_i(\boldsymbol{\beta}, \Sigma)$ where

$$\ell_i(\boldsymbol{\beta}, \Sigma) \equiv \log \int_{\mathbb{R}^{d^{\text{R}}}} \left\{ \prod_{j=1}^{n_i} F\big((2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}^{\text{F}}_{ij} + \boldsymbol{u}^T\boldsymbol{x}^{\text{R}}_{ij})\big) \right\} \\ \times |2\pi\,\Sigma|^{-1/2} \exp(-\tfrac{1}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u})\, d\boldsymbol{u}.$$

Each of the $\ell_i(\boldsymbol{\beta}, \Sigma)$ are approximated individually and then summed to approximate $\ell(\boldsymbol{\beta}, \Sigma)$. The essence of the approximation of $\ell_i(\boldsymbol{\beta}, \Sigma)$ is replacement of each

$$F\big((2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}^{\text{F}}_{ij} + \boldsymbol{u}^T\boldsymbol{x}^{\text{R}}_{ij})\big), \quad 1 \le j \le n_i,$$

by an unnormalized Multivariate Normal density function, chosen according to an appropriate minimum Kullback–Leibler divergence criterion. The resultant integrand is then proportional to a product of Multivariate Normal density functions and admits an explicit form. The number of approximating density functions is of the same order of magnitude and, together with the properties of minimum Kullback–Leibler divergence, leads to accurate and statistically consistent approximation of $\ell(\boldsymbol{\beta}, \Sigma)$. In the probit case, where $F = \Phi$, the minimum Kullback–Leibler divergence steps are explicit. This leads to accurate

approximation of $\ell(\boldsymbol{\beta}, \Sigma)$ without the need for any numerical integration – just some fixed-point iteration. The expectation propagation-approximate log-likelihood, which we denote by $\widetilde{\ell}(\boldsymbol{\beta}, \Sigma)$, can be evaluated quite rapidly and maximized using established derivative-free methods such as the Nelder–Mead algorithm (Nelder and Mead 1965) or quasi-Newton optimization methods such as the Broyden–Fletcher–Goldfarb–Shanno approach with numerical derivatives. The latter also facilitates Hessian matrix approximation at the maximum, which can be used to construct approximate confidence intervals.

We now provide the details, with subsections on Kullback–Leibler projection onto unnormalized Multivariate Normal density functions, message passing formulation for organizing the required versions of these projections and quasi-Newton-based approximate inference. The upcoming subsections require some specialized matrix notation. If $\boldsymbol{A}$ is $d \times d$ matrix then $\text{vec}(\boldsymbol{A})$ is the $d^2 \times 1$ vector obtained by stacking the columns of $\boldsymbol{A}$ underneath each other in order from left to right. Also, $\text{vech}(\boldsymbol{A})$ is $\frac{1}{2}d(d+1)+1$ vector defined similarly to $\text{vec}(\boldsymbol{A})$ but only involving entries on and below the diagonal. The *duplication matrix of order d*, denoted by $\boldsymbol{D}_d$, is the unique $d^2 \times \frac{1}{2}d(d+1)$ matrix of zeros and ones such that

$$\boldsymbol{D}_d \,\text{vech}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A}) \quad \text{for} \quad \boldsymbol{A} = \boldsymbol{A}^T.$$

The Moore–Penrose inverse of $\boldsymbol{D}_d$ is

$$\boldsymbol{D}_d^+ \equiv (\boldsymbol{D}_d^T \boldsymbol{D}_d)^{-1} \boldsymbol{D}_d^T.$$

3.1. Projection onto Unnormalized Multivariate Normal Density Functions

Let $L_1(\mathbb{R}^d)$ denote the set of absolutely integrable functions on $\mathbb{R}^d$. For $f_1, f_2 \in L_1(\mathbb{R}^d)$ such that $f_1, f_2 \geq 0$, the Kullback-Leibler divergence of f_2 from f_1 is

$$\text{KL}(f_1 \| f_2) = \int_{\mathbb{R}^d} \left[f_1(\boldsymbol{x}) \log\{f_1(\boldsymbol{x})/f_2(\boldsymbol{x})\} + f_2(\boldsymbol{x}) - f_1(\boldsymbol{x}) \right] d\boldsymbol{x} \quad (3)$$

(e.g., Minka 2005). In the special case where f_1 and f_2 are density functions the right-hand side of Equation (3) reduces to the more common Kullback–Leibler divergence expression. However, we require this more general form that caters for *unnormalized* density functions.

Now consider the family of functions on $\mathbb{R}^d$ of the form

$$f_{\text{UN}}(\boldsymbol{x}) \equiv \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \begin{bmatrix} \eta_0 \\ \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} \right\} \quad (4)$$

where $\eta_0 \in \mathbb{R}$, $\boldsymbol{\eta}_1$ is a $d \times 1$ vector and $\boldsymbol{\eta}_2$ is a $\frac{1}{2}d(d+1) \times 1$ vector restricted in such a way that $f_{\text{UN}} \in L_1(\mathbb{R}^d)$. Then (4) is the family of unnormalized Multivariate Normal density functions written in exponential family form with natural parameters η_0, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$.

Expectation propagation for generalized linear mixed models with Gaussian random effects has the following notion at its core:

$$\begin{aligned} &\text{given } f_{\text{input}} \in L_1(\mathbb{R}^d), \text{ determine the } \eta_0, \boldsymbol{\eta}_1 \text{ and } \boldsymbol{\eta}_2 \text{ that} \\ &\quad \text{minimizes } \text{KL}(f_{\text{input}} \| f_{\text{UN}}). \end{aligned} \quad (5)$$

The solution is termed the (Kullback–Leibler) projection onto the family of Multivariate Normal density functions and we write

$$\text{proj}[f_{\text{input}}](\boldsymbol{x}) \equiv \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \begin{bmatrix} \eta_0^* \\ \boldsymbol{\eta}_1^* \\ \boldsymbol{\eta}_2^* \end{bmatrix} \right\}$$

where

$$(\eta_0^*, \boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*) = \text{argmin}_{(\eta_0, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \in H} \text{KL}\big(f_{\text{input}} \| f_{\text{UN}}\big),$$

with H denoting the set of all allowable natural parameters. Note that the special case of Kullback-Leibler projection onto the unnormalized Multivariate Normal family has a simple moment-matching representation, with $(\eta_0^*, \boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*)$ being the unique vector such that zeroth-, first- and second-order moments of f_{UN} match those of f_{input}.

For the binary mixed model (1), expectation propagation requires repeated projection of the form

$$f_{\text{input}}(\boldsymbol{x}) = F(c_0 + \boldsymbol{c}_1^T \boldsymbol{x}) \exp\left\{ \begin{bmatrix} \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\eta}_1^{\text{input}} \\ \boldsymbol{\eta}_2^{\text{input}} \end{bmatrix} \right\}$$

onto the unnormalized Multivariate Normal family. An important observation is that in the case of probit mixed models, $\text{proj}[f_{\text{input}}](\boldsymbol{x})$ has an exact solution.

Let $\zeta(x) \equiv \log\{2\Phi(x)\}$. It follows that

$$\zeta'(x) = \phi(x)/\Phi(x) \quad \text{and} \quad \zeta''(x) = -\zeta'(x)\{x + \zeta'(x)\},$$

where $\phi(x) \equiv (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$ is the $N(0,1)$ density function. We are now in a position to define two algebraic functions which are fundamental for approximate likelihood-based inference in probit mixed models based on expectation propagation:

Definition 1. For primary arguments $\boldsymbol{a}_1$ $(d \times 1)$ and $\boldsymbol{a}_2$ $(\frac{1}{2}d(d+1) \times 1)$ such that $\text{vec}^{-1}(-\boldsymbol{D}_d^{+T}\boldsymbol{a}_2)$ is symmetric and positive definite, and auxiliary arguments $c_0 \in \mathbb{R}$ and $\boldsymbol{c}_1$ $(d \times 1)$ the function K_{probit} is given by

$$K_{\text{probit}}\left(\begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix}; c_0, \boldsymbol{c}_1 \right) \equiv \begin{bmatrix} \boldsymbol{R}_5^T(\boldsymbol{a}_1 + r_3\boldsymbol{c}_1) \\ \boldsymbol{D}_d^T \text{vec}(\boldsymbol{R}_5^T \boldsymbol{A}_2) \end{bmatrix}$$

with

$$\begin{aligned} \boldsymbol{A}_2 &\equiv \text{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{a}_2), \quad r_1 \equiv \sqrt{2(2 - \boldsymbol{c}_1^T \boldsymbol{A}_2^{-1} \boldsymbol{c}_1)}, \\ r_2 &\equiv (2c_0 - \boldsymbol{c}_1^T \boldsymbol{A}_2^{-1} \boldsymbol{a}_1)/r_1, \quad r_3 \equiv 2\zeta'(r_2)/r_1, \\ r_4 &\equiv -2\zeta''(r_2)/r_1^2 \quad \text{and} \quad \boldsymbol{R}_5 \equiv \left(\boldsymbol{A}_2 + r_4 \boldsymbol{c}_1 \boldsymbol{c}_1^T\right)^{-1} \boldsymbol{A}_2 \end{aligned}$$

and the function A_N is given by

$$A_N\left(\begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix} \right) \equiv -\frac{1}{4}\boldsymbol{a}_1^T \boldsymbol{A}_2^{-1} \boldsymbol{a}_1 - \frac{1}{2}\log\left| -2\boldsymbol{A}_2 \right|.$$

In addition, for primary arguments $\boldsymbol{a}_1$, $\boldsymbol{b}_1$ (each $d \times 1$) and $\boldsymbol{a}_2$, $\boldsymbol{b}_2$ (each $\frac{1}{2}d(d+1) \times 1$) such that both $\text{vec}^{-1}(-\boldsymbol{D}_d^{+T}\boldsymbol{a}_2)$ and $\text{vec}^{-1}(-\boldsymbol{D}_d^{+T}\boldsymbol{b}_2)$ are symmetric and positive definite, and

auxiliary arguments $c_0 \in \mathbb{R}$ and $\boldsymbol{c}_1$ $(d \times 1)$, the function C_{probit} is given by

$$C_{\text{probit}}\left(\begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}; c_0, \boldsymbol{c}_1\right) \equiv \log \Phi(r_2) + \frac{1}{4}\boldsymbol{b}_1^T \boldsymbol{B}_2^{-1}\boldsymbol{b}_1 - \frac{1}{4}\boldsymbol{a}_1^T \boldsymbol{A}_2^{-1}\boldsymbol{a}_1 + \frac{1}{2}\log\{|\boldsymbol{B}_2|/|\boldsymbol{A}_2|\}$$

with $\boldsymbol{B}_2 \equiv \text{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{b}_2)$.

Inspection of Definition 1 reveals that the K_{probit} and C_{probit} functions are simple functions up to evaluations of $\log(\Phi)$ and $\zeta' = \phi/\Phi$. Even though software for Φ is widely available, direct computation of $\log(\Phi)$ and ζ' can be unstable and software such as the function `zeta()` in the R package `sn` (Azzalini 2017) is recommended. Another option is use of continued fraction representation and Lentz's Algorithm (e.g., Wand and Ormerod 2012).

Expectation propagation for probit mixed models relies heavily upon:

Theorem 1. If

$$f_{\text{input}}(\boldsymbol{x}) = \Phi(c_0 + \boldsymbol{c}_1^T\boldsymbol{x}) \exp\left\{\begin{bmatrix} \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\eta}_1^{\text{input}} \\ \boldsymbol{\eta}_2^{\text{input}} \end{bmatrix}\right\}$$

then

$$\text{proj}[f_{\text{input}}](\boldsymbol{x}) = \exp\left\{\begin{bmatrix} 1 \\ \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \begin{bmatrix} \eta_0^* \\ \boldsymbol{\eta}_1^* \\ \boldsymbol{\eta}_2^* \end{bmatrix}\right\}$$

where

$$\begin{bmatrix} \boldsymbol{\eta}_1^* \\ \boldsymbol{\eta}_2^* \end{bmatrix} = K_{\text{probit}}\left(\begin{bmatrix} \boldsymbol{\eta}_1^{\text{input}} \\ \boldsymbol{\eta}_2^{\text{input}} \end{bmatrix}; c_0, \boldsymbol{c}_1\right) \quad \text{and}$$

$$\eta_0^* = C_{\text{probit}}\left(\begin{bmatrix} \boldsymbol{\eta}_1^{\text{input}} \\ \boldsymbol{\eta}_2^{\text{input}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\eta}_1^* \\ \boldsymbol{\eta}_2^* \end{bmatrix}; c_0, \boldsymbol{c}_1\right).$$

A proof of Theorem 1 is given in Section S.1 of the online supplement.

3.2. Message Passing Formulation

The ith summand of $\ell(\boldsymbol{\beta}, \Sigma)$ can be written as

$$\ell_i(\boldsymbol{\beta}, \Sigma) = \log \int_{\mathbb{R}^{d^R}} \left\{\prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\beta})\right\} p(\boldsymbol{u}_i; \Sigma)\, d\boldsymbol{u}_i \quad (6)$$

where, for $1 \le j \le n_i$,

$$p(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\beta}) \equiv F\big((2y_{ij} - 1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}} + \boldsymbol{u}_i^T\boldsymbol{x}_{ij}^{\text{R}})\big) \quad \text{and}$$

$$p(\boldsymbol{u}_i; \Sigma) \equiv |2\pi\,\Sigma|^{-1/2} \exp\big(-\tfrac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i\big)$$

are, respectively, the conditional density functions of each response given its random effect and the density function of that random effect. Note that product structure of the integrand in Equation (6) can be represented using *factor graph* shown in Figure 1. The circle in Figure 1 corresponds to the random vector $\boldsymbol{u}_i$ and factor graph parlance is a *stochastic variable node*. The solid rectangles correspond to each of the $n_i + 1$ *factors* in the Equation (6) integrand. Each of these factors depend on $\boldsymbol{u}_i$, which is signified by an edge connecting each factor node to the lone stochastic variable node.

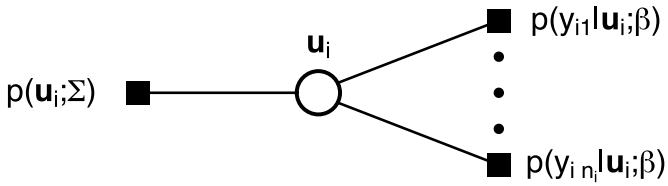

Figure 1. Factor graph representation of the product structure of the integrand in Equation (6). The open circle corresponds to the random effect vector $\boldsymbol{u}_i$ and the solid rectangles indicate factors. Edges indicate dependence of each factor on $\boldsymbol{u}_i$.

Expectation propagation approximation of $\ell_i(\boldsymbol{\beta}, \Sigma)$ involves projection onto the unnormalized Multivariate Normal family. Suppose that:

$$\underset{\sim}{p}(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\beta}) = \exp\left\{\begin{bmatrix} 1 \\ \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \end{bmatrix}^T \boldsymbol{\eta}_{ij}\right\}, \quad 1 \le j \le n_i \quad (7)$$

are initialized to be unnormalized Multivariate Normal density functions in $\boldsymbol{u}_i$. Then, for each $j = 1, \ldots, n_i$, the $\boldsymbol{\eta}_{ij}$ update involves minimization of

$$\text{KL}\left(p(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\beta}) \left\{\prod_{j' \ne j}^{n_i} \underset{\sim}{p}(y_{ij'}|\boldsymbol{u}_i; \boldsymbol{\beta})\right\} \times p(\boldsymbol{u}_i; \Sigma) \,\middle\|\, \left\{\prod_{j'=1}^{n_i} \underset{\sim}{p}(y_{ij'}|\boldsymbol{u}_i; \boldsymbol{\beta})\right\} p(\boldsymbol{u}_i; \Sigma)\right) \quad (8)$$

as functions of $\boldsymbol{u}_i$. Noting that this problem has the form of Equation (5), Theorem 1 can be used to perform the update explicitly in the case of a probit link. This procedure is then iterated until the $\boldsymbol{\eta}_{ij}$s converge.

A convenient way to keep track of the updates and compartmentalize the algebra and coding is to call upon the notion of *message passing*. Minka (2005) shows how to express expectation propagation as a message passing algorithm in the Bayesian graphical models context, culminating in his equation (54) and (83) update formulae. Exactly the same formulae arise here, as is made clear in Section S.2 of the online supplement. In particular, in keeping with (83) of Minka (2005), (8) can be expressed as

$$m_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}(\boldsymbol{u}_i) \longleftarrow \frac{\text{proj}\big[m_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)\, p(y_{ij}|u_i; \boldsymbol{\beta})\big](\boldsymbol{u}_i)}{m_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)}, \quad 1 \le j \le n_i, \quad (9)$$

where $m_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}(\boldsymbol{u}_i)$ is the *message* passed from the factor $p(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\beta})$ to the stochastic node $\boldsymbol{u}_i$ and $m_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)$ is the message passed from $\boldsymbol{u}_i$ back to $p(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\beta})$. The message passed from $p(\boldsymbol{u}_i; \Sigma)$ to $\boldsymbol{u}_i$ is

$$m_{p(\boldsymbol{u}_i;\Sigma)\to\boldsymbol{u}_i}(\boldsymbol{u}_i) \longleftarrow \frac{\text{proj}\big[m_{\boldsymbol{u}_i\to p(\boldsymbol{u}_i;\Sigma)}(\boldsymbol{u}_i)\, p(u_i; \Sigma)\big](\boldsymbol{u}_i)}{m_{\boldsymbol{u}_i\to p(\boldsymbol{u}_i;\Sigma)}(\boldsymbol{u}_i)}. \quad (10)$$

In keeping with Equation (54) of Minka (2005), the stochastic node to factor messages are updated according to

$$m_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i) = m_{p(\boldsymbol{u}_i;\Sigma)\to \boldsymbol{u}_i}(\boldsymbol{u}_i) \left\{ \prod_{j' \neq j}^{n_i} m_{p(y_{ij'}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}(\boldsymbol{u}_i) \right\}, \quad 1 \leq j \leq n_i, \tag{11}$$

and

$$m_{\boldsymbol{u}_i \to p(\boldsymbol{u}_i;\Sigma)}(\boldsymbol{u}_i) = \prod_{j=1}^{n_i} m_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}(\boldsymbol{u}_i). \tag{12}$$

As laid out at the end of Section 6 of Minka (2005), the expectation message passing protocol is:

Initialize all factor to stochastic node messages.
Cycle until all factor to stochastic node messages converge:

For each factor:

Compute the messages passed to the factor using (11) or (12).
Compute the messages passed from the factor using (9) or (10).

Upon convergence, the expectation propagation approximation to $\ell_i(\boldsymbol{\beta}, \Sigma)$ is

$$\underset{\sim}{\ell}_i(\boldsymbol{\beta}, \Sigma) = \log \int_{\mathbb{R}^{d^{\mathrm{R}}}} \left\{ \prod_{j=1}^{n_i} m_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}(\boldsymbol{u}_i) \right\} \times m_{p(\boldsymbol{u}_i;\Sigma)\to \boldsymbol{u}_i}(\boldsymbol{u}_i)\, d\boldsymbol{u}_i, \tag{13}$$

where the integrand is in keeping with the general form given by (44) of Minka and Winn (2008). The success of expectation propagation hinges on the fact that each of the messages in Equation (13) is an unnormalized Multivariate Normal density function and the integral over $\mathbb{R}^{d^{\mathrm{R}}}$ can be obtained exactly as follows:

$$\begin{aligned}
&\int_{\mathbb{R}^{d^{\mathrm{R}}}} \left\{ \prod_{j=1}^{n_i} m_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}(\boldsymbol{u}_i) \right\} m_{p(\boldsymbol{u}_i;\Sigma)\to \boldsymbol{u}_i}(\boldsymbol{u}_i)\, d\boldsymbol{u}_i \\
&= \int_{\mathbb{R}^{d^{\mathrm{R}}}} \left[\prod_{j=1}^{n_i} \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{u}_i \\ \mathrm{vech}(\boldsymbol{u}_i \boldsymbol{u}_i^T) \end{bmatrix}^T \boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i} \right\} \right] \\
&\quad \times \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{u}_i \\ \mathrm{vech}(\boldsymbol{u}_i \boldsymbol{u}_i^T) \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{u}_i;\Sigma)\to \boldsymbol{u}_i} \right\} d\boldsymbol{u}_i \\
&= (2\pi)^{d^{\mathrm{R}}/2} \exp\left\{ \left(\boldsymbol{\eta}_\Sigma + \mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}\} \right)_0 \right. \\
&\quad \left. + A_N \left(\left(\boldsymbol{\eta}_\Sigma + \mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}\} \right)_{-0} \right) \right\}
\end{aligned}$$

where

$$\boldsymbol{\eta}_\Sigma \equiv \begin{bmatrix} -\frac{1}{2}\log|2\pi\Sigma| \\ \mathbf{0}_{d^{\mathrm{R}}} \\ -\frac{1}{2}\boldsymbol{D}_{d^{\mathrm{R}}}^T \mathrm{vec}(\Sigma^{-1}) \end{bmatrix},$$

$$\mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}\} \equiv \sum_{j=1}^{n_i} \boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i},$$

A_N is as defined in Definition 1 and, for an unnormalized Multivariate Normal natural parameter vector $\boldsymbol{\eta}$, $\boldsymbol{\eta}_0$ denotes the first entry (the zero subscript is indicative of the first entry being the coefficient of 1) and $\boldsymbol{\eta}_{-0}$ denotes the remaining entries.

The full algorithm for expectation propagation approximation of $\ell(\boldsymbol{\beta}, \Sigma)$ is summarized as Algorithm 1. The derivational details are given in Section S.2. A key point is that each of the message passing updates Equations (9)–(12) is expressed in Algorithm 1 in terms of updates to natural parameter vectors.

We have carried out extensive simulated data tests on Algorithm 1 using the starting values described in Section 3.3 and found convergence to be rapid. Moreover, each of updates in Algorithm 1 involve explicit calculations and low-level language implementation, used in our R package `glmmEP`, affords very fast evaluation of the approximate log-likelihood surface. As explained in (3.4), quasi-Newton methods can be used for maximization of $\underset{\sim}{\ell}(\boldsymbol{\beta}, \Sigma)$ and approximate likelihood-based inference.

3.3. Recommended Starting Values for Algorithm 1

In Section S.3 of the online supplement, we use a Taylor series argument to justify the following starting values for $\boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}$ in Algorithm 1:

$$\boldsymbol{\eta}^{\mathrm{start}}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i} \equiv \begin{bmatrix} 0 \\ (2y_{ij}-1)\zeta'(\widehat{a}_{ij})\boldsymbol{x}_{ij}^{\mathrm{R}} - \zeta''(\widehat{a}_{ij})\boldsymbol{x}_{ij}^{\mathrm{R}}(\boldsymbol{x}_{ij}^{\mathrm{R}})^T\widehat{\boldsymbol{u}}_i \\ \frac{1}{2}\zeta''(\widehat{a}_{ij})\boldsymbol{D}_{d^{\mathrm{R}}}^T \mathrm{vec}\left(\boldsymbol{x}_{ij}^{\mathrm{R}}(\boldsymbol{x}_{ij}^{\mathrm{R}})^T\right) \end{bmatrix}, \quad 1 \leq j \tag{14}$$

where

$$\widehat{a}_{ij} \equiv (2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\mathrm{F}} + \widehat{\boldsymbol{u}}_i^T\boldsymbol{x}_{ij}^{\mathrm{R}})$$

and $\widehat{\boldsymbol{u}}_i$ is a prediction of $\boldsymbol{u}_i$. A convenient choice for $\widehat{\boldsymbol{u}}_i$ is that based on Laplace approximation. In the R computing environment, the function `glmer()` in the package `lme4` (Bates et al. 2015) provides fast Laplace approximation-based predictions for the $\boldsymbol{u}_i$. In our numerical experiments, we found convergence of the cycle loop of Algorithm 1 to be quite rapid, with convergents of

$$\left(\boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}\right)_{-0} \quad \text{relatively close to} \quad \left(\boldsymbol{\eta}^{\mathrm{start}}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to \boldsymbol{u}_i}\right)_{-0}.$$

Therefore, we strongly recommend the starting values (14).

3.4. Quasi-Newton Optimization and Approximate Inference

Even though Algorithm 1 provides fast approximate evaluation of the probit mixed model likelihood surface, we still need to maximize over $(\boldsymbol{\beta}, \Sigma)$ to obtain the expectation propagation-approximate maximum likelihood estimators $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\Sigma}})$. This is also the issue of approximate inference based on Fisher information theory.

Algorithm 1 Expectation propagation approximation of the log-likelihood for the probit mixed model (1) with $F = \Phi$ via message passing on the Figure 1 factor graph.

Inputs: $y_{ij}, \boldsymbol{x}_{ij}^{\mathrm{F}}, \boldsymbol{x}_{ij}^{\mathrm{R}}$, $1 \le i \le m$, $1 \le j \le n_i$;
$\boldsymbol{\beta}$ ($d^{\mathrm{F}} \times 1$), Σ ($d^{\mathrm{R}} \times d^{\mathrm{R}}$, symmetric and positive definite).
Set constants: $c_{0,ij} \longleftarrow (2y_{ij} - 1)(\boldsymbol{\beta}^T \boldsymbol{x}_{ij}^{\mathrm{F}})$; $\boldsymbol{c}_{1,ij} \longleftarrow (2y_{ij} - 1)\boldsymbol{x}_{ij}^{\mathrm{R}}$, $\quad 1 \le i \le m,\ 1 \le j \le n_i$;

$$\boldsymbol{\eta}_{p(\boldsymbol{u}_i;\Sigma)\to\boldsymbol{u}_i} \longleftarrow \boldsymbol{\eta}_{\Sigma} \equiv \begin{bmatrix} -\frac{1}{2}\log|2\pi\Sigma| \\ \boldsymbol{0}_{d^{\mathrm{R}}} \\ -\frac{1}{2}\boldsymbol{D}_{d^{\mathrm{R}}}^T \mathrm{vec}(\Sigma^{-1}) \end{bmatrix}, \quad 1 \le i \le m.$$

For $i = 1, \ldots, m$:

Initialize: $\boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}$, $\quad 1 \le j \le n_i$ (see Section 3.3 for a recommendation)
Cycle:

$$\mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\} \longleftarrow \sum_{j=1}^{n_i} \boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}$$

For $j = 1, \ldots, n_i$:

$$\boldsymbol{\eta}_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{u}_i;\Sigma)\to\boldsymbol{u}_i} + \mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\} - \boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}$$

$$\left(\boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\right)_{-0} \longleftarrow K_{\mathrm{probit}}\left(\left(\boldsymbol{\eta}_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\right)_{-0}; c_{0,ij}, \boldsymbol{c}_{1,ij}\right) - \left(\boldsymbol{\eta}_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\right)_{-0}$$

until all natural parameter vectors converge.
For $j = 1, \ldots, n_i$:

$$\left(\boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\right)_{0} \longleftarrow C_{\mathrm{probit}}\left(\left(\boldsymbol{\eta}_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\right)_{-0}, \left(\boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\right)_{-0} + \left(\boldsymbol{\eta}_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\right)_{-0}; c_{0,ij}, \boldsymbol{c}_{1,ij}\right)$$

$$\mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\} \longleftarrow \sum_{j=1}^{n_i} \boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}$$

Output: The expectation propagation approximate log-likelihood given by

$$\underset{\sim}{\ell}(\boldsymbol{\beta}, \Sigma) = \frac{1}{2} m\, d^{\mathrm{R}} \log(2\pi) + \sum_{i=1}^{m} \left\{ \left(\boldsymbol{\eta}_{\Sigma} + \mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\}\right)_0 + A_N\left(\left(\boldsymbol{\eta}_{\Sigma} + \mathrm{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\}\right)_{-0}\right)\right\}$$

Since $\underset{\sim}{\ell}(\boldsymbol{\beta}, \Sigma)$ is defined implicitly via an iterative scheme, differentiation for use in derivative-based optimization techniques is not straightforward. A practical workaround involves the employment of optimization methods such as those of the quasi-Newton variety for which derivatives are approximated numerically. In the R computing environment, the function `optim()` supports several derivative-free optimization implementations. The Matlab computing environment (The Mathworks Incorporated 2018) has similar capabilities via functions such as `fminunc()`. In the `glmmEP` package and the examples in Section 4, we use the Broyden–Fletcher–Goldfarb–Shanno quasi-Newton method (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970) with Nelder-Mead starting values. Section 2.2.2.3 of Givens and Hoetig (2005) provides a concise summary of the Broyden–Fletcher–Goldfarb–Shanno method.

Since Σ is constrained to be symmetric and positive definite, we instead perform quasi-Newton optimization over the unconstrained parameter vector $(\boldsymbol{\beta}, \boldsymbol{\theta})$ where

$$\boldsymbol{\theta} \equiv \mathrm{vech}\left(\frac{1}{2}\log(\Sigma)\right)$$

and $\log(\Sigma)$ is the matrix logarithm of Σ (e.g., Section 2.2 of Pinheiro and Bates 2000). Note that $\log(\Sigma)$ can be obtained

using

$$\log(\Sigma) = \boldsymbol{U}_\Sigma \text{diag}\{\log(\boldsymbol{\lambda}_\Sigma)\}\boldsymbol{U}_\Sigma^T \quad \text{where}$$
$$\Sigma = \boldsymbol{U}_\Sigma \text{diag}(\boldsymbol{\lambda}_\Sigma)\boldsymbol{U}_\Sigma^T$$

is the spectral decomposition of Σ and $\log(\boldsymbol{\lambda}_\Sigma)$ denotes element-wise evaluation of the logarithm to the entries of $\boldsymbol{\lambda}_\Sigma$. If $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\theta}}})$ is the maximizer of $\underset{\sim}{\ell}$ then the expectation propagation-approximate maximum likelihood estimate of Σ is

$$\widehat{\underset{\sim}{\Sigma}} = \boldsymbol{U}_{\widehat{\underset{\sim}{\boldsymbol{\theta}}}} \text{diag}\{\exp(2\boldsymbol{\lambda}_{\widehat{\underset{\sim}{\boldsymbol{\theta}}}})\}\boldsymbol{U}_{\widehat{\underset{\sim}{\boldsymbol{\theta}}}}^T \quad \text{where}$$
$$\text{vech}^{-1}(\widehat{\underset{\sim}{\boldsymbol{\theta}}}) = \boldsymbol{U}_{\widehat{\underset{\sim}{\boldsymbol{\theta}}}} \text{diag}(\boldsymbol{\lambda}_{\widehat{\underset{\sim}{\boldsymbol{\theta}}}})\boldsymbol{U}_{\widehat{\underset{\sim}{\boldsymbol{\theta}}}}^T$$

is the spectral decomposition of the $\text{vech}^{-1}(\widehat{\underset{\sim}{\boldsymbol{\theta}}})$. Note that $\text{vech}^{-1}(\boldsymbol{a})$ is the symmetric matrix $\boldsymbol{A}$ of appropriate dimension such that $\text{vech}(\boldsymbol{A}) = \boldsymbol{a}$.

The `optim()` function in R and the `fminunc()` function in Matlab each have the option of computing an approximation to the Hessian matrix at the optimum, which can be used for approximate likelihood-based inference. In particular, we can use the approximate Hessian matrix to construct confidence intervals for the entries of $\boldsymbol{\beta}$ and the standard deviation and correlation parameters of Σ. The full details are given in Section S.4 of the online supplement. Here, we sketch the idea for the special case of $d^{\text{R}} = 2$, for which

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

For confidence interval construction, it is appropriate (e.g., Section 2.4 of Pinheiro and Bates) to work with the parameter vector

$$\boldsymbol{\omega} \equiv \begin{bmatrix} \log(\sigma_1) \\ \log(\sigma_2) \\ \tanh^{-1}(\rho) \end{bmatrix}.$$

Approximate $100(1-\alpha)\%$ confidence intervals for the entries of $(\boldsymbol{\beta}, \boldsymbol{\omega})^T$ are

$$\begin{bmatrix} \widehat{\underset{\sim}{\boldsymbol{\beta}}} \\ \widehat{\underset{\sim}{\boldsymbol{\omega}}} \end{bmatrix} \pm \Phi^{-1}(1-\frac{1}{2}\alpha)\sqrt{-\text{diagonal}\left(\{\text{H}\,\underset{\sim}{\ell}(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\omega}}})\}^{-1}\right)} \quad (15)$$

where $\text{H}\,\underset{\sim}{\ell}(\boldsymbol{\beta}, \boldsymbol{\omega})$ is the Hessian matrix of $\underset{\sim}{\ell}$ with respect to the $(\boldsymbol{\beta}, \boldsymbol{\omega})$ parameter vector. Confidence intervals for the entries of $\boldsymbol{\beta}$, σ_1, σ_2 and ρ follow from standard inversion manipulations.

Note that $(\boldsymbol{\beta}, \theta)$ is an unconstrained parameterization whilst $(\boldsymbol{\beta}, \boldsymbol{\omega})$ is a constrained parameterization. Hence, the optimization should be performed with respect to the former parameterization whereas the Hessian matrix in (15) is respect to the latter parameterization. In the examples of Section 4 and the **R** package `glmmEP` we use the following strategy:

- Obtain $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\theta}}})$ using `optim()` with the $(\boldsymbol{\beta}, \boldsymbol{\theta})$ parameterization in the function being maximized and the `hessian` argument set to `FALSE`.
- Compute $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\omega}}})$ and use this as an initial value with a call to `optim()` with the $(\boldsymbol{\beta}, \boldsymbol{\omega})$ parameterization in the function being maximized and the `hessian` argument set to `TRUE`.

Full details of confidence interval calculations for the general multivariate random effects situation are given in Section S.4 of the online supplement.

In our numerical experiments, we have found Nelder-Mead followed by Broyden-Fletcher-Goldfarb-Shanno optimization of expectation propagation approximate log-likelihood, with confidence intervals based on the approximate Hessian matrix, to be very effective. In Section 4, we present simulation results that show this strategy producing fast and accurate inference for binary mixed models.

3.5. Expectation Propagation Approximate Best Prediction

The best predictors of $\boldsymbol{u}_i$ are

$$\text{BP}(\boldsymbol{u}_i) \equiv E(\boldsymbol{u}_i|\boldsymbol{y}), \quad 1 \le i \le m.$$

We now show that Algorithm 1 provides, as by-products, straightforward empirical best predictions of the $\boldsymbol{u}_i$.

Let

$$\widehat{\underset{\sim}{\boldsymbol{\eta}}}_i \equiv \boldsymbol{\eta}_\Sigma + \text{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\} = \begin{bmatrix} \widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i1} \\ \widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i2} \end{bmatrix} \quad (16)$$

where $\boldsymbol{\eta}_\Sigma$ and $\text{SUM}\{\boldsymbol{\eta}_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\}$ are as in Algorithm 1 with $(\boldsymbol{\beta}, \Sigma) = (\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\Sigma}})$, $\widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i1}$ is the subvector of $\widehat{\underset{\sim}{\boldsymbol{\eta}}}_i$ corresponding to the first d^{R} entries and $\widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i2}$ contains the remaining entries. Then in Section S.5 of the online supplement we show that a suitable empirical approximation to $\text{BP}(\boldsymbol{u}_i)$, based on the expectation propagation estimate, is

$$\widehat{\underset{\sim}{\text{BP}}}(\boldsymbol{u}_i) = -\frac{1}{2}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i2}\right)\right\}^{-1}\widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i1}. \quad (17)$$

The corresponding covariance matrix empirical approximation is

$$\widehat{\underset{\sim}{\text{cov}}}(\boldsymbol{u}_i|\boldsymbol{y}) = -\frac{1}{2}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i2}\right)\right\}^{-1}. \quad (18)$$

In view of equation (13.7) of McCulloch, Searle, and Neuhaus (2008), $\text{cov}\{\widehat{\underset{\sim}{\text{BP}}}(\boldsymbol{u}_i) - \boldsymbol{u}_i\}$ is approximated by $E_{\boldsymbol{y}_i}\{\widehat{\underset{\sim}{\text{cov}}}(\boldsymbol{u}_i|\boldsymbol{y}_i)\}$. Approximate prediction interval construction is hindered by this expectation over the sampling distribution of the responses. See, for example, Carlin and Gelfand (1991), for discussion and access to some of the relevant literature concerning valid prediction interval construction in the more general empirical Bayes' context.

4. Numerical Evaluation and Illustration

We now demonstrate the impressive accuracy and speed of Algorithm 1 combined with quasi-Newton methods for approximate likelihood-based inference for probit mixed models. First, we report the results of some studies involving simulated data. Analysis of actual data is discussed later in this section.

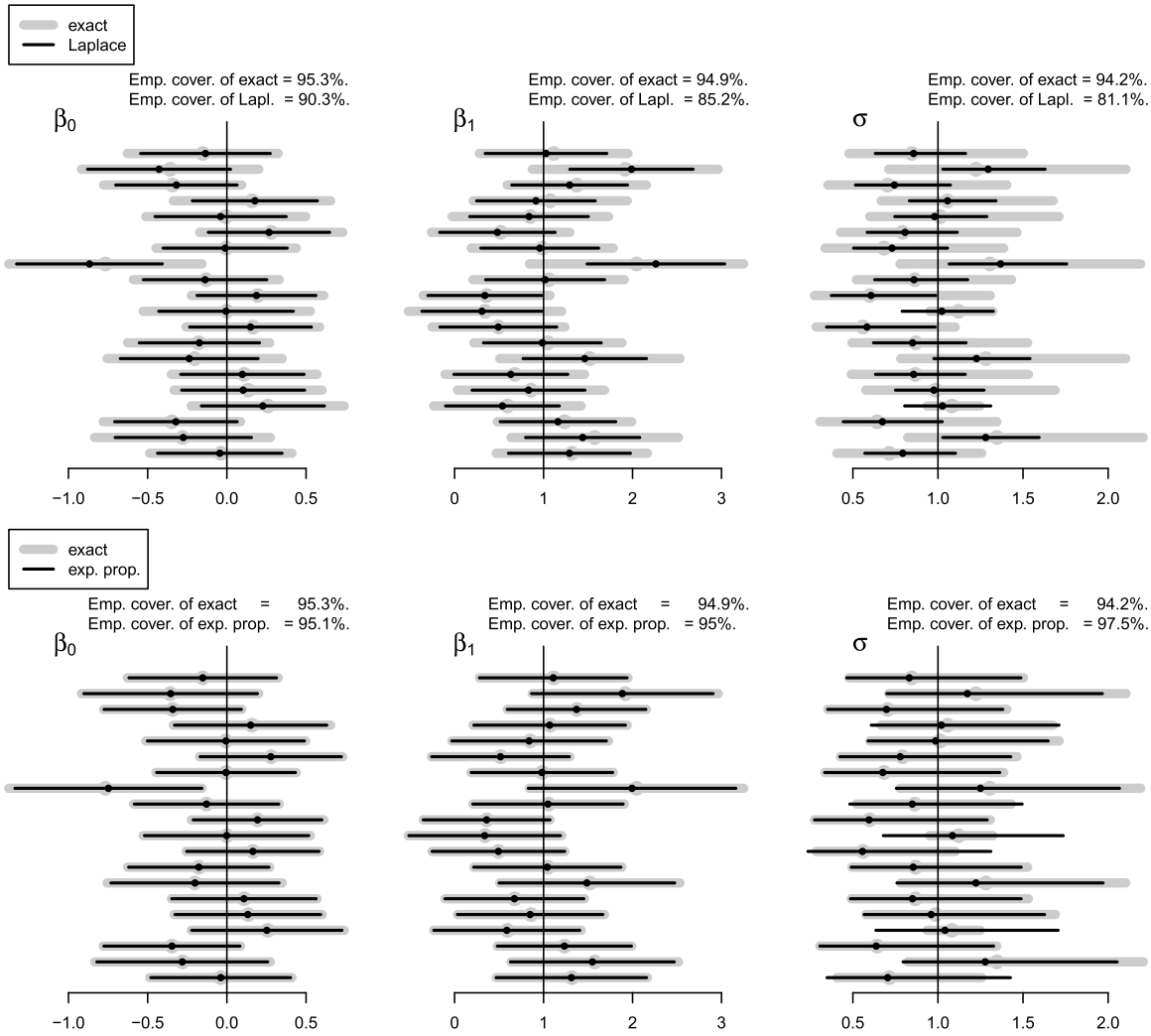

Figure 2. Comparison of point estimation and 95% confidence interval coverage for the first simulation study with true parameter values given by (19). The upper row of panels compares exact maximum likelihood with Laplace approximation. The lower row of panels compares exact maximum likelihood with expectation propagation approximation. The horizontal lines indicate expectation propagation-based confidence intervals for 20 randomly chosen replications of the simulation study described in the text. The points indicate the corresponding approximate maximum likelihood estimates. The vertical lines indicate true parameter values. The percentages displayed at the top of each panel are empirical coverages over all 1000 replications for each method involved in the comparison.

4.1. Simulations

Our simulations involved (1) comparison with exact maximum likelihood for the $d^{\mathrm{R}} = 1$ situation for which quadrature is univariate, and (2) evaluation of inferential accuracy and speed for a larger model involving bivariate random effects.

4.1.1. Comparison with Exact Maximum Likelihood for Univariate Random Effects

Our first simulation study involved simulation of 1000 datasets according to the $d^{\mathrm{R}} = 1$ version of (1) with true parameter values:

$$\boldsymbol{\beta}_{\mathrm{true}} = [0, 1]^T \quad \text{and} \quad \Sigma_{\mathrm{true}} = \sigma^2_{\mathrm{true}} = 1. \tag{19}$$

The sample sizes were set to $m = 100$ and $n_i = 2$. The $\boldsymbol{x}^{\mathrm{F}}_{ij}$ and $\boldsymbol{x}^{\mathrm{R}}_{ij}$ vectors were of the form

$$\boldsymbol{x}^{\mathrm{F}}_{ij} = [1, x_{ij}]^T \quad \text{and} \quad \boldsymbol{x}^{\mathrm{R}}_{ij} = 1 \tag{20}$$

where x_{ij} was generated independently from a Uniform distribution on the unit interval.

For each simulated dataset, the probit mixed model defined by Equation (20) was fit using each of the following approaches:

1. Exact maximum likelihood with adaptive Gauss–Hermite quadrature used for the univariate intractable integrals. This was achieved using the function `glmer()` in the R package `lme4` (Bates et al. 2015). The number of points for evaluation of the adaptive Gauss–Hermite approximation was fixed at 100.
2. The Laplace approximation used by `glmer()`.
3. Expectation propagation as described in Section 3.
4. Data cloning as used by the R package `dclone`, with 10 clones and inference as described in Sólymos (2010).

Of interest is comparison of quadrature-free approximations (2)–(4) against the exact maximum likelihood benchmark. Figure 2 contrasts the point estimates and confidence intervals produced by Laplace approximation and expectation propagation against those produced by exact maximum likelihood. The first row of Figure 2 shows that Laplace approximation results in poor statistical inference, with the empirical coverage

values falling well below the advertized 95% level. The gray line segments for exact likelihood confidence intervals and black line segments for their Laplace approximations have very noticeable discrepancies. In the second row of Figure 2, we repeat the empirical coverage percentages and gray line segments for exact likelihood inference and, instead, compare these results with those produced by expectation propagation. For the fixed effects, β_0 and β_1, the empirical coverage of expectation propagation is seen to be very close to 95%. For the standard deviation parameter, σ, expectation propagation delivers slightly more coverage than advertized (97.5% versus 95%). However, the relatively low sample sizes in this study should be kept in mind. The simulation study in the next subsection uses higher sample sizes and expectation propagation is seen to be particularly accurate in terms of confidence interval coverage. The empirical coverage values for data cloning were 95.7%, 95.1%, and 97.4%. These are very close to those of expectation propagation.

Figure 3 compares the approaches via estimated mean squared error and mean squared error of prediction. The latter comparison involved randomly selecting 5 of the one hundred u_i random intercepts and recording their predictions for each approach. In Figure 3, we have also plotted corresponding t-based 95% confidence intervals, which provide an indication of the inherent variability of simulation-based mean squared error estimation. Expectation propagation is shown to perform well in comparison with exact maximum likelihood and best prediction, and generally improves upon Laplace approximation and data cloning for this particular yardstick.

Table 1 compares the computing times of the four approaches when run on a MacBook Air laptop with 8 gigabytes of random access memory and a 2.2 gigahertz processor. Even though such comparison necessarily is obscured by factors such as the computer language in which an approach is implemented, Table 1 provides a reasonable indication of computing times in practice. Laplace approximation and expectation propagation take less than a fifth of a second on average. Exact computation takes about 10–20 sec, with confidence interval construction for σ (not provided by `glmer()`) accounting for most of that time. Data cloning, with an average of about 2.4 minutes, is much slower than the three other approaches.

Table 1. Average (standard deviation) computing times in seconds for fitting and inference for the four approaches using in the first simulation study.

Exact	Laplace	Expec. propag.	Data cloning
16.10 (5.25)	0.158 (0.0163)	0.1960 (0.0198)	143 (4.38)

4.1.2. Accuracy and Speed Assessment for Bivariate Random Effects

In this study, we simulated 1000 datasets according to a $d^{\mathrm{R}} = 2$ version of (1) with true parameter values

$$\boldsymbol{\beta}_{\text{true}} = [0.37, 0.93, -0.46, 0.08, -1.34, 1.09]^T \quad \text{and}$$
$$\Sigma_{\text{true}} = \begin{bmatrix} 0.53 & -0.36 \\ -0.36 & 0.92 \end{bmatrix}. \tag{21}$$

The number of groups was fixed at $m = 250$ and each n_i value selected randomly from a discrete Uniform distribution on $\{20, 21, \ldots, 30\}$. The $\boldsymbol{x}_{ij}^{\mathrm{F}}$ and $\boldsymbol{x}_{ij}^{\mathrm{R}}$ vectors were of the form

$$\boldsymbol{x}_{ij}^{\mathrm{F}} = [1, x_{1,ij}, x_{2,ij}, x_{3,ij}, x_{4,ij}, x_{5,ij}]^T \quad \text{and} \quad \boldsymbol{x}_{ij}^{\mathrm{R}} = [1, x_{1,ij}]^T$$

where each $x_{k,ij}$ was generated independently from a Uniform distribution on the unit interval. All relative tolerance values were set to 10^{-5} and the maximum number of iteration values were set to 100, which is relevant for the upcoming speed assessment.

The points and horizontal line segments in Figure 4 are displays of estimates and corresponding 95% confidence intervals for each of the interpretable model parameters, for 50 randomly chosen replications. The numbers in the top right-hand

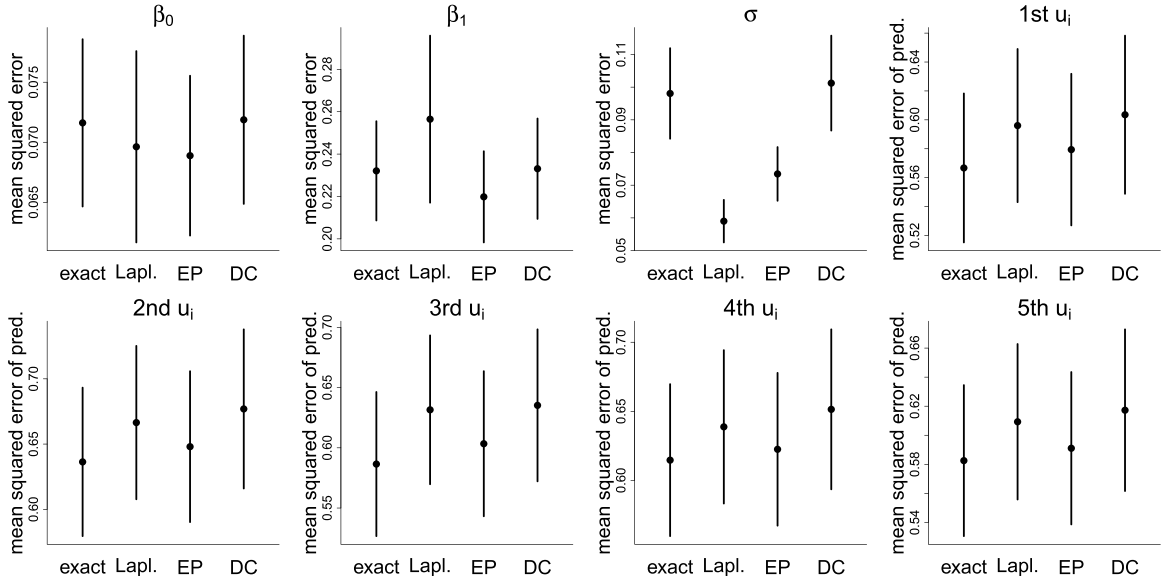

Figure 3. First three upper panels: Estimated mean squared errors for the parameters in $d^{\mathrm{R}} = 1$ simulation study with true parameter values given by Equation (19) for four different approaches: exact maximum likelihood (exact), Laplace approximation (Lapl.), expectation propagation (EP) and data cloning (DC). The estimates are the average squared error values over the 1000 replications in the simulation study. The vertical line segments indicate corresponding t-based 95% confidence intervals for the mean squared error. Remaining panels: the same as the first three panels but with mean squared errors of prediction for 5 randomly chosen u_i values.

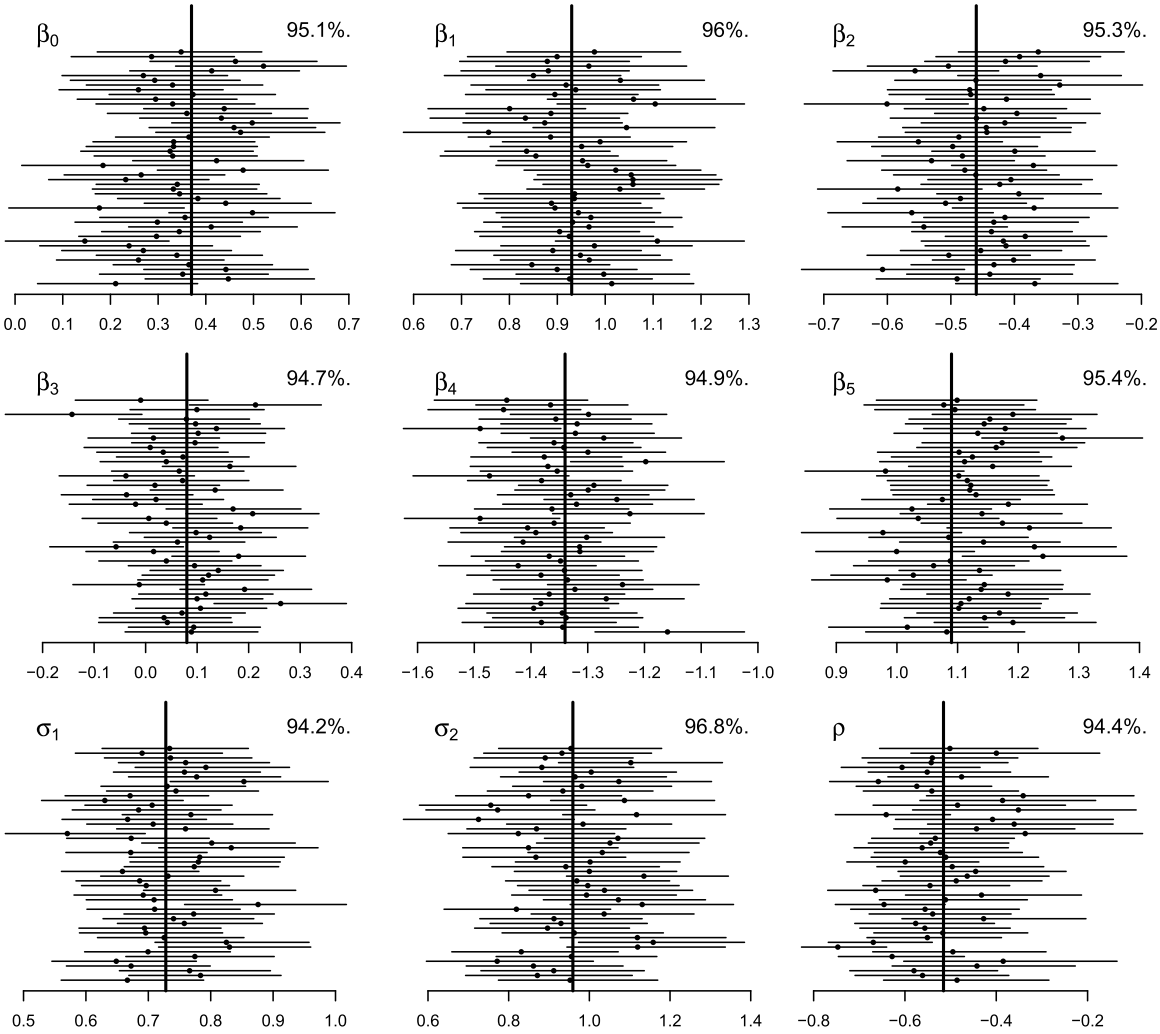

Figure 4. Summary of confidence interval coverage for the second simulation study with true parameter values given by Equation (21). The horizontal lines indicate expectation propagation-based confidence intervals for 50 randomly chosen replications of the simulation study described in the text. The solid circular points indicate the corresponding point estimates. The vertical lines indicate true parameter values. The percentage in the top right-hand corner of each panel is the empirical coverage over all 1000 replications.

corner of each panel are the empirical coverage values based on all 1000 replications. For all nine parameters, the empirical coverage values are in keeping with the advertized coverage of 95%, and is an indication of excellent accuracy for this setting.

Despite the higher samples and complexity of the model, we have reduced the fitting times to tens of seconds in the `glmmEP` package within the R computing environment. This has been achieved by implementation of Algorithm 1 in a low-level language so that approximate likelihood evaluations are very rapid. The computing speed depends upon various relative tolerance values and upper bounds on numbers of iterations for the various iterative schemes as well as attributes of the computer. This simulation study was run on a MacBook Air laptop with 8 gigabytes of random access memory and a 2.2 gigahertz processor. The convergence stopping criteria values are given earlier in this section. Over the 1000 replications the median computing time was 18 sec, the upper quartile was 20 sec and the maximum was 34 sec. Such speed is impressive given that each dataset contained tens of thousands of observations and bivariate random effects are accurately handled.

4.2. Application to Data from a Immunization Study

Data from a 1987 Guatemala childhood immunization study are stored in the data frame `guImmun` within the R package `mlmRev` (Bates, Maechler, and Bolker 2014). Rodríguez and Goldman (1995) presented details of the study and some multilevel analyses. Variables in the `guImmun` data frame include

`immun` a two-level factor variable indicating whether a child received a complete set of immunizations at the time of the survey, with levels `Y` for complete set and `N` for incomplete set,

`pcInd81` percentage of indigenous population in the community in which the child lived at the time of the 1981 census,

`kid2p` a two-level factor variable indicating whether or not the child was two years or older at the time of the survey, with levels `Y` for two years or older and `N` for younger than two years,

`momEd` a three-level factor variable indicating the mother's level of education, with levels `N` for not finished primary school, `P` for finished primary school and `S` for finished secondary school,

husEd a four-level factor variable indicating the husband's level of education, with levels N for not finished primary school, P for finished primary school, S for finished secondary school and U for unknown,

momWork a two-level factor variable indicating whether or not the child's mother had ever worked outside the home, with levels Y for worked outside of the home and N for never worked outside of the home,

rural a two-level factor variable indicating whether or not the child's location is considered rural or urban, with levels Y for rural and N for urban,

mom a multilevel factor variable that codes the children's mothers, out of 1,595 mothers in total.

A random intercepts and slopes probit mixed model for these data is

$$
\begin{aligned}
&I(\texttt{immun}_{ij} = \texttt{Y}) | u_{0i}, u_{1i} \overset{\text{ind.}}{\sim} \\
&\quad \text{Bernoulli}\Big(\Phi\big(\beta_0 + u_{0i} + (\beta_1 + u_{1i})\,\texttt{pcInd81}_{ij} \\
&\qquad + \beta_2\, I(\texttt{kid2p}_{ij} = \texttt{Y}) + \beta_3\, I(\texttt{momEd}_{ij} = \texttt{S}) \\
&\qquad + \beta_4\, I(\texttt{husEd}_{ij} = \texttt{S}) + \beta_5\, I(\texttt{momWork}_{ij} = \texttt{Y})\big) \\
&\qquad + \beta_6\, I(\texttt{rural}_{ij} = \texttt{Y})\Big)\Big) \qquad (22)
\end{aligned}
$$

where $I(\mathcal{P}) = 1$ if $\mathcal{P}$ is true and 0 otherwise. Also, $\texttt{immun}_{ij}$ denotes the value of immun for the jth child of the ith mother, $1 \le i \le 1,595$, with the other variables defined analogously. The bivariate random effects vectors are assumed to satisfy

$$
\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \overset{\text{ind.}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right). \qquad (23)
$$

The variables in Equation (22) were selected using a least absolute shrinkage selection operator (Tibshirani 1996) approach.

We fitted this model using our expectation propagation approximate likelihood inference scheme. It took about 10 seconds on the fourth author's MacBook Air laptop (2.2 gigahertz processor and 8 gigabytes of random access memory) to produce the inferential summary given in Table 2.

Table 2. Expectation propagation approximate maximum likelihood estimates and corresponding 95% confidence interval (C.I.) lower and upper limits for the parameters in model (22) and (23).

Parameter	95% C.I. low.	Estimate	95% C.I. upp.
β_0	−0.6711	−0.3373	−0.0035
β_1	−1.0783	−0.7663	−0.4543
β_2	0.7018	0.9291	1.1565
β_3	−0.4090	0.0653	0.5396
β_4	−0.3388	0.0523	0.4434
β_5	0.0531	0.2591	0.4650
β_6	−0.7895	−0.5345	−0.2795
σ_1	1.1622	1.5370	2.0328
σ_2	1.5407	2.5887	4.3494
ρ	−0.9486	−0.7821	−0.2766

With the exception of those involving parental education, each of the parameters is seen to be statistically significantly different from zero. As examples, the 95% confidence interval for β_1 of $(-1.08, -0.454)$ indicates a lower prevalence of immunization in communities with higher percentages of indigenous people and the 95% confidence interval for σ_2 of $(1.54, 4.35)$ shows that there is significant heterogeneity in the indigenous percentage effect across the 1,595 families.

Figure 5 provides a visual display of the fixed effects estimates and approximate 95% confidence intervals in Table 2. For comparison, we also include the results obtained from the default call to the glmer() function in the package lme4 (Bates et al. 2015), in which a Laplace approximation is used, data cloning via the package dclone (Sólymos 2010) with 10 clones and a Markov chain Monte Carlo fitting via the function stan() in the R language package rstan (Stan Development Team 2018). The last of these involves a Bayesian version of (22) with diffuse priors and therefore is close to likelihood-based

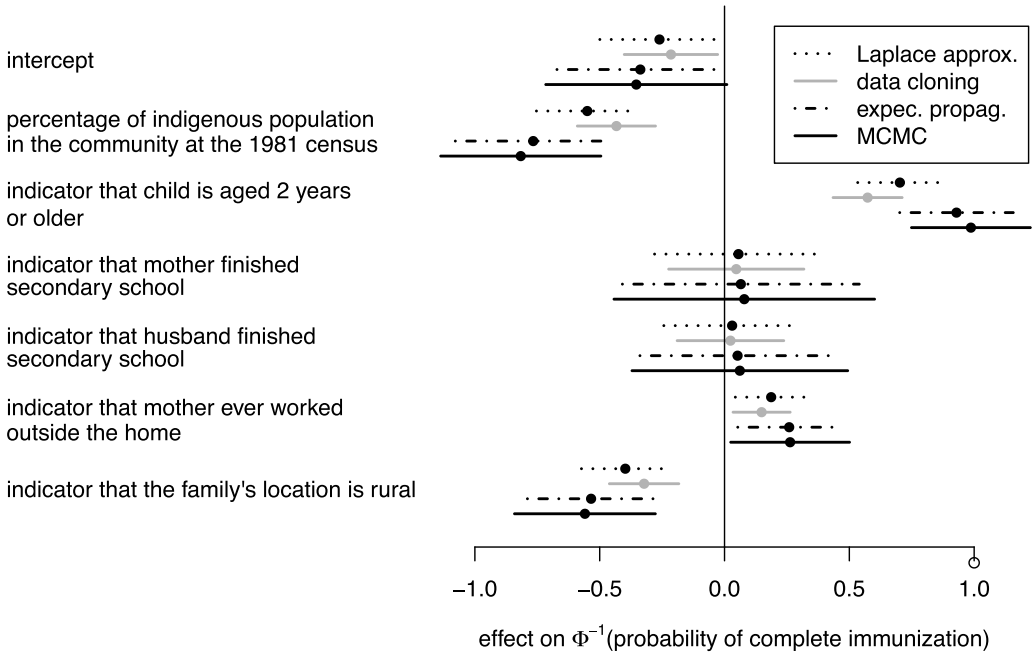

Figure 5. Visual comparison of approximate 95% confidence/credible intervals for $\beta_0, \ldots, \beta_6$ for three approaches to fitting the probit mixed model (22) to the Guatemala immunization data. The approaches are Laplace approximation, expectation propagation and Markov chain Monte Carlo (MCMC) with details given in the text.

inference. The actual diffuse priors are independent $N(0, 10^{10})$ distributions for $\beta_0, \ldots, \beta_6$ and a member of the marginally noninformative family of covariance matrix priors described in Huang and Wand (2013) for the 2×2 covariance matrix in (23). In the notation of Huang and Wand (2013) the hyperparameters were set to $\nu = 2$ and $A_1 = A_2 = 10^5$. A warm-up of size 200,000 was used followed by samples of size 10,000 retained for inference. Under the important and nontrivial assumption—perhaps plausible here—that the Markov chain Monte Carlo-based 95% credible intervals are close to the 95% confidence intervals based on exact maximum likelihood, Figure 5 shows good accuracy of expectation propagation. Laplace approximation is seen to lead to fixed effect estimates with considerable bias and reduced standard errors. Similar comments apply to the version of data cloning used here. We note that data cloning has quite a few tuning parameters such as the number of clones and prior hyperparameter values. Also, data cloning is also very slow for this example,taking about 3 hr on a contemporary laptop computer, making it difficult to assess sensitivity to tuning parameter choice. On the same laptop expectation propagation took only 12 seconds.

5. Theoretical Considerations

We now discuss the question regarding whether the excellent inferential accuracy of the Section 3 methodology is supported by theory. A fuller theoretical analysis is the subject of ongoing work involving the first four authors and, upon completion, will be reported elsewhere. In this section, we provide a heuristic explanation for the accuracy of expectation propagation in the binary response mixed model context.

First note that the ith log-likelihood summand is

$$\ell_i(\boldsymbol{\beta}, \Sigma) = \log \int_{\mathbb{R}^{d^R}} \left\{ \prod_{j=1}^{n_i} \frac{p(\boldsymbol{y}_j|\boldsymbol{u}_i; \boldsymbol{\beta})}{\widetilde{p}(\boldsymbol{y}_j|\boldsymbol{u}_i; \boldsymbol{\beta})} \right\} \times \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i \boldsymbol{u}_i^T) \end{bmatrix}^T \underset{\sim}{\widehat{\boldsymbol{\eta}}}_i \right\} d\boldsymbol{u}_i$$

where $\widetilde{p}(\boldsymbol{y}_j|\boldsymbol{u}_i; \boldsymbol{\beta})$ is given by expression (7) with the $\boldsymbol{\eta}_{ij}$ set to the converged $\boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}$ values. We also have

$$\underset{\sim}{\ell}_i(\boldsymbol{\beta}, \Sigma) = \log \int_{\mathbb{R}^{d^R}} \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i \boldsymbol{u}_i^T) \end{bmatrix}^T \underset{\sim}{\widehat{\boldsymbol{\eta}}}_i \right\} d\boldsymbol{u}_i.$$

Now make the change of variables

$$\boldsymbol{v} = \Delta_i^{-1}\{\boldsymbol{u}_i - \underset{\sim}{\text{BP}}(\boldsymbol{u}_i)\} \quad \text{where} \quad \Delta_i \equiv \underset{\sim}{\text{cov}}(\boldsymbol{u}_i|\boldsymbol{y})^{1/2}$$

involving the expectation propagation-approximate best predictor quantities given by Equations (17) and (18). Straightforward manipulations then lead to the discrepancy between $\ell_i(\boldsymbol{\beta}, \Sigma)$ and $\underset{\sim}{\ell}_i(\boldsymbol{\beta}, \Sigma)$ equalling

$$\ell_i(\boldsymbol{\beta}, \Sigma) - \underset{\sim}{\ell}_i(\boldsymbol{\beta}, \Sigma) = \log \int_{\mathbb{R}^{d^R}} \left\{ \prod_{j=1}^{n_i} A_{ij}(\Delta_i \boldsymbol{v}) \right\} \phi_I(\boldsymbol{v})\, d\boldsymbol{v} \quad (24)$$

where, for any $\boldsymbol{x} \in \mathbb{R}^{d^R}$, $\phi_I(\boldsymbol{x}) \equiv (2\pi)^{-d^R/2} \exp(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x})$ and

$$A_{ij}(\boldsymbol{x}) \equiv F\Big((2y_{ij} - 1)\big(\boldsymbol{\beta}^T \boldsymbol{x}_{ij}^{\text{F}} + (\underset{\sim}{\text{BP}}(\boldsymbol{u}_i) + \boldsymbol{x})^T \boldsymbol{x}_{ij}^{\text{R}}\big)\Big) \times \exp\left\{ -\begin{bmatrix} 1 \\ \underset{\sim}{\text{BP}}(\boldsymbol{u}_i) + \boldsymbol{x} \\ \text{vech}\Big(\big(\underset{\sim}{\text{BP}}(\boldsymbol{u}_i) + \boldsymbol{x}\big)\big(\underset{\sim}{\text{BP}}(\boldsymbol{u}_i) + \boldsymbol{x}\big)^T\Big) \end{bmatrix}^T \times \boldsymbol{\eta}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i} \right\}.$$

Using the same change of variables, the moment-matching conditions corresponding to the Kullback–Leibler projection (8) are

$$\int_{\mathbb{R}^{d^R}} \boldsymbol{v}^{\otimes k} A_{ij}(\Delta_i \boldsymbol{v}) \phi_I(\boldsymbol{v})\, d\boldsymbol{v} = \int_{\mathbb{R}^{d^R}} \boldsymbol{v}^{\otimes k} \phi_I(\boldsymbol{v})\, d\boldsymbol{v}, \quad k = 0, 1, 2, \quad (25)$$

where $\boldsymbol{v}^{\otimes 0} \equiv 1$, $\boldsymbol{v}^{\otimes 1} \equiv \boldsymbol{v}$ and $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$.

To aid intuition, for the remainder of this section, we restrict attention to $d^R = 1$ and write δ_i instead of Δ_i to signify the fact that this quantity is scalar in this special case. Next, we make the

$$\text{working assumption:} \quad \delta_i = O_p(n_i^{-1/2}). \quad (26)$$

This assumption is in keeping with the fact that δ_i is the expectation propagation approximation to the empirical standard deviation of $\underset{\sim}{\text{BP}}(\boldsymbol{u}_i) - \boldsymbol{u}_i$. Then Taylor series expansion of A_{ij} about zero and substitution into the $d^R = 1$ version of (25) leads to

$$A_{ij}(0) = 1 + O(\delta_i^4), \quad A'_{ij}(0) = O(\delta_i^2) \quad \text{and} \quad A''_{ij}(0) = O(\delta_i^2).$$

Plugging these into Equation (24) and using $\log(1 + \varepsilon) \approx \varepsilon$ for small ε we obtain

$$\ell_i(\boldsymbol{\beta}, \Sigma) - \underset{\sim}{\ell}_i(\boldsymbol{\beta}, \Sigma) = O_p(n_i^{-1/2}) \quad \text{under (26).}$$

These heuristics suggest that expectation propagation provides consistent estimation of the log-likelihood summands as the number of measurements in the ith group increases. The deeper question concerning the asymptotic statistical properties of the expectation propagation-based estimators $(\underset{\sim}{\widehat{\boldsymbol{\beta}}}, \underset{\sim}{\widehat{\Sigma}})$ requires more delicate theoretical analysis. As mentioned earlier in this section, this question is being pursued by authors of this article.

Before closing this section, we mention that there is a small but emerging body of research concerning the large sample behavior of expectation propagation for approximation Bayesian inference. A recent contribution of this type is Dehaene and Barthelmé (2018) which provides Bernstein–von Mises theory for Bayesian expectation propagation.

6. Higher Level and Crossed Random Effects Extensions

The binary mixed model given by (1) is adequate for the common situation of there being only one grouping mechanism. However, more elaborate models are required for situations such as hierarchical and cross-tabulated grouping mechanisms. Goldstein (2010), for example, provides an extensive treatment of mixed models with higher levels of nesting. A major reference

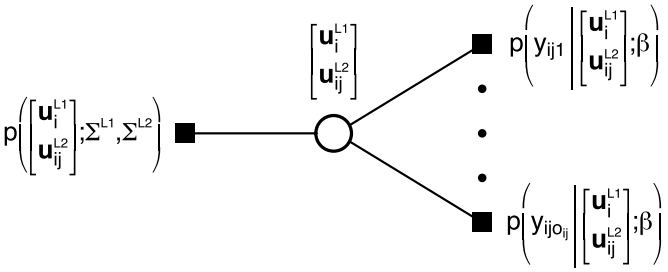

Figure 6. Factor graph representation of the product structure of the integrand in Equation (6). The open circle corresponds to the random effect vector $[\boldsymbol{u}_i^{\mathrm{L1}}\ \boldsymbol{u}_{ij}^{\mathrm{L2}}]^T$ and the solid rectangles indicate factors in the integrand of Equation (28). Edges indicate dependence of each factor on $[\boldsymbol{u}_i^{\mathrm{L1}}\ \boldsymbol{u}_{ij}^{\mathrm{L2}}]^T$.

for crossed random effects mixed models is Baayen, Davidson, and Bates (2008). Here, we provide advice regarding extension our expectation propagation approach to these settings.

The *two levels of nesting* extension of (1) is

$$\begin{aligned} y_{ijk}|\boldsymbol{u}_i^{\mathrm{L1}},\boldsymbol{u}_{ij}^{\mathrm{L2}} &\overset{\text{ind.}}{\sim} \\ &\text{Bernoulli}\Big(F\big(\boldsymbol{\beta}^T\boldsymbol{x}_{ijk}^{\mathrm{F}}+(\boldsymbol{u}_i^{\mathrm{L1}})^T\boldsymbol{x}_{ijk}^{\mathrm{R1}}+(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\boldsymbol{x}_{ijk}^{\mathrm{R2}}\big)\Big), \\ &\boldsymbol{u}_i^{\mathrm{L1}} \overset{\text{ind.}}{\sim} N(\boldsymbol{0},\Sigma^{\mathrm{L1}}) \text{ independently of } \boldsymbol{u}_{ij}^{\mathrm{L2}} \overset{\text{ind.}}{\sim} N(\boldsymbol{0},\Sigma^{\mathrm{L2}}), \\ &1\le i\le m,\quad 1\le j\le n_i,\quad 1\le k\le o_{ij}. \end{aligned} \tag{27}$$

The response y_{ijk} and predictor vectors $\boldsymbol{x}_{ijk}^{\mathrm{F}}$, $\boldsymbol{x}_{ijk}^{\mathrm{R1}}$ and $\boldsymbol{x}_{ijk}^{\mathrm{R2}}$ correspond to the kth set of measurements within the jth inner group within the ith outer group. The number of outer groups is m and the number of inner groups in the ith outer group is n_i. The sample size of the jth group in the ith outer group is o_{ij}. Also, $\boldsymbol{x}_{ijk}^{\mathrm{R1}}$ is $d^{\mathrm{R1}}\times 1$ and $\boldsymbol{x}_{ijk}^{\mathrm{R2}}$ is $d^{\mathrm{R2}}\times 1$. The log-likelihood of $(\boldsymbol{\beta},\Sigma^{\mathrm{L1}},\Sigma^{\mathrm{L2}})$ may be written as

$$\begin{aligned} \ell(\boldsymbol{\beta},\Sigma^{\mathrm{L1}},\Sigma^{\mathrm{L2}}) &= \sum_{i=1}^{m}\log\int_{\mathbb{R}^{d^{\mathrm{R}}}}\prod_{j=1}^{n_i}\prod_{k=1}^{o_{ij}} p\left(y_{ijk}\middle|\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{bmatrix};\boldsymbol{\beta}\right) \\ &\quad\times p\left(\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{bmatrix};\Sigma^{\mathrm{L1}},\Sigma^{\mathrm{L2}}\right)d\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{bmatrix} \end{aligned} \tag{28}$$

where $d^{\mathrm{R}}=d^{\mathrm{R1}}+d^{\mathrm{R2}}$,

$$\begin{aligned} p\left(y_{ijk}\middle|\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{bmatrix};\boldsymbol{\beta}\right) &\equiv F\Big((2y_{ijk}-1)\big(\boldsymbol{\beta}^T\boldsymbol{x}_{ijk}^{\mathrm{F}}+(\boldsymbol{u}_i^{\mathrm{L1}})^T\boldsymbol{x}_{ijk}^{\mathrm{R1}} \\ &\quad+(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\boldsymbol{x}_{ijk}^{\mathrm{R2}}\big)\Big),\ y_{ijk}=0,1, \end{aligned}$$

and

$$\begin{aligned} &p\left(\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{bmatrix};\Sigma^{\mathrm{L1}},\Sigma^{\mathrm{L2}}\right) \\ &\quad\equiv|2\pi\Sigma^{\mathrm{L1}}|^{-1/2}|2\pi\Sigma^{\mathrm{L2}}|^{-1/2} \\ &\qquad\times\exp\left\{-\frac{1}{2}\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{bmatrix}^T\begin{bmatrix}\Sigma^{\mathrm{L1}} & \boldsymbol{0}\\ \boldsymbol{0} & \Sigma^{\mathrm{L2}}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{bmatrix}\right\}. \end{aligned}$$

Expectation propagation approximation of $\ell(\boldsymbol{\beta},\Sigma^{\mathrm{L1}},\Sigma^{\mathrm{L2}})$ then proceeds by message passing on the factor graph displayed in Figure 6. In the probit case, Theorem 1 can be called upon to obtain closed-form updates for the message natural parameter vectors leading to an algorithm analogous to Algorithm 1.

A *crossed random effects* extension of (1) is

$$\begin{aligned} y_{ii'j}|\boldsymbol{u}_i,\boldsymbol{u}'_{i'} &\overset{\text{ind.}}{\sim} \text{Bernoulli}\Big(F\big(\boldsymbol{\beta}^T\boldsymbol{x}_{ii'j}^{\mathrm{F}}+(\boldsymbol{u}_i)^T\boldsymbol{x}_{ii'j}^{\mathrm{R}}+(\boldsymbol{u}'_{i'})^T\boldsymbol{x}_{ii'j}^{\mathrm{R}\prime}\big)\Big), \\ &\boldsymbol{u}_i \overset{\text{ind.}}{\sim} N(\boldsymbol{0},\Sigma) \text{ independently of } \boldsymbol{u}'_{i'} \overset{\text{ind.}}{\sim} N(\boldsymbol{0},\Sigma'), \\ &1\le i\le m,\quad 1\le i'\le m',\quad 1\le j\le n_{ii'} \end{aligned} \tag{29}$$

where the data are cross-tabulated according to membership of two groups of sizes m and m' indexed according to the pair $(i,i')\in\{1,\dots,m\}\times\{1,\dots,m'\}$, with $n_{ii'}$ denoting the sample size within group (i,i'). Note that $n_{ii'}=0$ is a possibility for some (i,i'). The response $y_{ii'j}$ and the predictor vectors $\boldsymbol{x}_{ii'j}^{\mathrm{F}}$, $\boldsymbol{x}_{ii'j}^{\mathrm{R}}$ and $\boldsymbol{x}_{ii'j}^{\mathrm{R}\prime}$ correspond to the jth set of measurements within group (i,i'). The $\boldsymbol{u}_i$, $1\le i\le m$, are $d^{\mathrm{R}}\times 1$ random effects for group-specific departures from the fixed effects for the first group. The $\boldsymbol{u}'_i$, $1\le i\le m'$, are $d^{\mathrm{R}\prime}\times 1$ random effects for group-specific departures from the fixed effects for the second group. The log-likelihood of $(\boldsymbol{\beta},\Sigma,\Sigma')$ is

$$\begin{aligned} &\ell\big(\boldsymbol{\beta},\Sigma,\Sigma'\big) \\ &\quad=\log\int_{\mathbb{R}^{md^{\mathrm{R}}+m'd^{\mathrm{R}\prime}}}\left\{\prod_{(i,i'):n_{ii'}>0}\prod_{j=1}^{n_{ii'}}p\left(y_{ii'j}\middle|\boldsymbol{u}_i,\boldsymbol{u}'_{i'};\boldsymbol{\beta}\right)\right\} \\ &\quad\times p\left(\boldsymbol{u},\boldsymbol{u}';\Sigma,\Sigma'\right)d\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{u}'\end{bmatrix} \end{aligned} \tag{30}$$

where

$$\begin{aligned} &p\left(y_{ii'j}\middle|\boldsymbol{u}_i,\boldsymbol{u}'_{i'};\boldsymbol{\beta}\right) \\ &\quad=F\Big((2y_{ii'j}-1)\big(\boldsymbol{\beta}^T\boldsymbol{x}_{ii'j}^{\mathrm{F}}+(\boldsymbol{u}_i)^T\boldsymbol{x}_{ii'j}^{\mathrm{R}}+(\boldsymbol{u}'_{i'})^T\boldsymbol{x}_{ii'j}^{\mathrm{R}\prime}\big)\Big), \end{aligned}$$

$\boldsymbol{u}\equiv[\boldsymbol{u}_1^T\cdots\boldsymbol{u}_m^T]^T$, $\boldsymbol{u}'\equiv[\boldsymbol{u}_1'^T\cdots\boldsymbol{u}_{m'}'^T]^T$ and

$$\begin{aligned} &p\big(\boldsymbol{u},\boldsymbol{u}';\Sigma,\Sigma'\big) \\ &\quad\text{is the } N\left(\begin{bmatrix}\boldsymbol{0}_{md^{\mathrm{R}}}\\ \boldsymbol{0}_{m'd^{\mathrm{R}\prime}}\end{bmatrix},\begin{bmatrix}\boldsymbol{I}_m\otimes\Sigma & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}_{m'}\otimes\Sigma'\end{bmatrix}\right) \\ &\quad\text{density function.} \end{aligned}$$

Likelihood-based inference for the $(\boldsymbol{\beta},\Sigma,\Sigma')$ is particularly challenging for crossed random effects since the dimensions of the intractable integrals grow with the number of groups. See, for example, Section 3.3 of Jiang (2017) for discussion about some of the challenges that arise in asymptotic analysis for generalized linear mixed models with crossed random effects.

Expectation propagation approximation of $\ell(\boldsymbol{\beta},\Sigma,\Sigma')$ can be carried out via message passing on the factor graph given in Figure 7. A message passing algorithm analogous to Algorithm 1 results.

We tested the viability of expectation propagation for a $d^{\mathrm{R}}=d^{\mathrm{R}\prime}=1$ version of (29) with sample sizes $m=10$, $m'=6$ and $n_{ii'}=3$ for all (i,i') pairs. The true parameter values were

$$\boldsymbol{\beta}_{\text{true}}=[-0.58,1.07]^T,\quad \Sigma_{\text{true}}=0.32\quad\text{and}\quad\Sigma'_{\text{true}}=0.47.$$

with predictor data such that the first entry of the 2×1 vector $\boldsymbol{x}_{ii'j}^{\mathrm{F}}$ was 1 and the second entry was a uniform random number between 0 and 1. One thousand replications were simulated and

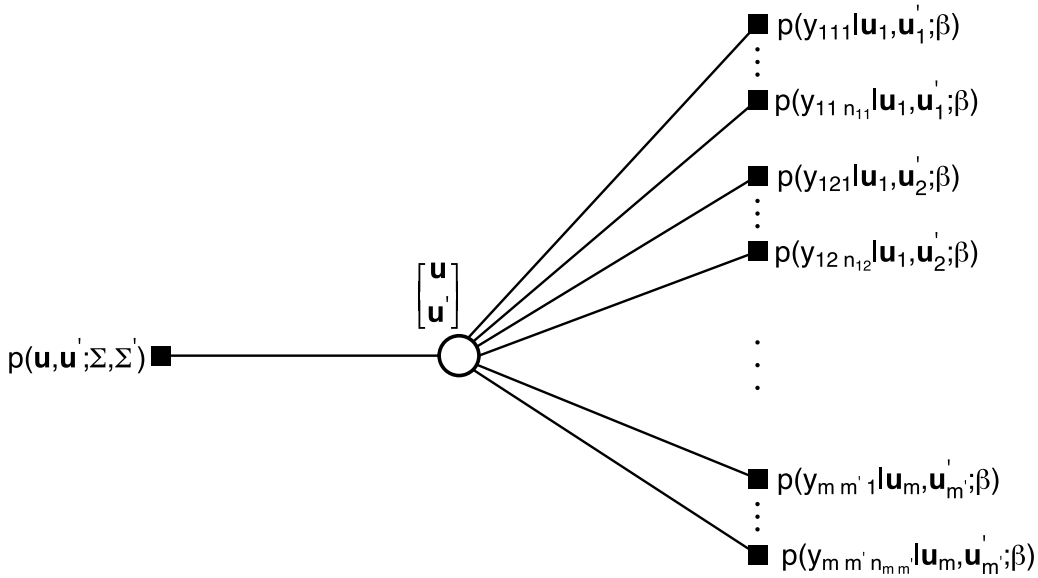

Figure 7. Factor graph representation of the product structure of the integrand in Equation (30). The open circle corresponds to the random effect vector $[\boldsymbol{u}^T\ (\boldsymbol{u}')^T]^T$ and the solid rectangles indicate factors in the integrand of Equation (30). Edges indicate dependence of each factor on $[\boldsymbol{u}^T\ (\boldsymbol{u}')^T]^T$.

95% confidence intervals for the four model parameters were obtained. The empirical coverage for the fixed effects intercept was 92.6%, whilst that for the slope was 94.4%. The empirical coverage for the two variance parameters Σ and Σ' were 63.1% and 89.2%, respectively. More research is required to assess the general viability of expectation propagation for crossed random effects generalized linear mixed models.

7. Transferral to Other Mixed Models

Until now we have mainly focused on the special case of probit mixed models with Gaussian random effects since the requisite Kullback–Leibler projections have closed form solutions. However, our approach is quite general and, at least in theory, applies to other mixed models. We now briefly describe transferral to other mixed models.

7.1. Logistic Mixed Models

As we mention in Section 2, the probit and logistic cases are distinguished according to whether $F = \Phi$ or $F =$ expit. Therefore, transferral from probit to logistic mixed models involves replacement of f_{input} in Theorem 1 by

$$f_{\text{input}}(\boldsymbol{x}) = \text{expit}(c_0 + \boldsymbol{c}_1^T\boldsymbol{x}) \exp\left\{\begin{bmatrix} \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\eta}_1^{\text{input}} \\ \boldsymbol{\eta}_2^{\text{input}} \end{bmatrix}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{31}$$

In view of Lemma 1 of the online supplement, Kullback–Leibler projection of f_{input} onto the unnormalized Normal family involves univariate integrals of the form

$$\int_{-\infty}^{\infty} x^p \exp\{qx - rx^2 - \log(1 + e^x)\}\, dx, \quad p = 0, 1, 2,\ q \in \mathbb{R},\ r > 0. \tag{32}$$

In the Bayesian context, Gelman et al. (2014; sec. 13.8) and Kim and Wand (2018) describe quadrature-based approaches to evaluation of Equation (32), each of which transfers to the frequentist context dealt with here. However, there is a significant speed cost compared with the probit case.

Details on the mechanics and performance of expectation propagation for logistic mixed models are reported in Yu (2020).

7.2. Other Generalized Linear Mixed Models

Whilst we have focused on the binary response situation in this article, we quickly point out that the principles apply to other generalized linear mixed models such as those based on the Gamma and Poisson families. Note that (2) with $F =$ expit generalizes to

$$\ell(\boldsymbol{\beta}, \Sigma) = \sum_{i=1}^{m} \log \int_{\mathbb{R}^{d^R}} \left[\prod_{j=1}^{n_i} \exp\left\{ y_{ij}(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}} + \boldsymbol{u}^T\boldsymbol{x}_{ij}^{\text{R}}) - b\big(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}} + \boldsymbol{u}^T\boldsymbol{x}_{ij}^{\text{R}}\big) + c(y_{ij}) \right\} \right] \times |2\pi\,\Sigma|^{-1/2} \exp(-\tfrac{1}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u})\, d\boldsymbol{u}$$

where the functions b and c are as given in Table 2.1 of McCullagh and Nelder (1989). Setting $b(x) = \log(1+e^x)$ and $c(x) = 0$ gives the $F =$ expit logistic mixed model while putting $b(x) = e^x$ and $c(x) = -\log(x!)$ gives the corresponding Poisson mixed model. The family of integrals

$$\int_{-\infty}^{\infty} x^p \exp\{qx - rx^2 - b(x)\}\, dx, \quad p = 0, 1, 2,\ q \in \mathbb{R},\ r > 0,$$

is required to facilitate the required Kullback-Leibler projections. Yu (2020) contains a detailed account of the practicalities and performance of expectation propagation for this class of models.

Acknowledgments

We are grateful for assistance from Jim Booth, Omar Ghattas, Alan Huang, Subhash Lele and Peter Sólymos on aspects of this research. We are also thankful for comments from an associate editor and two referees.

Funding

This research was supported by Australian Research Council Discovery Project DP180100597. I. M. Johnstone was supported by National Science Foundation Division of Mathematical Sciences grant 1811614 and National Institutes of Health grant R01 EB001988.

ORCID

I. M. Johnstone https://orcid.org/0000-0002-2865-3076
J. T. Ormerod https://orcid.org/0000-0002-4650-7507
M. P. Wand https://orcid.org/0000-0003-2555-896X
J. C. F. Yu https://orcid.org/0000-0003-3583-6564

References

Azzalini, A. (2017), "The R Package `sn`: The Skew-normal and Skew-t Distributions (version 1.5)," available at *http://azzalini.stat.unipd.it/SN* [4]

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008), "Mixed-effects Modeling With Crossed Random Effects for Subjects and Items," *Journal of Memory and Language*, 59, 390–412. [13]

Bates, D., Maechler, M., and Bolker, B. (2014), "`mlmRev`: Examples From Multilevel Modelling Software Review," R package version 1.0. Available at *http://cran.r-project.org*. [10]

Baltagi, B. H. (2013), *Econometric Analysis of Panel Data* (5th ed.), Chichester: Wiley. [1]

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015), "Fitting Linear Mixed-effects Models Using `lme4`," *Journal of Statistical Software*, 67, 1–48. [1,5,8,11]

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer. [1]

Broyden, C. G. (1970), "The Convergence of a Class of Double-rank Minimization Algorithms," *Journal of the Institute of Mathematics and Its Applications*, 6, 76–90. [6]

Carlin, B. P., and Gelfand, A. E. (1991), "A Sample Reuse Method for Accurate Parametric Empirical Bayes Confidence Intervals," *Journal of the Royal Statistical Society*, Series B, 53, 189–200. [7]

Dehaene, G., and Barthelmé, S. (2018), "Expectation Propagation in the Large-data Limit," *Journal of the Royal Statistical Society*, Series B, 80, 199–217. [12]

Diggle, P., Heagerty, P., Liang, K.-L., and Zeger, S. (2002), *Analysis of Longitudinal Data* (2nd ed.), Oxford: Oxford University Press. [1]

Fletcher, R. (1970), "A New Approach to Variable Metric Algorithms," *Computer Journal*, 13, 317–322. [6]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B.,Vehtari, A., and Rubin, D. B. (2014), *Bayesian Data Analysis* (3rd ed.), Boca Raton, Florida: CRC Press. [14]

Gelman, A., and Hill, J. (2007), *Data Analysis using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press. [1]

Givens, G. H., and Hoetig, J. A. (2005), *Computational Statistics*. Hoboken, NJ: Wiley. [6]

Goldfarb, D. (1970), "A Family of Variable Metric Updates Derived by Variational Means," *Mathematics of Computation*, 24, 23–26. [6]

Goldstein, H. (2010), *Multilevel Statistical Models* (4th ed.), Chichester: Wiley. [1,12]

Huang, A., and Wand, M. P. (2013), "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Analysis*, 8, 439–452. [12]

Jeon, M., Rijmen, F., and Rabe-Hesketh, S. (2017), "A Variational Maximization–Maximization Algorithm for Generalized Linear Mixed Models With Crossed Random Effects," *Psychometrika*, 82, 693–716. [2]

Jiang, J. (2017). *Asymptotic Analysis of Mixed Effects Models*, Boca Raton, FL: CRC Press. [2,13]

Kim, A. S. I., and Wand, M. P. (2018), "On Expectation Propagation for Generalised, Linear and Mixed Models," *Australian and New Zealand Journal of Statistics*, 60, 75–102. [14]

Lele, S.R., Nadeem, K., and Schmuland, B. (2010), "Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning," *Journal of the American Statistical Association*, 105, 1617–1625. [2]

The Mathworks Incorporated (2018), Natick, MA, USA. [6]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall. [14]

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models* (2nd ed.). New York: Wiley. [2,7]

Minka, T. P. (2001), "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, eds. J.S. Breese and D. Koller, pp. 362–369. Burlington, MA: Morgan Kaufmann. [1]

Minka, T. (2005), "Divergence Measures and Message Passing," Microsoft Research Technical Report Series, MSR-TR-2005-173, 1–17. [3,4,5]

Minka, T., and Winn, J. (2008), "Gates: A Graphical Notation for Mixture Models," Microsoft Research Technical Report Series, MSR-TR-2008-185, 1–16. [5]

Nelder, J. A., and Mead, R. (1965), "A Simplex Method for Function Minimization," *Computer Journal*, 7, 308–313. [3]

Ogden, H. E. (2015), "A Sequential Reduction Method for Inference in Generalized Linear Mixed Models," *Electronic Journal of Statistics*, 9, 135–152. [2]

Pinheiro, J. C., and Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*. New York: Springer. [6]

R Core Team (2019), R: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: *https://www.R-project.org/*. [1,2]

Rao, J. N. K., and Molina, I. (2015), *Small Area Estimation* (2nd ed.), Hoboken, NJ: Wiley. [1]

Rodríguez, G., and Goldman, N. (1995), "An Assessment of Estimation Procedures for Multilevel Models With Binary Responses," *Journal of the Royal Statistical Society*, Series A, 158, 73–89. [10]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations (with Discussion)," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [1]

Shanno, D. F. (1970), "Conditioning of Quasi-Newton Methods for Function Minimization," *Mathematics of Computation*, 24, 647–656. [6]

Sólymos, P. (2010), "`dclone`: Data Cloning in R," *The R Journal*, 2/2, 29–37. [8,11]

`Stan` Development Team (2018), "`RStan`: the R interface to `Stan`," R package version 2.18.2. Available at: *https://mc-stan.org/*. [11]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [11]

Wand, M. P., and Ormerod, J. T. (2012), "Continued Fraction Enhancement of Bayesian Computing," *Statistics*, 1, 31–41. [2,4]

Wand, M. P., and Yu, J. C. F. (2019), "`glmmEP`: Fast and Accurate Likelihood-based Inference in Generalized Linear Mixed Models Via Expectation Propagation," R package version 1.0. Available at: *http://cran.r-project.org*. [2]

Yu, J. C. F. (2020), "Fast and Accurate Frequentist Generalized Linear Mixed Model Analysis Via Expectation Propagation," Doctor of Philosophy thesis, University of Technology Sydney. [14]