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:*

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

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

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

Proof of Lemma 1. Note that the integrals on the left-hand side in (S.1)–(S.3) 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.3). Simplification 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

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

We then note that,

$$\begin{aligned} E\{g(s_1)s_2s_3\} &= \int_{-\infty}^{\infty} g(s_1) \left\{ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s_2s_3 p(s_2s_3|s_1) ds_2 ds_3 \right\} p(s_1) ds_1 \\ &= \int_{-\infty}^{\infty} g(s_1) \left\{ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\text{Cov}(s_2, s_3|s_1) + E(s_2|s_1)E(s_3|s_1)\} ds_2 ds_3 \right\} p(s_1) ds_1 \end{aligned}$$

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

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

Result (S.3) 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.3), the (i, j) entry of the right-hand side of (S.6) is

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

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.

∎

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

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

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

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

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

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

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

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

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

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

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 $\mathrm{KL}\big(f\|g(\cdot;\boldsymbol{\eta})\big)$. Therefore,

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

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

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

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

$$\begin{cases} \boldsymbol{\eta}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\eta}_2 = -\frac{1}{2}\boldsymbol{D}_d^T \text{vec}(\boldsymbol{\Sigma}^{-1}) \end{cases} \quad \text{and} \quad \begin{cases} \boldsymbol{\mu} = -\frac{1}{2}\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\right)\}^{-1}\boldsymbol{\eta}_1 \\ \boldsymbol{\Sigma} = -\frac{1}{2}\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2\right)\}^{-1} \end{cases} \tag{S.9}$$

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

we have

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

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

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

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:

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

where the linear and quadratic coefficient updates are

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

and the constant coefficient update is

$$\begin{aligned}\left(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\right)_{0} \longleftarrow C_{\text{probit}}\Big(&\left(\eta_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\right)_{-0}, \left(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})\to\boldsymbol{u}_i}\right)_{-0} \\ &+\left(\eta_{\boldsymbol{u}_i\to p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta})}\right)_{-0}; (2y_{ij}-1)(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^{\text{F}}), (2y_{ij}-1)\boldsymbol{x}_{ij}^{\text{R}}\Big).\end{aligned}$$

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

The second definition in (S.11) gives

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

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

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

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

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

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

which leads to

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

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

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

and

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

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

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

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 $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$:

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

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

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

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

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

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

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

with the proj[·] 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 $\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}$ that arises from this substitution is then given by

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

By matching coefficients of like terms we arrive at

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

where

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

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

$$\Big(\eta_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}\Big)_{-0}$$

implying that the first entry of $\eta^{\text{start}}_{p(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\beta}) \to \boldsymbol{u}_i}$ 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 4.

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

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

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 $\text{diagonal}(\boldsymbol{\Sigma}) = \exp(2\boldsymbol{\omega}_1)$.
      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 $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\theta}}})$ 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 $\widehat{\underset{\sim}{\boldsymbol{\omega}}}$ corresponding to $\widehat{\underset{\sim}{\boldsymbol{\theta}}}$ 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 $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\omega}}})$. In this call, request that the Hessian matrix $\mathsf{H}\underset{\sim}{\ell}(\boldsymbol{\beta}, \boldsymbol{\omega})$ at the maximum $(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\omega}}})$ 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
$$\begin{bmatrix} \widehat{\underset{\sim}{\boldsymbol{\beta}}} \\ \widehat{\underset{\sim}{\boldsymbol{\omega}}} \end{bmatrix} \pm \Phi^{-1}(1 - \tfrac{1}{2}\alpha)\sqrt{-\text{diagonal}\big(\{\mathsf{H}\underset{\sim}{\ell}(\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \widehat{\underset{\sim}{\boldsymbol{\omega}}})\}^{-1}\big)}.$$
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

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

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

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

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

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

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

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

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

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

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

Additional References

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

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