Fast and Accurate Binary Response Mixed Model Analysis via Expectation Propagation

BY P. HALL[1], I.M. JOHNSTONE[2], J.T. ORMEROD[3], M.P. WAND[4] AND J.C.F. YU[4]

[1]*University of Melbourne*, [2]*Stanford University*, [3]*University of Sydney* *and* [4]*University of Technology Sydney*

16th August, 2018

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 theory 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.

Keywords: Best prediction; Generalized linear mixed models; Maximum likelihood; Kullback-Leibler projection; 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 & 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 & 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, 2018) 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.*, Section 14.4). An embellishment of Laplace approximation, known as integrated nested Laplace approximation (Rue, Martino & Chopin, 2009), has been successful in various Bayesian inference contexts.

Expectation propagation (e.g. Minka, 2001) is 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 towards approximate inference for Bayesian graphical models (e.g. Chapter 10, Bishop, 2006). 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.

To facilitate widespread use of the new approach, a new package in the R language (R Core Team, 2018) has been launched. The package, glmmEP (Wand & Yu, 2018), 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 summarised in Section 5. Higher level and random effects extensions are touched upon in Section 6. Lastly, we briefly discuss transferral of 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

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

where F, the *inverse link*, is a pre-specified cumulative distribution function and y_{ij} is the jth response for the ith group, where number of groups is m and the number of responses 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 $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ 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

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

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

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

The d^{R}-dimensional integrals in the $\ell(\boldsymbol{\beta},\boldsymbol{\Sigma})$ and $\text{BP}(\boldsymbol{u}_i)$ expressions cannot be reduced further and multivariate numerical integration must be called upon for their evaluation. In addition, $\ell(\boldsymbol{\beta},\boldsymbol{\Sigma})$ 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. Lastly, there is the problem of obtaining approximate confidence intervals for the entries of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ 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 & Neuhaus (2008) provides a summary of the main approaches up until the mid-2000s. Some more recent contributions, loosely related to the approach proposed here, include Jeon, Rijmen & Rabe-Hesketh (2017), Ogden (2015) and Wand & Ormerod (2012). 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 $\ell(\boldsymbol{\beta},\boldsymbol{\Sigma})$. Approximation of $\text{BP}(\boldsymbol{u}_i)$ follows relatively quickly. First note that $\ell(\boldsymbol{\beta},\boldsymbol{\Sigma}) = \sum_{i=1}^{m} \ell_i(\boldsymbol{\beta},\boldsymbol{\Sigma})$ where

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

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

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

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 approximating density functions of the same order of magnitude and, together with the properties of minimum Kullback-Leibler divergence, leads to accurate and statistically consistent approximation of $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. In probit case, where $F = \Phi$, the minimum Kullback-Leibler divergence steps are explicit. This leads to accurate approximation of $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ without the need for any numerical integration – just some fixed-point iteration. The expectation propagation-approximate log-likelihood, which we denote by $\underset{\sim}{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, can be evaluated quite rapidly and maximized using established derivative-free methods such as the Nelder-Mead algorithm (Nelder & 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 each of 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 $\mathrm{vec}(\boldsymbol{A})$ 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, $\mathrm{vech}(\boldsymbol{A})$ is $\frac{1}{2}\,d(d+1)+1$ vector defined similarly to $\mathrm{vec}(\boldsymbol{A})$ 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

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

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

$$\mathrm{KL}(f_1\|f_2) = \int_{\mathbb{R}^d} \big[f_1(\boldsymbol{x})\log\{f_1(\boldsymbol{x})/f_2(\boldsymbol{x})\} + f_2(\boldsymbol{x}) - f_1(\boldsymbol{x})\big]\,d\boldsymbol{x} \tag{3}$$

(e.g. Minka, 2005). In the special case where f_1 and f_2 are density functions the right-hand side of (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

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

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:

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

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

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

where

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

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 case of probit mixed models, 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

$$\boldsymbol{A}_2 \equiv \text{vec}^{-1}(\boldsymbol{D}_d^{+T} \boldsymbol{a}_2), \quad r_1 \equiv \sqrt{2(2 - \boldsymbol{c}_1^T \boldsymbol{A}_2^{-1} \boldsymbol{c}_1)}, \quad r_2 \equiv (2c_0 - \boldsymbol{c}_1^T \boldsymbol{A}_2^{-1} \boldsymbol{a}_1)/r_1,$$

$$r_3 \equiv 2\zeta'(r_2)/r_1, \quad r_4 \equiv -2\zeta''(r_2)/r_1^2 \quad \textit{and} \quad \boldsymbol{R}_5 \equiv \left(\boldsymbol{A}_2 + r_4 \boldsymbol{c}_1 \boldsymbol{c}_1^T\right)^{-1} \boldsymbol{A}_2$$

and the function A_N is given by

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

In addition, for primary arguments $\boldsymbol{a}_1$,$\boldsymbol{b}_1$ $(d \times 1)$ and $\boldsymbol{a}_2, \boldsymbol{b}_2$ $(\frac{1}{2}\, d(d+1) \times 1)$ 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

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

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 & Ormerod, 2012).

Expectation propagation for probit mixed models relies heavily upon:

Theorem 1. *If*

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

then

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

where

$$\left[\begin{array}{c}\boldsymbol{\eta}_1^*\\ \boldsymbol{\eta}_2^*\end{array}\right] = K_{\text{probit}}\left(\left[\begin{array}{c}\boldsymbol{\eta}_1^{\text{input}}\\ \boldsymbol{\eta}_2^{\text{input}}\end{array}\right]; c_0, \boldsymbol{c}_1\right) \quad \textit{and} \quad \eta_0^* = C_{\text{probit}}\left(\left[\begin{array}{c}\boldsymbol{\eta}_1^{\text{input}}\\ \boldsymbol{\eta}_2^{\text{input}}\end{array}\right], \left[\begin{array}{c}\boldsymbol{\eta}_1^*\\ \boldsymbol{\eta}_2^*\end{array}\right]; c_0, \boldsymbol{c}_1\right).$$

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

3.2 Message Passing Formulation

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

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

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

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

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 (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 (6) integrand. Each of these factors depend on $\boldsymbol{u}_i$, which is signified by an edges connecting each factor node to the lone stochastic variable node.

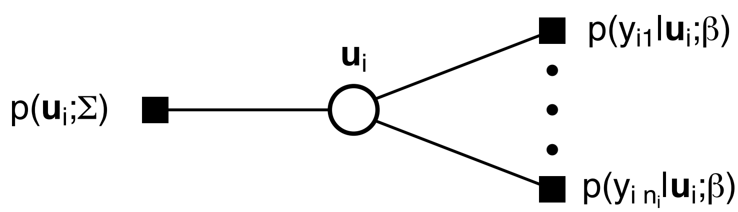

Figure 1: *Factor graph representation of the product structure of the integrand in (6). The open circle corresponds to the random effect vector $\boldsymbol{u}_i$ and the solid rectangles indicate factors. Edges indicate dependence of each factor on $\boldsymbol{u}_i$.*

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

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

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

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

as functions of $\boldsymbol{u}_i$. Noting that this problem has the form (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

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

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 $p(\boldsymbol{u}_i;\boldsymbol{\Sigma})$ to $\boldsymbol{u}_i$ is

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

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

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

and

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

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 propagation approximation to $\ell_i(\boldsymbol{\beta},\boldsymbol{\Sigma})$ is

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

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

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

where

$$\eta_{\boldsymbol{\Sigma}} \equiv \begin{bmatrix} -\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}| \\ \mathbf{0}_{d^{\mathrm{R}}} \\ -\frac{1}{2}\boldsymbol{D}_{d^{\mathrm{R}}}^T \mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}, \quad \mathrm{SUM}\{\eta_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\} \equiv \sum_{j=1}^{n_i} \eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i},$$

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 $\ell(\boldsymbol{\beta},\boldsymbol{\Sigma})$ is summarized as Algorithm 1. The derivational details are given in Section S.2.

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 $\underset{\sim}{\ell}(\boldsymbol{\beta},\boldsymbol{\Sigma})$ and approximate likelihood-based inference.

3.3 Recommended Starting Values for Algorithm 1

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

Inputs: $y_{ij}, \boldsymbol{x}_{ij}^{\text{F}}, \boldsymbol{x}_{ij}^{\text{R}}$, $1 \le i \le m$, $1 \le j \le n_i$;

$\boldsymbol{\beta}$ ($d^{\text{F}} \times 1$), $\boldsymbol{\Sigma}$ ($d^{\text{R}} \times d^{\text{R}}$, symmetric and positive definite).

Set constants: $c_{0,ij} \longleftarrow (2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}})$; $\boldsymbol{c}_{1,ij} \longleftarrow (2y_{ij}-1)\boldsymbol{x}_{ij}^{\text{R}}$, $\quad 1 \le i \le m,\ 1 \le j \le n_i$;

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

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

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

Cycle:

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

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

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

$$\begin{aligned}\Big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\Big)_{-0} \longleftarrow K_{\text{probit}}\Big(&\big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}; c_{0,ij}, \boldsymbol{c}_{1,ij}\Big) \\ &- \big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}\end{aligned}$$

until all natural parameter vectors converge.

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

$$\begin{aligned}\Big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\Big)_0 \longleftarrow C_{\text{probit}}\Big(&\big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}, \big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\big)_{-0} \\ &+ \big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}; c_{0,ij}, \boldsymbol{c}_{1,ij}\Big)\end{aligned}$$

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

Output: The expectation propagation approximate log-likelihood given by

$$\begin{aligned}\underset{\sim}{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \;=\; \tfrac{1}{2}\, m \log(2\pi) + \sum_{i=1}^{m} \Big\{ &\Big(\eta_{\boldsymbol{\Sigma}} + \text{SUM}\{\eta_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\}\Big)_0 \\ &+ A_N\Big(\Big(\eta_{\boldsymbol{\Sigma}} + \text{SUM}\{\eta_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\}\Big)_{-0}\Big)\Big\}\end{aligned}$$

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

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

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

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

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 $(\boldsymbol{\beta},\boldsymbol{\Sigma})$ to obtain the expectation propagation-approximate maximum likelihood estimators $(\widehat{\underset{\sim}{\boldsymbol{\beta}}},\widehat{\underset{\sim}{\boldsymbol{\Sigma}}})$. This is also the issue of approximate inference based on Fisher information theory.

Since $\underset{\sim}{\ell}(\boldsymbol{\beta},\boldsymbol{\Sigma})$ 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 & Hoetig (2005) provides a concise summary of the Broyden-Fletcher-Goldfarb-Shanno method.

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

$$\boldsymbol{\theta} \equiv \text{vech}\big(\tfrac{1}{2}\log(\boldsymbol{\Sigma})\big)$$

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

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

is the spectral decomposition of $\boldsymbol{\Sigma}$ and $\log(\boldsymbol{\lambda}_{\boldsymbol{\Sigma}})$ denotes element-wise evaluation of the logarithm to the entries of $\boldsymbol{\lambda}_{\boldsymbol{\Sigma}}$. 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 $\boldsymbol{\Sigma}$ is

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

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 $\mathbf{\Sigma}$. The full details are given in Section S.4 of the online supplement. Here we sketch the idea for the special case of $d^{\mathrm{R}} = 2$, for which

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

For confidence interval construction it is appropriate (e.g. Section 2.4 of Pinheiro & 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

$$\begin{bmatrix} \widehat{\underset{\sim}{\boldsymbol{\beta}}} \\ \widehat{\underset{\approx}{\boldsymbol{\omega}}} \end{bmatrix} \pm \Phi^{-1}(1 - \tfrac{1}{2}\,\alpha)\sqrt{-\text{diagonal}\big(\{\mathsf{H}\,\underset{\approx}{\ell}(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\approx}{\boldsymbol{\omega}}})\}^{-1}\big)} \tag{15}$$

where $\mathsf{H}\,\underset{\approx}{\ell}(\boldsymbol{\beta}, \boldsymbol{\omega})$ is the Hessian matrix of $\underset{\approx}{\ell}$ 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 parametrization whilst $(\boldsymbol{\beta}, \boldsymbol{\omega})$ is a constrained parametrization. Hence, the optimization should be performed with respect to the former parametrization whereas the Hessian matrix in (15) is respect to the latter parametrization. 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 $(\boldsymbol{\beta}, \boldsymbol{\theta})$ parametrization in the function being maximized and the `hessian` argument set to `FALSE`.

- Compute $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\approx}{\boldsymbol{\omega}}})$ and use this as a initial value with a call to `optim()` with the $(\boldsymbol{\beta}, \boldsymbol{\omega})$ parametrization 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

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

where $\eta_{\mathbf{\Sigma}}$ and $\text{SUM}\{\eta_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\}$ are as in Algorithm 1 with $(\boldsymbol{\beta}, \mathbf{\Sigma}) = (\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\mathbf{\Sigma}}})$, $\widehat{\underset{\sim}{\boldsymbol{\eta}}}_{i1}$ is the sub-vector 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 $\mathrm{BP}(\boldsymbol{u}_i)$, based on the expectation propagation estimate, is

$$\underset{\sim}{\mathrm{BP}}(\boldsymbol{u}_i) = -\tfrac{1}{2}\Big\{\mathrm{vec}^{-1}\Big(\boldsymbol{D}_d^{+T}\underset{\sim}{\widehat{\boldsymbol{\eta}}}_{i2}\Big)\Big\}^{-1}\underset{\sim}{\widehat{\boldsymbol{\eta}}}_{i1}. \tag{17}$$

The corresponding covariance matrix empirical approximation is

$$\underset{\sim}{\mathrm{Cov}}(\boldsymbol{u}_i|\boldsymbol{y}) = -\tfrac{1}{2}\Big\{\mathrm{vec}^{-1}\Big(\boldsymbol{D}_d^{+T}\underset{\sim}{\widehat{\boldsymbol{\eta}}}_{i2}\Big)\Big\}^{-1}. \tag{18}$$

In view of equation (13.7) of McCulloch, Searle & Neuhaus (2008), $\mathrm{Cov}\{\underset{\sim}{\mathrm{BP}}(\boldsymbol{u}_i) - \boldsymbol{u}_i\}$ is approximated by $E_{\boldsymbol{y}_i}\{\underset{\sim}{\mathrm{Cov}}(\boldsymbol{u}_i|\boldsymbol{y}_i)\}$. Approximate prediction interval construction is hindered by this expectation over the sampling distribution of the responses. See, for example, Carlin & 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. Firstly, we report the results of some studies involving simulated data. Analysis of actual data is discussed later in this section.

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 1,000 datasets according to the $d^{\mathrm{R}} = 1$ version of (1) with true parameter values:

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

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

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

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

For each simulated dataset, the probit mixed model defined by (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.

Of interest is comparison of quadrature-free approximations (2) and (3) 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 shoddy 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.

4.1.2 Accuracy and Speed Assessment for Bivariate Random Effects

In this study we simulated $1,000$ datasets according to a $d^{\mathrm{R}} = 2$ version of (1) with true parameter values:

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

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 3 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 corner of each panel are the empirical coverage values based on all $1,000$ 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 gotten the fitting times down 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 1,000 replications the median computing time was 18 seconds, the upper quartile was 20 seconds and the maximum was 34 seconds. Such speed is impressive given that each data set contained tens of thousands of observations and bivariate random effects are accurately handled.

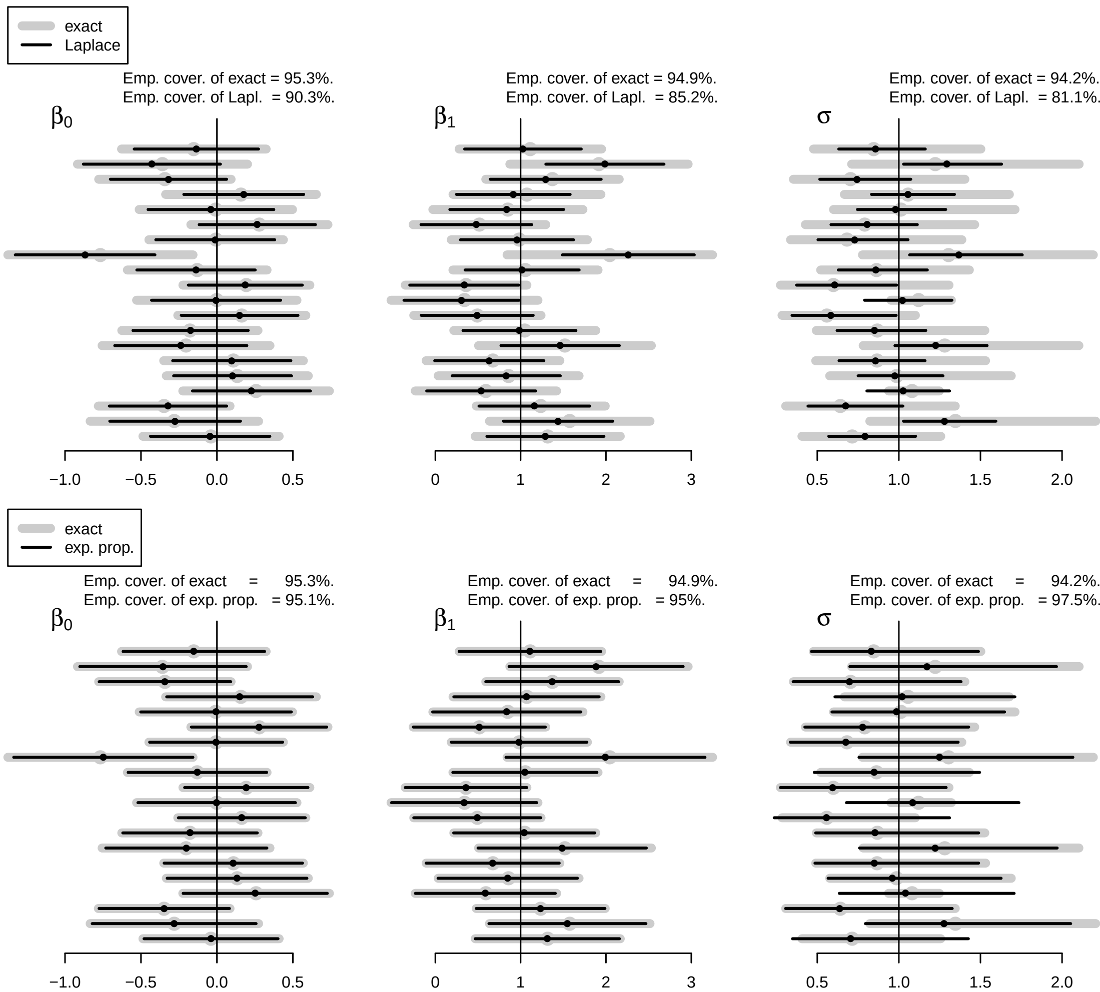

Figure 2: *Comparison of point estimation and 95% confidence interval coverage for the first simulation study with true parameter values given by (19). The upper row of panels compares exact maximum likelihood with Laplace approximation. The low row of panels compares exact maximum likelihood with expectation 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* $1,000$ *replications for each method involved in the comparison.*

4.2 Application to Data from a Fertility Study

Data from a 1988 Bangladesh fertility study are stored in the data frame `Contraception` within the R package mlmRev (Bates, Maechler and Bolker, 2014). Steele, Diamond and Amin (1996) contains details of the study and some multilevel analyses. Variables in the `Contraception` data frame include:

`use` a two-level factor variable indicating whether a woman is a user of contraception at the time of the survey, with levels `Y` for use and `N` for non-use,

`age` age of the woman in years at the time of the survey, centred about the average age of all women in the study,

`district` a multi-level factor variable that codes the district, out of 60 districts in total, in which the woman lives,

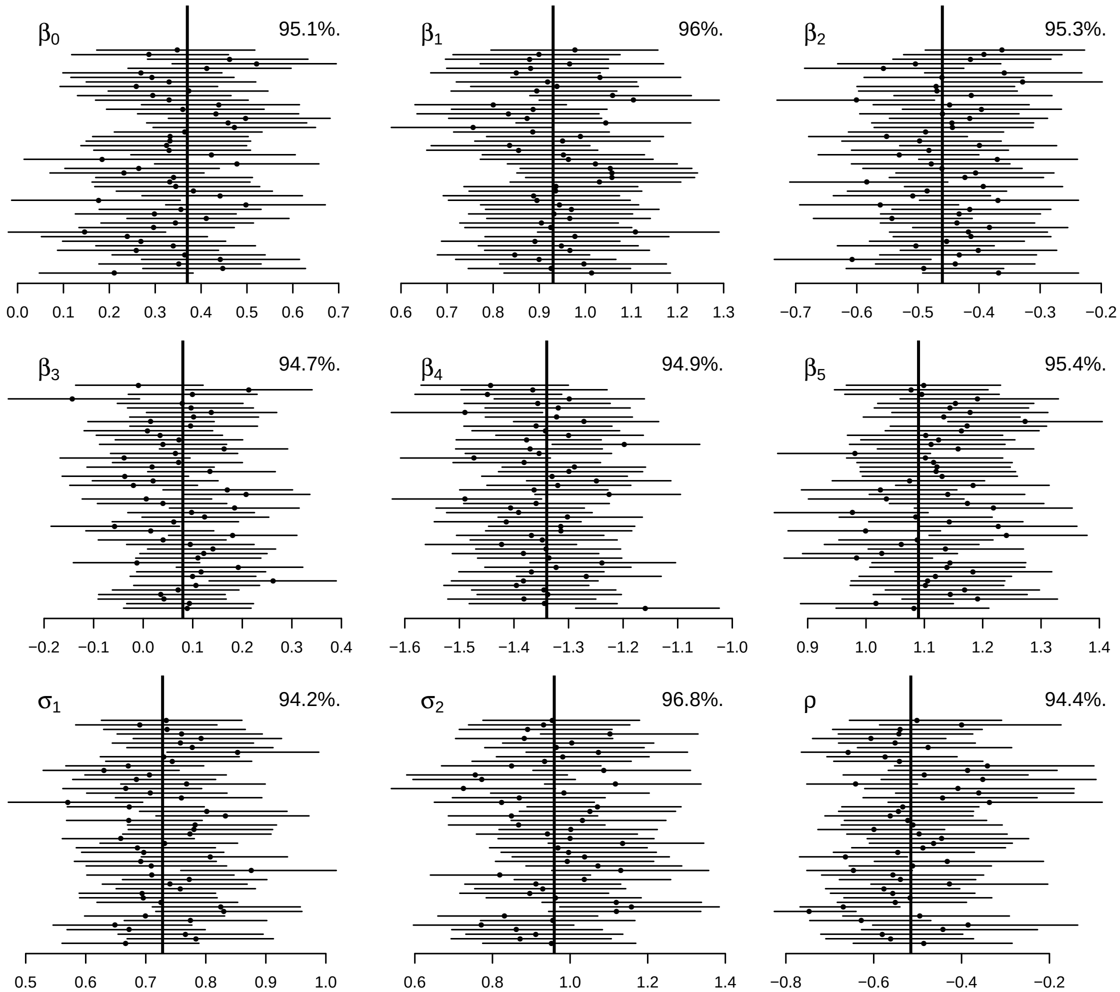

Figure 3: *Summary of confidence interval coverage for the second simulation study with true parameter values given by (21). The horizontal lines indicate expectation propagation-based confidence intervals for 50 randomly chosen replications of the simulation study described in the text. The solid circular 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* $1,000$ *replications.*

urban a two-level factor variable indicating whether or not the district in which the woman lives is urban, with levels `Y` for urban dwelling and `N` for rural dwelling, and

livch a four-level factor variable that indicates the number of living children of the woman, with levels `0` for no children, `1` for one child, `2` for two children and `3+` for three or more children.

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

$$\begin{aligned}I(\texttt{use}_{ij}=\texttt{Y})|u_{0i},u_{1i} \stackrel{\text{ind.}}{\sim} \text{Bernoulli}\Big(\Phi\big(&\beta_0+u_{0i}+(\beta_1+u_{1i})I(\texttt{urban}_{ij}=\texttt{Y})\\ &+\beta_2\texttt{age}_{ij}+\beta_3\,I(\texttt{livch}_{ij}=\texttt{1})+\beta_4\,I(\texttt{livch}_{ij}=\texttt{2})+\beta_5\,I(\texttt{livch}_{ij}=\texttt{3+})\big)\Big)\end{aligned} \tag{22}$$

where $I(\mathcal{P})=1$ if $\mathcal{P}$ is true and 0 otherwise. Also, $\texttt{use}_{ij}$ denotes the value of `use` for the jth woman within the ith district, $1\le i\le 60$, with the other variables defined analogously.

parameter	95% C.I. low.	estimate	95% C.I. upp.
β_0	-1.2185	-1.0418	-0.8651
β_1	0.2956	0.5003	0.7049
β_2	-0.0259	-0.0164	-0.0068
β_3	0.4934	0.6815	0.8698
β_4	0.6223	0.8306	1.0389
β_5	0.6102	0.8244	1.0387
σ_1	0.2748	0.3785	0.5214
σ_2	0.3096	0.4965	0.7962
ρ	-0.9367	-0.7984	-0.4446

Table 1: *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).*

The bivariate random effects vectors are assumed to satisfy

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

We fitted this model using our expectation propagation approximate likelihood inference scheme. It took about 35 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 1.

Each of the parameters is seen to be statistically significantly different from zero. As examples, the 95% confidence interval for β_2 of $(0.296, 0.705)$ indicates a higher use of contraception in urban districts ad the 95% confidence interval for σ_2 of $(0.310, 0.796)$ shows that their is signficant heterogeneity in the urban versus rural effect across the 60 districts.

We also used expectation propagation approximate best prediction to obtain predictions of the u_{0i} and u_{1i} values. The results are plotted in Figure 4 and provide a visualization of between-district heterogeneity.

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

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

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

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

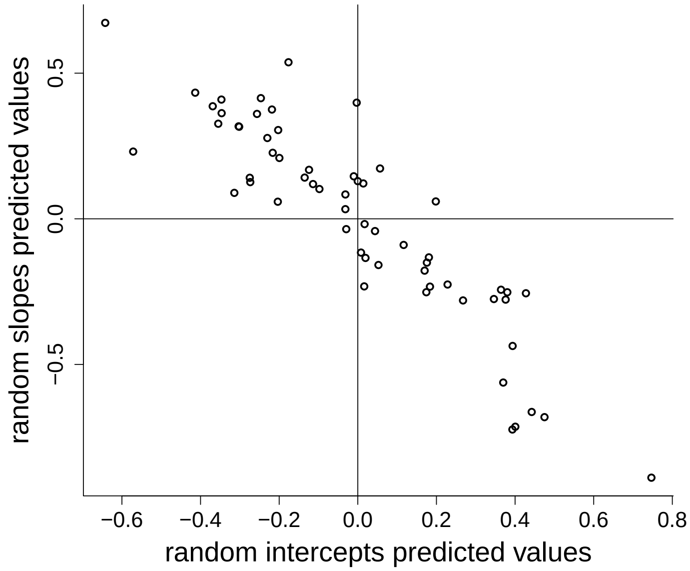

Figure 4: *Scatterplot of the expectation propagation-approximate best predictions of the random slopes and corresponding random intercepts for the fit of model given by (22) and (23) to data from a 1998 Bangladesh fertility study.*

Now make the change of variables

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

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

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

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

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

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

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

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 $d^{\mathrm{R}} = 1$ and write δ_i instead of $\boldsymbol{\Delta}_i$ to signify the fact that this quantity is scalar in this special case. Next, we make the

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

This assumption is in keeping with the fact that δ_i is the expectation propagation approximation to the sample standard deviation of $\underset{\sim}{\mathrm{BP}}(\boldsymbol{u}_i) - \boldsymbol{u}_i$. Then Taylor series expansion of A_{ij} about zero and substitution into the $d^{\mathrm{R}} = 1$ 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 (24) and using $\log(1+\varepsilon) \approx \varepsilon$ for small ε we obtain

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

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 $(\underset{\sim}{\widehat{\boldsymbol{\beta}}}, \underset{\sim}{\widehat{\boldsymbol{\Sigma}}})$ 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 & 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 for crossed random effects mixed models is Baayen, Davidson & Bates (2008). Here we provide advice regarding extension our expectation propagation approach to these settings.

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

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

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 $(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{\mathrm{L1}}, \boldsymbol{\Sigma}^{\mathrm{L2}})$ may be written as

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

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

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

and

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

Expectation propagation approximation of $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{\text{L1}}, \boldsymbol{\Sigma}^{\text{L2}})$ then proceeds by message passing on the factor graph displayed in Figure 5. 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.

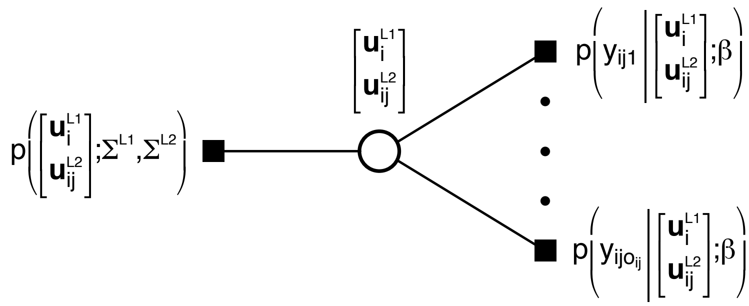

Figure 5: *Factor graph representation of the product structure of the integrand in (6). The open circle corresponds to the random effect vector* $[\boldsymbol{u}_i^{\text{L1}}\ \boldsymbol{u}_{ij}^{\text{L2}}]^T$ *and the solid rectangles indicate factors in the integrand of (28). Edges indicate dependence of each factor on* $[\boldsymbol{u}_i^{\text{L1}}\ \boldsymbol{u}_{ij}^{\text{L2}}]^T$*.*

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

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

where the data are cross-tabulated according to membership of two groups of sizes m and m' indexed according to the pair $(i, i') \in \{1, \ldots, m\} \times \{1, \ldots, m'\}$, 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 $\boldsymbol{x}^{\text{F}}_{ii'j}$, $\boldsymbol{x}^{\text{R}}_{ii'j}$ and $\boldsymbol{x}^{\text{R}\prime}_{ii'j}$ correspond to the jth set of measurements within group (i, i'). The $\boldsymbol{u}_i$, $1 \le i \le m$, are $d^{\text{R}} \times 1$ random effects for group-specific departures from the fixed effects for the first group. The $\boldsymbol{u}'_i$, $1 \le i \le m'$, are $d^{\text{R}\prime} \times 1$ random effects for group-specific departures from the fixed effects for the second group. The log-likelihood of $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}')$ is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}') = \sum_{i=1}^{m}\sum_{i'=1}^{m'} I(n_{ii'} > 0) \log \int_{\mathbb{R}^{d^{\text{R}}+d^{\text{R}\prime}}} \prod_{j=1}^{n_{ii'}} p\left(y_{ii'j}\middle| \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}'_{i'} \end{bmatrix}; \boldsymbol{\beta}\right) p\left(\begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}'_{i'} \end{bmatrix}; \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'\right) d\begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}'_{i'} \end{bmatrix}$$

where

$$p\left(y_{ii'j}\middle| \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}'_{i'} \end{bmatrix}; \boldsymbol{\beta}\right) \equiv F\Big((2y_{ijk} - 1)\big(\boldsymbol{\beta}^T\boldsymbol{x}^{\text{F}}_{ii'j} + (\boldsymbol{u}_i)^T\boldsymbol{x}^{\text{R}}_{ii'j} + (\boldsymbol{u}'_{i'})^T\boldsymbol{x}^{\text{R}\prime}_{ii'j}\big)\Big),\ y_{ijk} = 0, 1,$$

and

$$p\left(\begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}'_{i'} \end{bmatrix}; \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'\right) \equiv |2\pi\boldsymbol{\Sigma}|^{-1/2}|2\pi\boldsymbol{\Sigma}'|^{-1/2} \exp\left\{-\tfrac{1}{2}\begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}'_{i'} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}' \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}'_{i'} \end{bmatrix}\right\}.$$

Expectation propagation approximation of $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}')$ can be achieved via message passing on a factor graph similar to that shown in Figure 5.

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 $F = \text{expit}$. Therefore, transferral from probit to logistic mixed models involves replacement of f_{input} in Theorem 1 by

$$f_{\text{input}}(\boldsymbol{x}) = \text{expit}(c_0 + \boldsymbol{c}_1^T \boldsymbol{x}) \exp\left\{ \begin{bmatrix} \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\eta}_1^{\text{input}} \\ \boldsymbol{\eta}_2^{\text{input}} \end{bmatrix} \right\}, \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{29}$$

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

$$\int_{-\infty}^{\infty} x^p \exp\{qx - rx^2 - \log(1 + e^x)\}\, dx, \quad p = 0, 1, 2,\ q \in \mathbb{R},\ r > 0. \tag{30}$$

In the Bayesian context, Gelman *et al.* (2014; Section 13.8) and Kim & Wand (2017) describe quadrature-based approaches to evaluation of (30), 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 is to be reported in Yu (2019).

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 $F = \text{expit}$ generalizes to

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

where the functions b and c are as given in Table 2.1 of McCullagh & Nelder (1989). Setting $b(x) = \log(1 + e^x)$ and $c(x) = 0$ gives the $F = \text{expit}$ 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 (2019) will contain a detailed account of the practicalities and performance of expectation propagation for this class of models.

Acknowledgments

This research was supported by Australian Research Council Discovery Project DP180100597. We are grateful for advice from Jim Booth, Omar Ghattas and Alan Huang on aspects of this research.

References

Azzalini, A. (2017). The R package `sn`: The skew-normal and skew-t distributions (version 1.5). `http://azzalini.stat.unipd.it/SN`

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.

Bates, D., Maechler, M. and Bolker, B. (2014). mlmRev: Examples from multilevel modelling software review. R package version 1.0.
`http://cran.r-project.org`.

Baltagi, B.H. (2013). *Econometric Analysis of Panel Data, Fifth Edition*. Chichester, U.K.: John Wiley & Sons.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67(1)**, 1–48.

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

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.

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.

Dehaene, G. and Barthelmé, S. (2018). Expectation propagation in the large-data limit. *Journal of the Royal Statistical Society, Series B*, **80**, 199–217.

Diggle, P., Heagerty, P., Liang, K.-L. and Zeger, S. (2002). *Analysis of Longitudinal Data, Second Edition*. Oxford, U.K.: Oxford University Press.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, **13**, 317–322.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B.,Vehtari, A. and Rubin, D.B. (2014). *Bayesian Data Analysis, Third Edition*, Boca Raton, Florida: CRC Press.

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

Givens, G.H. and Hoetig, J.A. (2005). *Computational Statistics*, Hoboken, New Jersey: John Wiley & Sons.

Goldfarb, D. (1970). A family of variable metric updates derived by variational means.

Mathematics of Computation, **24**, 23–26.

Goldstein, H. (2010). *Multilevel Statistical Models, Fourth Edition*. Chichester, U.K.: John Wiley & Sons.

Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

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.

Kim, A.S.I. and Wand, M.P. (2017). On expectation propagation for generalised, linear and mixed models. *Australian and New Zealand Journal of Statistics*, **59**, in press.

Magnus, J.R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition*. Chichester U.K.: Wiley.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

The Mathworks Incorporated (2018). Natick, Massachusetts, U.S.A.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*. London: Chapman and Hall.

McCulloch, C.E., Searle, S.R. and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models, Second Edition*. New York: John Wiley & Sons.

Minka, T.P. (2001). Expectation propagation for approximate Bayesian inference. In J.S. Breese & D. Koller (eds), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Burlington, Massachusetts: Morgan Kaufmann.

Minka, T. (2005). Divergence measures and message passing. *Microsoft Research Technical Report Series*, **MSR-TR-2005-173**, 1–17.

Minka, T. & Winn, J. (2008). Gates: A graphical notation for mixture models. *Microsoft Research Technical Report Series*, **MSR-TR-2008-185**, 1–16.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Ogden, H.E. (2015). A sequential reduction method for inference in generalized linear mixed models. *Electronic Journal of Statistics*, **9**, 135–152.

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

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation, Second Edition*. Hoboken, New Jersey: John Wiley & Sons.

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.

Shanno, D.F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, **24**, 647–656.

Steele, F., Diamond, I. and Amin, S. (1996). Immunization uptake in rural Bangladesh: a multilevel analysis. *Journal of the Royal Statistical Society, Series A*, **159**, 289–299.

The Mathworks Incorporated (2018). Natick, Massachusetts, U.S.A.

Wainwright, M.J. and Jordan, M.I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, **1**, 1–305.

Wand, M.P. and Ormerod, J.T. (2012). Continued fraction enhancement of Bayesian computing. *Stat*, **1**, 31–41.

Wand, M.P. and Yu, J.C.F. (2018). glmmEP: Fast and accurate likelihood-based inference in generalized linear mixed models via expectation propagation. R package version 1.0. http://cran.r-project.org.

Yu, J.C.F. (2019). *Fast and Accurate Frequentist Generalized Linear Mixed Model Analysis via Expectation Propagation.* Doctor of Philosophy thesis, University of Technology Sydney.

Supplement for:

Fast and Accurate Binary Response Mixed Model Analysis via Expectation Propagation

By P. Hall[1], I.M. Johnstone[2], J.T. Ormerod[3], M.P. Wand[4] and J.C.F. Yu[4]

[1]*University of Melbourne,* [2]*Stanford University,* [3]*University of Sydney and* [4]*University of Technology Sydney*

S.1 Proof of Theorem 1

For $\boldsymbol{x} \in \mathbb{R}^d$, define

$$\phi_{\boldsymbol{\Sigma}}(\boldsymbol{x}) \equiv (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\big(-\tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\big)$$

so that

$$\phi_{\boldsymbol{I}}(\boldsymbol{x}) = (2\pi)^{-d/2}\exp\big(-\tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\big).$$

We continue to use an unadorned ϕ to denote the Univariate Normal density function:

$$\phi(x) = (2\pi)^{-1/2}\exp(-\tfrac{1}{2}\,x^2).$$

The notation $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}}$ for a column vector $\boldsymbol{v}$ is also used.

Lemma 1. *For any function $g : \mathbb{R} \to \mathbb{R}$ and $d \times 1$ vectors $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ such that the integrals exist:*

$$\int_{\mathbb{R}^d} g(\boldsymbol{\alpha}_1^T\boldsymbol{x})\phi_{\boldsymbol{I}}(\boldsymbol{x})\,d\boldsymbol{x} = \int_{-\infty}^{\infty} g(\|\boldsymbol{\alpha}_1\|\,z)\phi(z)\,dz, \tag{S.1}$$

$$\int_{\mathbb{R}^d} g(\boldsymbol{\alpha}_1^T\boldsymbol{x})(\boldsymbol{\alpha}_2^T\boldsymbol{x})\phi_{\boldsymbol{I}}(\boldsymbol{x})\,d\boldsymbol{x} = \{(\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2)/\|\boldsymbol{\alpha}_1\|\}\int_{-\infty}^{\infty} z\,g(\|\boldsymbol{\alpha}_1\|\,z)\phi(z)\,dz \tag{S.2}$$

and
$$\begin{aligned}\int_{\mathbb{R}^d} g(\boldsymbol{\alpha}_1^T\boldsymbol{x})(\boldsymbol{\alpha}_2^T\boldsymbol{x})(\boldsymbol{\alpha}_3^T\boldsymbol{x})\phi_{\boldsymbol{I}}(\boldsymbol{x})\,d\boldsymbol{x} &= (\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_3)\int_{-\infty}^{\infty} g(\|\boldsymbol{\alpha}_1\|\,z)\phi(z)\,dz \\ &\quad +\{(\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2)(\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_3)/\|\boldsymbol{\alpha}_1\|^2\}\int_{-\infty}^{\infty}(z^2-1)g(\|\boldsymbol{\alpha}_1\|\,z)\phi(z)\,dz.\end{aligned} \tag{S.3}$$

Proof of Lemma 1. Lemma 1 is a consequence of the fact that the integrals on the left-hand side are, respectively,

$$E\{g(\boldsymbol{\alpha}_1^T\boldsymbol{x})\},\ E\{g(\boldsymbol{\alpha}_1^T\boldsymbol{x})(\boldsymbol{\alpha}_2^T\boldsymbol{x})\}\quad\text{and}\quad E\{g(\boldsymbol{\alpha}_1^T\boldsymbol{x})(\boldsymbol{\alpha}_2^T\boldsymbol{x})(\boldsymbol{\alpha}_3^T\boldsymbol{x})\}$$

where

$$\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{0}_d, \boldsymbol{I}_d).$$

We now focus on simplification of the third integral (S.4). Simplication of the first and second integrals is similar and simpler. Make the change of variables

$$\boldsymbol{s} \equiv \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} = \boldsymbol{A}\boldsymbol{x} \quad\text{where}\quad \boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{\alpha}_1^T \\ \boldsymbol{\alpha}_2^T \\ \boldsymbol{\alpha}_3^T \end{bmatrix}$$

so that

$$E\{g(\boldsymbol{\alpha}_1^T\boldsymbol{x})(\boldsymbol{\alpha}_2^T\boldsymbol{x})(\boldsymbol{\alpha}_3^T\boldsymbol{x})\} = E\{g(s_1)s_2s_3\} \quad \text{where} \quad \boldsymbol{s} \sim N(\mathbf{0}_3, \boldsymbol{A}\boldsymbol{A}^T).$$

We then note that,

$$\begin{aligned} E\{g(s_1)s_2s_3\} &= \int_{-\infty}^{\infty} g(s_1)\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s_2s_3p(s_2s_3|s_1)ds_2ds_3\right\}p(s_1)ds_1 \\ &= \int_{-\infty}^{\infty} g(s_1)\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{\text{Cov}(s_2,s_3|s_1)+E(s_2|s_1)E(s_3|s_1)\}ds_2ds_3\right\}p(s_1)ds_1 \end{aligned}$$

and make use of the result (see e.g. Theorem 3.2.4 of Mardia, Kent & Bibby, 1979)

$$\begin{aligned} \begin{bmatrix} s_2 \\ s_3 \end{bmatrix}\Bigg| s_1 \sim N\Bigg(& (s_1/\|\boldsymbol{\alpha}_1\|^2)\begin{bmatrix} \boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_3 \end{bmatrix}, \\ & \begin{bmatrix} \|\boldsymbol{\alpha}_2\|^2 & \boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_3 \\ \boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_3 & \|\boldsymbol{\alpha}_3\|^2 \end{bmatrix} - (1/\|\boldsymbol{\alpha}_1\|^2)\begin{bmatrix} (\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2)^2 & (\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2)(\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_3) \\ (\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2)(\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_3) & (\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_3)^2 \end{bmatrix}\Bigg). \end{aligned}$$

Result (S.4) then follows via simple algebraic manipulations.

Lemma 2. *For all* $a \in \mathbb{R}$ *and* $d \times 1$ *vectors* $\boldsymbol{b}$

$$\int_{\mathbb{R}^d} \Phi(a+\boldsymbol{b}^T\boldsymbol{x})\phi_{\boldsymbol{I}}(\boldsymbol{x})\,dx = \Phi\left(\frac{a}{\sqrt{\boldsymbol{b}^T\boldsymbol{b}+1}}\right), \tag{S.4}$$

$$\int_{\mathbb{R}^d} \boldsymbol{x}\,\Phi(a+\boldsymbol{b}^T\boldsymbol{x})\phi_{\boldsymbol{I}}(\boldsymbol{x})\,dx = \frac{\boldsymbol{b}}{\sqrt{\boldsymbol{b}^T\boldsymbol{b}+1}}\,\phi\left(\frac{a}{\sqrt{\boldsymbol{b}^T\boldsymbol{b}+1}}\right) \quad \textit{and} \tag{S.5}$$

$$\int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^T\,\Phi(a+\boldsymbol{b}^T\boldsymbol{x})\phi_I(\boldsymbol{x})\,dx = \Phi\left(\frac{a}{\sqrt{\boldsymbol{b}^T\boldsymbol{b}+1}}\right)\boldsymbol{I} - \frac{a\boldsymbol{b}\boldsymbol{b}^T}{\sqrt{(\boldsymbol{b}^T\boldsymbol{b}+1)^3}}\phi\left(\frac{a}{\sqrt{\boldsymbol{b}^T\boldsymbol{b}+1}}\right). \tag{S.6}$$

Proof of Lemma 2.

Suppose that Z_1 and Z_2 are independent $N(0,1)$ random variables. As defined in Section 4.2, for a logical proposition $\mathcal{P}$, let $I(\mathcal{P}) = 1$ if $\mathcal{P}$ is true and $I(\mathcal{P}) = 0$ if $\mathcal{P}$ is false. Then

$$P(Z_1 \le a + \|\boldsymbol{b}\|Z_2) = E\{I(Z_1 \le a+\|\boldsymbol{b}\|Z_2)\} = E[E\{I(Z_1 \le a+\|\boldsymbol{b}\|Z_2)|\boldsymbol{Z}_2\}] = E\{h(\boldsymbol{Z}_2)\}$$

where $h(z_2) \equiv E\{I(Z_1 \le a+\|\boldsymbol{b}\|Z_2)|Z_2 = z_2\}$. But note that

$$h(z_2) = P(Z_1 \le a+\|\boldsymbol{b}\|z_2) = \Phi(a+\|\boldsymbol{b}\|z_2)$$

which implies that

$$P(Z_1 \le a+\|\boldsymbol{b}\|Z_2) = E\{\Phi(a+\|\boldsymbol{b}\|Z_2)\} = \int_{-\infty}^{\infty}\Phi(a+\|\boldsymbol{b}\|z)\phi(z)\,dz.$$

Then

$$P(Z_1 \le a+\|\boldsymbol{b}\|Z_2) = P(Z_1 - \|\boldsymbol{b}\|Z_2 \le a) = P(X_3 \le a)$$

where $X_3 \equiv Z_1 - \|\boldsymbol{b}\|Z_2 \sim N(0, 1+\boldsymbol{b}^T\boldsymbol{b})$ by independence of Z_1 and Z_2. Then (S.4) follows immediately.

Next, let $\boldsymbol{e}_i$ denote the $d \times 1$ vector with ith entry equal to 1 and with zeroes elsewhere. Then, using Lemma 1,the ith entry of the right-hand side of (S.5) is

$$\begin{aligned}\int_{\mathbb{R}^d} (\boldsymbol{e}_i^T \boldsymbol{x})\, \Phi(a + \boldsymbol{b}^T \boldsymbol{x}) \phi_{\boldsymbol{I}}(\boldsymbol{x})\, d\boldsymbol{x} &= \{(\boldsymbol{e}_i^T \boldsymbol{b})/\|\boldsymbol{b}\|\} \int_{-\infty}^{\infty} z\Phi(a + \|\boldsymbol{b}\| z)\phi(z)\, dz \\ &= -\{(\boldsymbol{e}_i^T \boldsymbol{b})/\|\boldsymbol{b}\|\} \int_{-\infty}^{\infty} \Phi(a + \|\boldsymbol{b}\| z)\phi'(z)\, dz \\ &= (\boldsymbol{e}_i^T \boldsymbol{b}) \int_{-\infty}^{\infty} \phi(a + \|\boldsymbol{b}\| z)\phi(z)\, dz\end{aligned}$$

where the last result follows via integration by parts. The last integrand is

$$(2\pi)^{-1} \exp\{-\tfrac{1}{2}(a + \|\boldsymbol{b}\| z)^2 - \tfrac{1}{2} z^2\} = \phi\left(\frac{a}{\sqrt{\boldsymbol{b}^T \boldsymbol{b} + 1}}\right) \phi\left(\frac{z + a\|\boldsymbol{b}\|/(1 + \boldsymbol{b}^T \boldsymbol{b})}{1/\sqrt{\boldsymbol{b}^T \boldsymbol{b} + 1}}\right)$$

and (S.5) is an immediate consequence.

Lastly, because of (S.4), the (i, j) entry of the right-hand side of (S.6) is

$$\begin{aligned}&\int_{\mathbb{R}^d} (\boldsymbol{e}_i^T \boldsymbol{x})(\boldsymbol{e}_j^T \boldsymbol{x})\, \Phi(a + \boldsymbol{b}^T \boldsymbol{x}) \phi_{\boldsymbol{I}}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= (\boldsymbol{e}_i^T \boldsymbol{e}_j) \int_{-\infty}^{\infty} \Phi(a + \|\boldsymbol{b}\|\, z)\phi(z)\, dz + \frac{(\boldsymbol{e}_i^T \boldsymbol{b})(\boldsymbol{e}_j^T \boldsymbol{b})}{\|\boldsymbol{b}\|^2} \int_{-\infty}^{\infty} \Phi(a + \|\boldsymbol{b}\|\, z)\phi''(z)\, dz \\ &= (\boldsymbol{e}_i^T \boldsymbol{e}_j)\Phi\left(\frac{a}{\sqrt{\boldsymbol{b}^T \boldsymbol{b} + 1}}\right) - \frac{(\boldsymbol{e}_i^T \boldsymbol{b})(\boldsymbol{e}_j^T \boldsymbol{b})}{\|\boldsymbol{b}\|} \int_{-\infty}^{\infty} \phi(a + \|\boldsymbol{b}\|\, z)\phi'(z)\, dz\end{aligned}$$

where the last result follows via integration by parts. The last integrand is

$$-(2\pi)^{-1} z \exp\{-\tfrac{1}{2}(a + \|\boldsymbol{b}\| z)^2 - \tfrac{1}{2} z^2\} = -z\phi\left(\frac{a}{\sqrt{\boldsymbol{b}^T \boldsymbol{b} + 1}}\right) \phi\left(\frac{z + a\|\boldsymbol{b}\|/(1 + \boldsymbol{b}^T \boldsymbol{b})}{1/\sqrt{\boldsymbol{b}^T \boldsymbol{b} + 1}}\right)$$

and (S.6) follows.

$\blacksquare$

Next, we note a key connection between Kullback-Leibler projection onto the unnormalized and normalized Multivariate Normal families. For the latter, we introduce the notation

$$\text{proj}_{\text{N}}[p](\boldsymbol{x}) = q(\boldsymbol{x})$$

where q is the Multivariate Normal density function minimizes $\text{KL}(p\|q)$.

Lemma 3. *Let $f \in L_1(\mathbb{R}^d)$ be such that $f \geq 0$ and define $C_f \equiv \int_{\mathbb{R}^d} f(\boldsymbol{x})\, d\boldsymbol{x}$. Then*

$$\text{proj}[f](\boldsymbol{x}) = C_f\, \text{proj}_{\text{N}}[f/C_f](\boldsymbol{x}).$$

Proof Lemma 3.

Let $g(\cdot; \boldsymbol{\eta})$ be a generic unnormalized Multivariate Normal density function with natural parameter vector $\boldsymbol{\eta}$:

$$g(\boldsymbol{x}; \boldsymbol{\eta}) = \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \boldsymbol{\eta} \right\}.$$

Then the Kullback-Leibler divergence of $g(\cdot;\boldsymbol{\eta})$ from f is

$$\text{KL}\big(f\|g(\cdot;\boldsymbol{\eta})\big) = \int_{\mathbb{R}^d} \big[f(\boldsymbol{x})\log\{f(\boldsymbol{x})/g(\boldsymbol{x};\boldsymbol{\eta})\} + g(\boldsymbol{x};\boldsymbol{\eta}) - f(\boldsymbol{x})\big]\,d\boldsymbol{x} = \mathcal{K}(\boldsymbol{\eta}) + \text{const}$$

where 'const' denotes terms not depending on $\boldsymbol{\eta}$ and

$$\mathcal{K}(\boldsymbol{\eta}) \equiv (2\pi)^{d/2}\exp\{\boldsymbol{\eta}_0 + A_N(\boldsymbol{\eta}_{-0})\} - \begin{bmatrix} \int_{\mathbb{R}^d} f(\boldsymbol{x})\,d\boldsymbol{x} \\ \int_{\mathbb{R}^d} \boldsymbol{x}\,f(\boldsymbol{x})\,d\boldsymbol{x} \\ \int_{\mathbb{R}^d} \text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\,f(\boldsymbol{x})\,d\boldsymbol{x} \end{bmatrix}^T \boldsymbol{\eta}.$$

The derivative vector of $\mathcal{K}(\boldsymbol{\eta})$ is

$$\mathsf{D}\mathcal{K}(\boldsymbol{\eta}) = (2\pi)^{d/2}\exp\{\boldsymbol{\eta}_0 + A_N(\boldsymbol{\eta}_{-0})\}\begin{bmatrix} 1 \\ \mathsf{D}\,A(\boldsymbol{\eta}_{-0})^T \end{bmatrix}^T - \begin{bmatrix} \int_{\mathbb{R}^d} f(\boldsymbol{x})\,d\boldsymbol{x} \\ \int_{\mathbb{R}^d} \boldsymbol{x}\,f(\boldsymbol{x})\,d\boldsymbol{x} \\ \int_{\mathbb{R}^d} \text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\,f(\boldsymbol{x})\,d\boldsymbol{x} \end{bmatrix}^T$$

so the stationary condition, $\mathsf{D}\mathcal{K}(\boldsymbol{\eta})^T = \mathbf{0}$, for the minimization of $\text{KL}\big(f\|g(\cdot;\boldsymbol{\eta})\big)$ is

$$(2\pi)^{d/2}\exp\{\boldsymbol{\eta}_0 + A_N(\boldsymbol{\eta}_{-0})\}\begin{bmatrix} 1 \\ \nabla A_N(\boldsymbol{\eta}_{-0}) \end{bmatrix} = \begin{bmatrix} \int_{\mathbb{R}^d} f(\boldsymbol{x})\,d\boldsymbol{x} \\ \int_{\mathbb{R}^d} \boldsymbol{x}\,f(\boldsymbol{x})\,d\boldsymbol{x} \\ \int_{\mathbb{R}^d} \text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\,f(\boldsymbol{x})\,d\boldsymbol{x} \end{bmatrix}. \tag{S.7}$$

with $\nabla A_N(\boldsymbol{\eta}_{-0}) \equiv \mathsf{D}\,A(\boldsymbol{\eta}_{-0})^T$ denoting the *gradient* vector of $A(\boldsymbol{\eta}_{-0})$. It is easily checked that (S.7) is satisfied by

$$\begin{aligned} (\boldsymbol{\eta}^*)_0 &= \log(C_f) - A_N(\boldsymbol{\eta}^*_{-0}) - \tfrac{1}{2}\,d\log(2\pi) \\ \text{where}\quad \boldsymbol{\eta}^*_{-0} &= (\nabla A_N)^{-1}\left(\begin{bmatrix} \int_{\mathbb{R}^d} \boldsymbol{x}\,\{f(\boldsymbol{x})/C_f\}\,d\boldsymbol{x} \\ \int_{\mathbb{R}^d} \text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\,\{f(\boldsymbol{x})/C_f\}\,d\boldsymbol{x} \end{bmatrix}\right) \end{aligned} \tag{S.8}$$

with existence and uniqueness of $(\nabla A_N)^{-1}$ being guaranteed by Proposition 3.2 of Wainwright & Jordan (2008). The Hessian matrix of $\mathcal{K}(\boldsymbol{\eta})$ is

$$\mathsf{H}\,\mathcal{K}(\boldsymbol{\eta}) = (2\pi)^{d/2}e^{\boldsymbol{\eta}_0 + A_N(\boldsymbol{\eta}_{-0})}\left\{\begin{bmatrix} 1 \\ \nabla A_N(\boldsymbol{\eta}_{-0}) \end{bmatrix}\begin{bmatrix} 1 \\ \nabla A_N(\boldsymbol{\eta}_{-0}) \end{bmatrix}^T + \begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathsf{H}A_N(\boldsymbol{\eta}_{-0}) \end{bmatrix}\right\}$$

From Proposition 3.1 of Wainwright & Jordan (2008), A_N is strictly convex on its domain and therefore $\mathsf{H}A_N(\boldsymbol{\eta}_{-0})$ is positive definite. Hence $\mathsf{H}\,\mathcal{K}(\boldsymbol{\eta})$ is positive definite for all $\boldsymbol{\eta}$ and so (S.8) is the unique minimizer of $\text{KL}\big(f\|g(\cdot;\boldsymbol{\eta})\big)$. Therefore,

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

where $\boldsymbol{\eta}^*$ is as given by (S.8). However, $\boldsymbol{\eta}^*_{-0}$ is the same natural parameter vector that arises via projection of f/C_f onto the family of Multivariate Normal density functions and so

$$\text{proj}_N[f/C_f](\boldsymbol{x}) = \exp\left\{\begin{bmatrix} \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}^T \boldsymbol{\eta}^*_{-0} - A_N(\boldsymbol{\eta}^*_{-0})\right\}(2\pi)^{-d/2}$$

which immediately leads to Lemma 3.

∎

The proof of Theorem 1 involves transferral between the common $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ parameters of the d-variate Normal distribution and the natural parameters corresponding to the sufficient statistics $\boldsymbol{x}$ and $\text{vech}(\boldsymbol{x}\boldsymbol{x}^T)$. The transformations in each direction are

$$\begin{cases}\boldsymbol{\eta}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\[1ex] \boldsymbol{\eta}_2 = -\tfrac{1}{2}\boldsymbol{D}_d^T\text{vec}(\boldsymbol{\Sigma}^{-1})\end{cases} \quad \text{and} \quad \begin{cases}\boldsymbol{\mu} = -\tfrac{1}{2}\big\{\text{vec}^{-1}\Big(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\Big)\big\}^{-1}\boldsymbol{\eta}_1\\[1ex] \boldsymbol{\Sigma} = -\tfrac{1}{2}\big\{\text{vec}^{-1}\Big(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\Big)\big\}^{-1}\end{cases} \tag{S.9}$$

Recall the notation

$$\boldsymbol{v}^{\otimes k} \equiv \begin{cases}1 & \text{for } k=0\\ \boldsymbol{v} & \text{for } k=1\\ \boldsymbol{v}\boldsymbol{v}^T & \text{for } k=2\end{cases}$$

and consider Kullback-Leibler projection of $f_{\text{input}}/C_{f_{\text{input}}}$ onto the family of d-variate Normal density functions where

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

and $C_{f_{\text{input}}} \equiv \int_{\mathbb{R}^d} f_{\text{input}}(\boldsymbol{x})\,d\boldsymbol{x}$. Then the projection has mean and covariance matrix

$$\boldsymbol{\mu}^* = \mathcal{M}_1/\mathcal{M}_0 \quad \text{and} \quad \boldsymbol{\Sigma}^* = \mathcal{M}_2/\mathcal{M}_0 - (\mathcal{M}_1/\mathcal{M}_0)(\mathcal{M}_1/\mathcal{M}_0)^T \tag{S.10}$$

where

$$\mathcal{M}_k \equiv \int_{\mathbb{R}^d}\boldsymbol{x}^{\otimes k}\Phi(c_0 + \boldsymbol{c}_1^T\boldsymbol{x})\exp\left\{\begin{bmatrix}\boldsymbol{x}\\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\eta}_1^{\text{input}}\\ \boldsymbol{\eta}_2^{\text{input}}\end{bmatrix}\right\}d\boldsymbol{x}.$$

Letting

$$\boldsymbol{\Sigma}^{\text{input}} \equiv -\tfrac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{D}_{d^{\text{R}}}^{+T}\boldsymbol{\eta}_2^{\text{input}})\right\}^{-1} \quad \text{and} \quad \boldsymbol{\mu}^{\text{input}} \equiv \boldsymbol{\Sigma}^{\text{input}}\boldsymbol{\eta}_1^{\text{input}}$$

be the common parameters corresponding to $\boldsymbol{\eta}^{\text{input}}$ and making the change of variable $\boldsymbol{z} = (\boldsymbol{\Sigma}^{\text{input}})^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}^{\text{input}})$ we obtain

$$\mathcal{M}_k = (2\pi)^{d/2}e^{A_N(\boldsymbol{\eta}^{\text{input}})}\int_{\mathbb{R}^d}\left(\boldsymbol{\mu}^{\text{input}} + (\boldsymbol{\Sigma}^{\text{input}})^{1/2}\boldsymbol{z}\right)^{\otimes k}\Phi\Big((c_0 + \boldsymbol{c}_1^T\boldsymbol{\mu}^{\text{input}}) + \{(\boldsymbol{\Sigma}^{\text{input}})^{1/2}\boldsymbol{c}_1\}^T\boldsymbol{z}\Big)\phi_{\boldsymbol{I}}(\boldsymbol{z})\,d\boldsymbol{z}.$$

Lemma 2 and simple algebraic manipulations then give

$$\mathcal{M}_1/\mathcal{M}_0 = \boldsymbol{\mu}^{\text{input}} + \frac{\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1\zeta'(r_2)}{\sqrt{\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1 + 1}}.$$

and

$$\begin{aligned}\mathcal{M}_2/\mathcal{M}_0 = \boldsymbol{\mu}^{\text{input}}(\boldsymbol{\mu}^{\text{input}})^T &+ \frac{\{\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1(\boldsymbol{\mu}^{\text{input}})^T + \boldsymbol{\mu}^{\text{input}}\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}\}\zeta'(r_2)}{\sqrt{\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1 + 1}}\\ &+\boldsymbol{\Sigma}^{\text{input}} - \frac{r\zeta'(r_2)\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}}{\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1 + 1}.\end{aligned}$$

where

$$r_2 \equiv \frac{2c_0 - \boldsymbol{c}_1^T\big\{\text{vec}^{-1}\big(\boldsymbol{D}_{d^{\text{R}}}^{+T}\boldsymbol{\eta}_2^{\text{input}}\big)\big\}^{-1}\boldsymbol{\eta}_1^{\text{input}}}{\sqrt{2\left[2 - \boldsymbol{c}_1^T\big\{\text{vec}^{-1}\big(\boldsymbol{D}_{d^{\text{R}}}^{+T}\boldsymbol{\eta}_2^{\text{input}}\big)\big\}^{-1}\boldsymbol{c}_1\right]}}.$$

Combining these last two results and noting (S.10) we obtain the common parameter solutions

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}^{\text{input}} + \frac{\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1\zeta'(r_2)}{\sqrt{\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1+1}}$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}^{\text{input}} + \left\{\frac{\zeta''(r_2)}{\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1+1}\right\}\boldsymbol{\Sigma}^{\text{input}}\boldsymbol{c}_1\boldsymbol{c}_1^T\boldsymbol{\Sigma}^{\text{input}}.$$

Transferral to natural parameters via (S.9) and some simple manipulations then lead to

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

Finally,

$$\begin{aligned}\eta_0^* &= \log(C_{f_{\text{input}}}) - \log\int_{\mathbb{R}^{d^{\text{R}}}}\exp\left\{\begin{bmatrix}\boldsymbol{x}\\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\eta}_1^*\\ \boldsymbol{\eta}_2^*\end{bmatrix}\right\}d\boldsymbol{x}\\ &= \log(\mathcal{M}_0) - \tfrac{1}{2}\,d^{\text{R}}\log(2\pi) - A_N(\boldsymbol{\eta}^*) = \log\Phi(r_2) + A_N(\boldsymbol{\eta}^{\text{input}}) - A_N(\boldsymbol{\eta}^*)\\ &= C_{\text{probit}}\left(\begin{bmatrix}\boldsymbol{\eta}_1^{\text{input}}\\ \boldsymbol{\eta}_2^{\text{input}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\eta}_1^*\\ \boldsymbol{\eta}_2^*\end{bmatrix}; c_0, \boldsymbol{c}_1\right)\end{aligned}$$

S.2 Derivation of Algorithm 1

We now provide full justification of Algorithm 1, starting with a derivation of the message passing representation used in Algorithm 1.

S.2.1 Message Passing Representation Derivation

The derivation of the message passing representation is based on the infrastructure and results laid out in Minka (2005). The treatment given there is for a generalization of Kullback-Leibler divergence, known as *α-divergence*, and for approximation of (normalized) density functions rather than general non-negative L_1 functions. The Kullback-Leibler divergence minimization problem given by (8) corresponds to $\alpha = 1$ in the notation of Minka (2005). Following Section 4.1 of Minka (2005) we then define the messages passed from the factors neighboring $\boldsymbol{u}_i$ in Figure 1 to be

$$m_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) \equiv \underset{\sim}{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \quad\text{and}\quad m_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) \equiv p(\boldsymbol{u}_i;\boldsymbol{\Sigma}). \tag{S.11}$$

Then, (54) of Minka (2005) invokes the definition

$$m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i) \equiv m_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i)\prod_{j'\neq j} m_{p(y_{ij'}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i). \tag{S.12}$$

Result (60) of Minka (2005) with $\alpha = 1$, $s' = 1$ (since we are working with unnormalized rather than normalized Kullback-Leibler divergence) and the simplification that there is only one stochastic node, namely $\boldsymbol{u}_i$, provides the main factor to stochastic node message passing updates:

$$m_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) \longleftarrow \frac{\text{proj}\big[m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\big](\boldsymbol{u}_i)}{m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)},\ 1\le j\le n_i. \tag{S.13}$$

The other factor to stochastic node message passing update is, trivially from (S.11),

$$m_{p(\boldsymbol{u}_i;\,\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) \longleftarrow p(\boldsymbol{u}_i;\boldsymbol{\Sigma}).$$

The stochastic node to factor updates are, from (S.12),

$$m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}(\boldsymbol{u}_i) \longleftarrow m_{p(\boldsymbol{u}_i;\,\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i)\prod_{j'\neq j} m_{p(y_{ij'}|\boldsymbol{u}_i;\,\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i), \quad 1\leq j\leq n_i.$$

Next, we simplify these message updates to a programmable form.

S.2.2 Simplification of the $m_{p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i)$ Updates

From (S.12) it is apparent that $m_{p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i)$ is an unnormalized Multivariate Normal density function and therefore

$$m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}(\boldsymbol{u}_i) = \exp\left\{\begin{bmatrix}1\\ \boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}\right\}$$

with natural parameter vector $\eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}$. Introducing the abbreviation:

$$\boldsymbol{\eta}^{\otimes} \equiv \eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}$$

we have

$$m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}(\boldsymbol{u}_i) = \exp\left(\eta_0^{\otimes}\right)\exp\left\{\begin{bmatrix}\boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \boldsymbol{\eta}_{-0}^{\otimes}\right\}$$

where $\eta_0^{\otimes}$ denotes the first entry of $\boldsymbol{\eta}^{\otimes}$ and $\boldsymbol{\eta}_{-0}^{\otimes}$ contains the remaining entries. Substitution info (S.13) leads to

$$\begin{aligned}
m_{p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) &\longleftarrow \frac{\text{proj}\left[\exp\left(\eta_0^{\otimes}\right)\exp\left\{\begin{bmatrix}\boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \boldsymbol{\eta}_{-0}^{\otimes}\right\}\Phi\big((2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}}+\boldsymbol{u}_i^T\boldsymbol{x}_{ij}^{\text{R}})\big)\right](\boldsymbol{u}_i)}{\exp\left(\eta_0^{\otimes}\right)\exp\left\{\begin{bmatrix}\boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \boldsymbol{\eta}_{-0}^{\otimes}\right\}}\\
&= \frac{\text{proj}\left[\Phi\big(c_{0,ij}+\boldsymbol{c}_{1,ij}^T\,\boldsymbol{u}_i\big)\exp\left\{\begin{bmatrix}\boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \boldsymbol{\eta}_{-0}^{\otimes}\right\}\right](\boldsymbol{u}_i)}{\exp\left\{\begin{bmatrix}\boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \boldsymbol{\eta}_{-0}^{\otimes}\right\}}
\end{aligned}$$

where

$$c_{0,ij} \equiv (2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}}) \quad \text{and} \quad \boldsymbol{c}_{1,ij} \equiv (2y_{ij}-1)\boldsymbol{x}_{ij}^{\text{R}}.$$

Using Theorem 1:

$$m_{p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) \leftarrow \exp\left\{\begin{bmatrix}1\\ \boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \eta_{p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\right\}$$

where the linear and quadratic coefficient updates are

$$\begin{aligned}
\big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\big)_{-0} \longleftarrow\, & K_{\text{probit}}\Big(\big(\eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}\big)_{-0};(2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}}),(2y_{ij}-1)\boldsymbol{x}_{ij}^{\text{R}}\Big)\\
&-\big(\eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\,\boldsymbol{\beta})}\big)_{-0}
\end{aligned}$$

and the constant coefficient update is

$$\big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\big)_0 \longleftarrow C_{\text{probit}}\Big(\big(\eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}, \big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\big)_{-0} + \big(\eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}; (2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}}), (2y_{ij}-1)\boldsymbol{x}_{ij}^{\text{R}}\Big).$$

S.2.3 Simplification of the $m_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i)$ Update

The second definition in (S.11) gives

$$m_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) \longleftarrow p(\boldsymbol{u}_i;\boldsymbol{\Sigma}) = \exp\left\{\begin{bmatrix}1\\ \boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \eta_{\boldsymbol{\Sigma}}\right\}.$$

Therefore, if $\eta_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}$ denotes the natural parameter vector of $m_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i)$ then it has the trivial update

$$\eta_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i} \longleftarrow \eta_{\boldsymbol{\Sigma}}.$$

S.2.4 Simplification of the $m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)$ Updates

Given the simplified forms of the messages in the two previous subsections we have from (S.12):

$$m_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i) \longleftarrow \exp\left\{\begin{bmatrix}1\\ \boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \eta_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i}\right\} \times \prod_{j'\neq j}\exp\left\{\begin{bmatrix}1\\ \boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\end{bmatrix}^T \eta_{p(y_{ij'}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\right\}$$

which leads to

$$\begin{aligned}\eta_{\boldsymbol{u}_i\,\rightarrow\,p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})} &\longleftarrow \eta_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i} + \sum_{j'\neq j}\eta_{p(y_{ij'}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\\ &= \eta_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{u}_i} + \text{SUM}\{\eta_{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\} - \eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}.\end{aligned}$$

S.2.5 Assembly of All Natural Parameter Updates

We now return to the message passing protocol given in Section 3.2:

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).

For the factors $p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})$:

computing the messages passed to each of these factors reduces to

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

and computing the messages passed from these factors reduces to

$$\begin{aligned}\Big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\Big)_{-0} \longleftarrow K_{\text{probit}}\Big(\big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}; c_{0,ij}, \boldsymbol{c}_{1,ij}\Big)&\\ -\big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}&\end{aligned}$$

and

$$\begin{aligned}\Big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\Big)_{0} \longleftarrow C_{\text{probit}}\Big(\big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}, \big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\big)_{-0}&\\ +\big(\eta_{\boldsymbol{u}_i \to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\big)_{-0}; c_{0,ij}, \boldsymbol{c}_{1,ij}\Big).&\end{aligned}$$

For the factors $p(\boldsymbol{u}_i;\boldsymbol{\Sigma})$:

computing the messages passed from these factors reduces to

$$\eta_{\boldsymbol{u}_i \to p(\boldsymbol{u}_i;\boldsymbol{\Sigma})} \longleftarrow \sum_{j=1}^{n_i} \eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}$$

and computing the messages passed to these factors reduces to

$$\eta_{p(\boldsymbol{u}_i;\boldsymbol{\Sigma}) \to \boldsymbol{u}_i} \to \eta_{\boldsymbol{\Sigma}}.$$

Algorithm 1 is essentially these natural parameter updates being cycled until convergence. The update for $\Big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\Big)_0$ can be moved outside of the cycle loop without affecting convergence. Also, the $\eta_{\boldsymbol{u}_i \to p(\boldsymbol{u}_i;\boldsymbol{\Sigma})}$ updates are redundant and are omitted from Algorithm 1.

S.3 Derivation of Starting Values Recommendation

We now derive useful starting values for the $\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}$ that have to be initialized in Algorithm 1. Note that

$$\log\, p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) = \sum_{j=1}^{n_i}\{\zeta(a_{ij}) - \log(2)\} \quad \text{where} \quad a_{ij} \equiv (2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}} + \boldsymbol{u}_i^T\boldsymbol{x}_{ij}^{\text{R}})$$

and ζ is as defined in Section 3.1. Let $\widehat{\boldsymbol{u}}_i$ be a prediction of $\boldsymbol{u}_i$ and consider the following expansion of the data-dependent component of $\ell(\boldsymbol{\beta},\boldsymbol{\Sigma})$:

$$\begin{aligned}\zeta(a_{ij}) &= \zeta\big(\widehat{a}_{ij} + (\boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i)^T\boldsymbol{x}_{ij}^{\text{R}}(2y_{ij}-1)\big)\\ &= \zeta(\widehat{a}_{ij}) + (\boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i)^T\boldsymbol{x}_{ij}^{\text{R}} y_{ij}^{\dagger}\zeta'(\widehat{a}_{ij}) + \tfrac{1}{2}\{(\boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i)^T\boldsymbol{x}_{ij}^{\text{R}}(2y_{ij}-1)\}^2\zeta''(\widehat{a}_{ij}) + \dots\\ &= \begin{bmatrix} 1 \\ \boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i \\ \text{vech}\big((\boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i)(\boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i)^T\big)\end{bmatrix}^T \check{\boldsymbol{\eta}}_{ij} + \dots\end{aligned}$$

where, as in Section 3.3, $\widehat{a}_{ij} \equiv (2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}} + \boldsymbol{u}_i^T\boldsymbol{x}_{ij}^{\text{R}})$, and

$$\check{\boldsymbol{\eta}}_{ij} \equiv \begin{bmatrix} \zeta(\widehat{a}_{ij}) \\ \boldsymbol{x}_{ij}^{\text{R}}(2y_{ij}-1)\zeta'(\widehat{a}_{ij}) \\ \frac{1}{2}\zeta''(\widehat{a}_{ij})\boldsymbol{D}_{d^{\text{R}}}^T\text{vec}\big(\boldsymbol{x}_{ij}^{\text{R}}(\boldsymbol{x}_{ij}^{\text{R}})^T\big) \end{bmatrix}.$$

It follows that the quadratic approximation to $\log p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})$ based on Taylor expansion about $\widehat{\boldsymbol{u}}_i$ is $\log \check{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})$ where

$$\check{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \equiv \exp\left\{\begin{bmatrix} 1 \\ \boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i \\ \text{vech}\big((\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i)(\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i)^T\big) \end{bmatrix}^T \check{\boldsymbol{\eta}}_{ij}\right\}.$$

The starting value recommendation for $\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}$ is based on replacement of $\check{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})$ by $p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})$ in (S.13):

$$m_{\check{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}(\boldsymbol{u}_i) \longleftarrow \frac{\text{proj}[m_{\boldsymbol{u}_i\,\rightarrow\,\check{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)\,\check{p}(y_{ij}|u_i;\boldsymbol{\beta})](\boldsymbol{u}_i)}{m_{\boldsymbol{u}_i\,\rightarrow\,\check{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}(\boldsymbol{u}_i)} = \check{p}(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})$$

with the proj[$\cdot$] being superfluous in this case due to $\check{p}(y_{ij}|u_i;\boldsymbol{\beta})$ being already in the Multivariate Normal family. The starting value for $\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}$ that arises from this substitution is then given by

$$\exp\left\{\begin{bmatrix} 1 \\ \boldsymbol{u}_i \\ \text{vech}\big(\boldsymbol{u}_i\boldsymbol{u}_i^T\big) \end{bmatrix}^T \eta^{\text{start}}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\right\} = \exp\left\{\begin{bmatrix} 1 \\ \boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i \\ \text{vech}\big((\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i)(\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i)^T\big) \end{bmatrix}^T \check{\boldsymbol{\eta}}_{ij}\right\}.$$

By matching coefficients of like terms we arrive at

$$\eta^{\text{start}}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i} = \begin{bmatrix} \eta_0^{\text{start}} \\ (2y_{ij}-1)\zeta'(\widehat{a}_{ij})\boldsymbol{x}_{ij}^{\text{R}} - \zeta''(\widehat{a}_{ij})\boldsymbol{x}_{ij}^{\text{R}}(\boldsymbol{x}_{ij}^{\text{R}})^T\widehat{\boldsymbol{u}}_i \\ \frac{1}{2}\zeta''(\widehat{a}_{ij})\boldsymbol{D}_{d^{\text{R}}}^T\text{vec}\big(\boldsymbol{x}_{ij}^{\text{R}}(\boldsymbol{x}_{ij}^{\text{R}})^T\big) \end{bmatrix}$$

where

$$\eta_0^{\text{start}} = \zeta(\widehat{a}_{ij}) - (2y_{ij}-1)\zeta'(\widehat{a}_{ij})(\boldsymbol{x}_{ij}^{\text{R}})^T\widehat{\boldsymbol{u}}_i + \tfrac{1}{2}\zeta''(\widehat{a}_{ij})\{(\boldsymbol{x}_{ij}^{\text{R}})^T\widehat{\boldsymbol{u}}_i\}^2.$$

In Algorithm 1 the cycle loop corresponds to determination of the natural parameter vector

$$\big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}\big)_{-0}$$

implying that the first entry of $\eta^{\text{start}}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\,\rightarrow\,\boldsymbol{u}_i}$ is not needed for these iterations. Hence, we can instead set $\eta_0^{\text{start}} = 0$ without affecting Algorithm 1. We now have (14).

S.4 Details of Confidence Interval Calculations

Here we provide full details of approximate confidence intervals calculations based on quasi-Newton maximization of $\underset{\sim}{\ell}(\boldsymbol{\beta},\boldsymbol{\Sigma})$. The calculations depend on the following ingredients:

- some additional convenient matrix notation.
- formulae for transformation from the parameter vector $\boldsymbol{\theta} \equiv \text{vech}\big(\frac{1}{2}\log(\boldsymbol{\Sigma})\big)$ to a parameter vector $\boldsymbol{\omega}$ that is more appropriate for confidence interval construction.

- formulae for the reverse transformation: from $\boldsymbol{\omega}$ to $\boldsymbol{\theta}$.
- a quasi-Newton optimization-based strategy for calculating confidence intervals for the entries of $\boldsymbol{\omega}$, which are then easily transformed to confidence intervals for interpretable covariance matrix parameters, as illustrated in Figures 2 and 3.

S.4.1 Additional Matrix Notation

For a $d \times d$ matrix $\boldsymbol{A}$ define diagonal$(\boldsymbol{A})$ to be the $d \times 1$ vector consisting of the diagonal entries of $\boldsymbol{A}$ and, provided $d \geq 2$, define vecbd$(\boldsymbol{A})$ to be the $\frac{1}{2}\,d(d-1)$ vector containing the entries of $\boldsymbol{A}$ that are below the diagonal of $\boldsymbol{A}$ in order from left to right and top to bottom. For example,

$$\text{diagonal}\left(\begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 6 \\ 11 \\ 16 \end{bmatrix} \quad \text{and} \quad \text{vecbd}\left(\begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 7 \\ 8 \\ 12 \end{bmatrix}.$$

In addition, if each of $\boldsymbol{a}$ and $\boldsymbol{b}$ are $d \times 1$ vectors then $\boldsymbol{a} \odot \boldsymbol{b}$ is $d \times 1$ vector of element-wise products and $\boldsymbol{a}/\boldsymbol{b}$ is $d \times 1$ vector of element-wise quotients. Similarly, $\log(\boldsymbol{a})$ and $\tanh(\boldsymbol{a})$ are obtained in an element-wise fashion.

S.4.2 Transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\omega}$

Given a $\frac{1}{2}d(d+1) \times 1$ vector $\boldsymbol{\theta}$, its corresponding $\boldsymbol{\omega}$ vector of the same length is found via the steps:

1. Obtain the spectral decomposition $\text{vech}^{-1}(\boldsymbol{\theta}) = \boldsymbol{U}_{\boldsymbol{\theta}}\,\text{diag}(\boldsymbol{\lambda}_{\boldsymbol{\theta}})\boldsymbol{U}_{\boldsymbol{\theta}}^T$.
2. Set $\boldsymbol{\Sigma} = \boldsymbol{U}_{\boldsymbol{\theta}}\text{diag}\{\exp(2\boldsymbol{\lambda}_{\boldsymbol{\theta}})\}\boldsymbol{U}_{\boldsymbol{\theta}}^T$.
3. (a) If $d = 1$ then $\boldsymbol{\omega} = \log(\sqrt{\boldsymbol{\Sigma}})$.

   (b) If $d > 1$ then

$$\boldsymbol{\omega} = \begin{bmatrix} \log\left(\sqrt{\text{diagonal}(\boldsymbol{\Sigma})}\right) \\ \tanh^{-1}\left(\text{vecbd}(\boldsymbol{\Sigma})\Big/\sqrt{\text{vecbd}\Big(\text{diagonal}(\boldsymbol{\Sigma})\text{diagonal}(\boldsymbol{\Sigma})^T\Big)}\right) \end{bmatrix}.$$

S.4.3 Transformation from $\boldsymbol{\omega}$ to $\boldsymbol{\theta}$

Given a $\frac{1}{2}d(d+1) \times 1$ vector $\boldsymbol{\omega}$, its corresponding $\boldsymbol{\theta}$ vector of the same length is found via the steps:

1. Form the $d \times d$ symmetric matrix $\boldsymbol{\Sigma}$ as follows:
   (a) If $d = 1$ then $\boldsymbol{\Sigma} = \exp(2\boldsymbol{\omega})$.
   (b) If $d > 1$ then let $\boldsymbol{\omega}_1$ denote the first d entries of $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_2$ denote the remaining $\frac{1}{2}d(d-1)$ entries of $\boldsymbol{\omega}$.
      i. Set diagonal$(\boldsymbol{\Sigma}) = \exp(2\boldsymbol{\omega}_1)$
      ii. Obtain the below-diagonal entries of $\boldsymbol{\Sigma}$ so that

$$\text{vecbd}(\boldsymbol{\Sigma}) = \tanh(\boldsymbol{\omega}_2) \odot \text{vecbd}\big(\exp(\boldsymbol{\omega}_1)\exp(\boldsymbol{\omega}_1)^T\big)$$

      holds. Obtain the above-diagonal entries of $\boldsymbol{\Sigma}$ such that symmetry of $\boldsymbol{\Sigma}$ is enforced.

2. Obtain the spectral decomposition: $\boldsymbol{\Sigma} = \boldsymbol{U}_{\boldsymbol{\Sigma}}\text{diag}(\boldsymbol{\lambda}_{\boldsymbol{\Sigma}})\boldsymbol{U}_{\boldsymbol{\Sigma}}^T$.

3. Obtain $\boldsymbol{\theta} = \text{vech}\Big(\frac{1}{2}\boldsymbol{U}_{\boldsymbol{\Sigma}}\text{diag}\{\log(\boldsymbol{\lambda}_{\boldsymbol{\Sigma}})\}\boldsymbol{U}_{\boldsymbol{\Sigma}}^T\Big)$.

S.4.4 Quasi-Newton Optimization-Based Confidence Interval Calculations

The steps for obtaining confidence intervals for each of the interpretable parameters are:

1. Obtain $(\underset{\sim}{\widehat{\boldsymbol{\beta}}}, \underset{\sim}{\widehat{\boldsymbol{\theta}}})$ using a quasi-Newton optimization routine applied the expectation propagation-approximate log-likelihood $\underset{\sim}{\ell}$ with unconstrained input parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$.

2. Obtain $\underset{\sim}{\widehat{\boldsymbol{\omega}}}$ corresponding to $\underset{\sim}{\widehat{\boldsymbol{\theta}}}$ using the steps given in Section S.4.2.

3. Call the quasi-Newton optimization routine with input parameters $(\boldsymbol{\beta}, \boldsymbol{\omega})$ instead of $(\boldsymbol{\beta}, \boldsymbol{\theta})$, and initial value $(\underset{\sim}{\widehat{\boldsymbol{\beta}}}, \underset{\sim}{\widehat{\boldsymbol{\omega}}})$. In this call, request that the Hessian matrix $\mathsf{H}\underset{\sim}{\ell}(\boldsymbol{\beta}, \boldsymbol{\omega})$ at the maximum $(\underset{\sim}{\widehat{\boldsymbol{\beta}}}, \underset{\sim}{\widehat{\boldsymbol{\omega}}})$ be computed. The steps given in Section S.4.3 are used to obtain the corresponding $(\boldsymbol{\beta}, \boldsymbol{\theta})$ vector for evaluation of $\underset{\sim}{\ell}$ via the version of $\underset{\sim}{\ell}$ used in 1. for the optimization.

4. Form $100(1-\alpha)\%$ confidence intervals for the entries of $(\boldsymbol{\beta}, \boldsymbol{\omega})$ using

$$\begin{bmatrix} \underset{\sim}{\widehat{\boldsymbol{\beta}}} \\ \underset{\sim}{\widehat{\boldsymbol{\omega}}} \end{bmatrix} \pm \Phi^{-1}(1 - \tfrac{1}{2}\,\alpha)\sqrt{-\text{diagonal}\big(\{\mathsf{H}\,\underset{\sim}{\ell}(\underset{\sim}{\widehat{\boldsymbol{\beta}}}, \underset{\sim}{\widehat{\boldsymbol{\omega}}})\}^{-1}\big)}.$$

5. Transform the confidence intervals limits for the $\boldsymbol{\omega}$ component, using the functions $\exp$ and $\tanh$, to instead correspond to the standard deviation and correlation parameters:

$$\begin{bmatrix} \sqrt{\text{diagonal}(\boldsymbol{\Sigma})} \\ \text{vecbd}(\boldsymbol{\Sigma})\Big/\sqrt{\text{vecbd}\Big(\text{diagonal}(\boldsymbol{\Sigma})\text{diagonal}(\boldsymbol{\Sigma})^T\Big)} \end{bmatrix}.$$

S.5 Details of Approximate Best Prediction

For the binary mixed model (1), the best prediction of $\boldsymbol{u}_i$ is

$$\begin{aligned}\mathrm{BP}(\boldsymbol{u}_i) = E(\boldsymbol{u}_i|\boldsymbol{y}) = E(\boldsymbol{u}_i|\boldsymbol{y}_i) &= \int_{\mathbb{R}^{d^{\mathrm{R}}}} \boldsymbol{u}_i\, p(\boldsymbol{u}_i|\boldsymbol{y}_i;\boldsymbol{\beta},\boldsymbol{\Sigma})\, d\boldsymbol{u}_i \\ &= \int_{\mathbb{R}^{d^{\mathrm{R}}}} \boldsymbol{u}_i \left\{ \frac{p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})p(\boldsymbol{u}_i;\boldsymbol{\Sigma})}{\int_{\mathbb{R}^{d^{\mathrm{R}}}} p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})p(\boldsymbol{u}_i;\boldsymbol{\Sigma})} \right\} d\boldsymbol{u}_i\end{aligned}$$

where $\boldsymbol{y}_i \equiv (y_{i1},\ldots,y_{in_i})$. Now note that Algorithm 1 involves replacement of

$$p(\boldsymbol{y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})p(\boldsymbol{u}_i;\boldsymbol{\Sigma}) \quad \text{by} \quad \exp\left\{ \begin{bmatrix} 1 \\ \boldsymbol{u}_i \\ \mathrm{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \end{bmatrix}^T \underset{\sim}{\widehat{\boldsymbol{\eta}}}_i \right\}$$

where $\underset{\sim}{\widehat{\boldsymbol{\eta}}}_i$ is defined by (16). This leads to the approximation

$$\begin{aligned}&\underset{\sim}{\mathrm{BP}}(\boldsymbol{u}_i) = E(\underset{\sim}{\widehat{\boldsymbol{u}}}_i) \text{ where } \underset{\sim}{\widehat{\boldsymbol{u}}}_i \text{ is Multivariate Normal with natural parameter } \underset{\sim}{\widehat{\boldsymbol{\eta}}}_i \\ &\quad = -\tfrac{1}{2}\left\{\mathrm{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\underset{\sim}{\widehat{\boldsymbol{\eta}}}_{i2}\right)\right\}^{-1}\underset{\sim}{\widehat{\boldsymbol{\eta}}}_{i1}.\end{aligned}$$

Using (13.7) of McCulloch, Searle & Neuhaus (2008), the covariance matrix of $\mathrm{BP}(\boldsymbol{u}_i) - \boldsymbol{u}_i$ is

$$\mathrm{Cov}\{\mathrm{BP}(\boldsymbol{u}_i) - \boldsymbol{u}_i\} = E_{\boldsymbol{y}_i}\{\mathrm{Cov}(\boldsymbol{u}_i|\boldsymbol{y}_i)\}.$$

The expectation propagation approximation of $\mathrm{Cov}(\boldsymbol{u}_i|\boldsymbol{y}_i)$ is

$$\underset{\sim}{\mathrm{Cov}}(\boldsymbol{u}_i|\boldsymbol{y}) = -\tfrac{1}{2}\left\{\mathrm{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\underset{\sim}{\widehat{\boldsymbol{\eta}}}_{i2}\right)\right\}^{-1}.$$

However, approximation of $\mathrm{Cov}\{\mathrm{BP}(\boldsymbol{u}_i) - \boldsymbol{u}_i\}$ is hindered by the expectation over the $\boldsymbol{y}_i$ vector.