Appendix to

"Gaussian Variational Approximate Inference for Generalized Linear Mixed Models"

BY J.T. ORMEROD AND M.P. WAND

3rd February, 2011

Appendix A: Computational Details

A.1 Gauss-Hermite Quadrature for Evaluation of $B^{(r)}$

Here we briefly describe numerical evaluation of $B^{(r)}$ using the adaptive Gauss-Hermite quadrature procedure developed by Liu & Pierce (1994) for approximating positive Gaussian integrals.

For $r = 0, 1, 2, \ldots$ we have

$$\begin{aligned} B^{(r)}(\mu,\sigma^2) &= \int_{-\infty}^{\infty} \frac{b^{(r)}(\mu+\sigma x)\phi(x)}{\phi_{\sigma^*}(x-\mu^*)}\phi_{\sigma^*}(x-\mu^*)dx \\ &= \int_{-\infty}^{\infty} \left[\sqrt{2}\sigma^* b^{(r)}(\mu+\sigma(\mu^*+\sqrt{2}\sigma^* x))\phi(\mu^*+\sqrt{2}\sigma^* x)e^{x^2}\right] e^{-x^2}dx \end{aligned}$$

for any μ^* and σ^*. We choose μ^* and σ^* so that the integrand of $B^{(0)}(\mu,\sigma^2) = B(\mu,\sigma^2)$ is "most Gaussian" in shape so that

$$\mu^* = \underset{x}{\text{argmax}}\,\{b(\mu+\sigma x)\phi(x)\} \quad \text{and} \quad \sigma^* = -\left\{\left[\frac{d^2}{dx^2}\log\{b(\mu+\sigma x)\phi(x)\}\right]_{x=\mu^*}\right\}^{-1/2}.$$

We use the values of μ^* and σ^* corresponding to $r = 0$ because it is both computationally cheaper to evaluate μ^* and σ^* once and because $b^{(r)}$ may not be positive everywhere (potentially making the corresponding evaluation of σ^* problematic). We then apply Gauss-Hermite quadrature which uses

$$\int_{-\infty}^{\infty} g(x)e^{-x^2}dx = \sum_{k=1}^{N} w_k g(x_k) + \frac{\sqrt{\pi}N!}{2^N(2N)!}g^{(2N)}(\xi), \ \text{for some } -\infty < \xi < \infty$$

and some integer N. This is exact when $g(x)$ is a polynomial of degree $2N$ or less. Hence, we may approximate $B^{(r)}(\mu,\sigma)$ by

$$B^{(r)}(\mu,\sigma) \approx \sum_{k=1}^{N} w_k^* b^{(r)}(\mu+\sigma x_k^*)\phi(x_k^*)$$

where $w_k^* = \sqrt{2}\sigma^* w_k e^{x_k^2}$ and $x_k^* = \mu^* + \sqrt{2}\sigma^* x_k$ and the w_k and x_k values are the weights and abscissa of standard (or non-adaptive) Gauss-Hermite quadrature respectively.

There are several ways of obtaining w_j and x_j in practice. Tables for these values can be obtained from Abramowitz & Stegun (1972, Chapter 25). The function `gauss.quad()` in the `R` package `statmod` (Smyth, 2009) may also be used to find the w_js and x_js.

A.2 Notation Useful for Derivative and Hessian Expressions

Let f be a real-valued function in the $d \times 1$ vector $\boldsymbol{x} = [x_1, \ldots, x_d]^T$. Then the derivative vector $\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})$, is the $1 \times d$ with ith entry $\partial f(\boldsymbol{x})/\partial x_i$. The corresponding Hessian matrix is given by $\mathsf{H}_{\boldsymbol{x}} f(\boldsymbol{x}) = \mathsf{D}_{\boldsymbol{x}}\{\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})\}^T$. We extend the B notation to higher derivatives as follows:

$$B^{(r)}(\mu, \sigma^2) \equiv \int_{-\infty}^{\infty} b^{(r)}(\sigma x + \mu)\phi(x)\,dx.$$

Define $\mathcal{Q}(\boldsymbol{A}) \equiv (\boldsymbol{A} \otimes \mathbf{1}^T) \odot (\mathbf{1}^T \otimes \boldsymbol{A})$ where $\boldsymbol{A} \odot \boldsymbol{B}$ is the element-wise product of two equi-sized matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Next, we let $\boldsymbol{D}_p$ denote the *duplication matrix* of order p defined through the relationship $\text{vec}(\boldsymbol{A}) = \boldsymbol{D}_p \text{vech}(\boldsymbol{A})$ for a symmetric $p \times p$ matrix $\boldsymbol{A}$. Lastly, for each $1 \leq i \leq m$, let $\mathcal{B}^{(r)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) \equiv B^{(r)}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_i, \text{dg}(\boldsymbol{Z}_i\boldsymbol{\Lambda}_i\boldsymbol{Z}_i^T))$.

A.3 Derivative Vector of Lower Bound on Log-Likelihood

The derivative vector of $\underline{\ell} \equiv \underline{\ell}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ with respect to $(\boldsymbol{\beta}, \text{vech}(\boldsymbol{\Sigma}), \boldsymbol{\mu}_1, \text{vech}(\boldsymbol{\Lambda}_1), \ldots, \boldsymbol{\mu}_m, \text{vech}(\boldsymbol{\Lambda}_m))$ is $\mathsf{D}\,\underline{\ell} = \left[\mathsf{D}_{\boldsymbol{\beta}}\,\underline{\ell}, \mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\underline{\ell}, \mathsf{D}_{\boldsymbol{\mu}_1}\,\underline{\ell}, \mathsf{D}_{\text{vech}(\boldsymbol{\Lambda}_1)}\,\underline{\ell}, \cdots, \mathsf{D}_{\boldsymbol{\mu}_m}\,\underline{\ell}, \mathsf{D}_{\text{vech}(\boldsymbol{\Lambda}_m)}\,\underline{\ell}\right]$. We now give matrix algebraic expressions for each of these components:

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\beta}}\,\underline{\ell} &= \sum_{i=1}^{m}\{\boldsymbol{y}_i - \mathcal{B}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}^T\boldsymbol{X}_i, \\
\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} &= \tfrac{1}{2}\sum_{i=1}^{m}\text{vec}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i)\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\}^T\boldsymbol{D}_K,
\end{aligned}$$

and for $1 \leq i \leq m$

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\mu}_i}\,\underline{\ell} &= \{\boldsymbol{y}_i - \mathcal{B}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}^T\boldsymbol{Z}_i - \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}, \\
\mathsf{D}_{\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= \tfrac{1}{2}\text{vec}[\boldsymbol{\Lambda}_i^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{Z}_i^T\text{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{Z}_i]^T\boldsymbol{D}_K.
\end{aligned}$$

Also, define $\boldsymbol{\theta} \equiv [\boldsymbol{\beta}^T, \text{vech}(\boldsymbol{\Sigma})^T]^T$ and $\boldsymbol{\xi} = [\boldsymbol{\xi}_1^T, \ldots, \boldsymbol{\xi}_m^T]^T$ to be the vectors containing the unique model and variational parameters, respectively where $\boldsymbol{\xi}_i = [\boldsymbol{\mu}_i^T, \text{vech}(\boldsymbol{\Lambda}_i)^T]^T$, $1 \leq i \leq m$. Finally, define the gradient vectors $\boldsymbol{g}_{\boldsymbol{\theta}} \equiv [(\mathsf{D}_{\boldsymbol{\beta}}\,\underline{\ell}), (\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\underline{\ell})]^T$ and
$\boldsymbol{g}_{\boldsymbol{\xi}_i} \equiv [(\mathsf{D}_{\boldsymbol{\mu}_i}\,\underline{\ell}), (\mathsf{D}_{\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell})^T]^T$.

A.4 Hessian Matrix of Lower Bound on Log-Likelihood

The Hessian matrix of $\underline{\ell}$ with respect to $(\boldsymbol{\beta}, \text{vech}(\boldsymbol{\Sigma}), \boldsymbol{\mu}_1, \text{vech}(\boldsymbol{\Lambda}_1), \ldots, \boldsymbol{\mu}_m, \text{vech}(\boldsymbol{\Lambda}_m))$ is $\mathsf{H}\,\underline{\ell} = \left[\begin{smallmatrix} \boldsymbol{H}_{\boldsymbol{\theta\theta}} & \boldsymbol{H}_{\boldsymbol{\theta\xi}} \\ \boldsymbol{H}_{\boldsymbol{\theta\xi}}^T & \boldsymbol{H}_{\boldsymbol{\xi\xi}} \end{smallmatrix}\right]$ where in keeping with notation given in Section Appendix A.3 for the $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the Hessian matrix components are defined as $\boldsymbol{H}_{\boldsymbol{\theta\theta}} \equiv \text{blockdiag}(\mathsf{H}_{\boldsymbol{\theta\theta}}\,\underline{\ell}, \mathsf{H}_{\text{vech}(\boldsymbol{\Sigma})\text{vech}(\boldsymbol{\Sigma})}\,\underline{\ell})$, $\boldsymbol{H}_{\boldsymbol{\theta\xi}} = [\boldsymbol{H}_{\boldsymbol{\theta\xi}_1}, \ldots, \boldsymbol{H}_{\boldsymbol{\theta\xi}_m}]$, $\boldsymbol{H}_{\boldsymbol{\xi\xi}} = \text{blockdiag}(\boldsymbol{H}_{\boldsymbol{\xi}_1\boldsymbol{\xi}_1}, \ldots, \boldsymbol{H}_{\boldsymbol{\xi}_m\boldsymbol{\xi}_m})$. For $1 \le i \le m$

$$\boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i} \equiv \left[\begin{array}{cc} \mathsf{H}_{\boldsymbol{\mu}_i\boldsymbol{\mu}_i}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\mu}_i\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} \\ \mathsf{H}_{\text{vech}(\boldsymbol{\Lambda}_i)\boldsymbol{\mu}_i}\,\underline{\ell} & \mathsf{H}_{\text{vech}(\boldsymbol{\Lambda}_i)\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} \end{array}\right] \quad \text{and} \quad \boldsymbol{H}_{\boldsymbol{\theta\xi}_i} \equiv \left[\begin{array}{cc} \mathsf{H}_{\boldsymbol{\beta\mu}_i}\,\underline{\ell} & \mathsf{H}_{\boldsymbol{\beta}\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} \\ \mathsf{H}_{\text{vech}(\boldsymbol{\Sigma})\boldsymbol{\mu}_i}\,\underline{\ell} & \mathsf{H}_{\text{vech}(\boldsymbol{\Sigma})\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} \end{array}\right],$$

where

$$\begin{aligned} \mathsf{H}_{\boldsymbol{\theta\theta}}\,\underline{\ell} &= -\sum_{i=1}^m \boldsymbol{X}_i^T \text{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{X}_i, \\ \mathsf{H}_{\text{vech}(\boldsymbol{\Sigma})\text{vech}(\boldsymbol{\Sigma})}\,\underline{\ell} &= \tfrac{1}{2}\boldsymbol{D}_K^T\Big(m(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) - \sum_{i=1}^m [\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{M}_i + \boldsymbol{M}_i \otimes \boldsymbol{\Sigma}^{-1}]\Big)\boldsymbol{D}_K, \end{aligned}$$

with $\boldsymbol{M}_i = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i)\boldsymbol{\Sigma}^{-1}$ and, for $1 \le i \le m$,

$$\begin{aligned} \mathsf{H}_{\boldsymbol{\beta\mu}_i}\,\underline{\ell} &= -\boldsymbol{X}_i^T \text{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{Z}_i, \\ \mathsf{H}_{\boldsymbol{\beta}\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= -\tfrac{1}{2}\boldsymbol{X}_i^T \text{diag}\{\mathcal{B}^{(3)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\mathcal{Q}(\boldsymbol{Z}_i)\boldsymbol{D}_K, \\ \mathsf{H}_{\text{vech}(\boldsymbol{\Sigma})\boldsymbol{\mu}_i}\,\underline{\ell} &= \boldsymbol{D}_K^T\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i) \otimes \boldsymbol{\Sigma}^{-1}\}, \\ \mathsf{H}_{\text{vech}(\boldsymbol{\Sigma})\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= \tfrac{1}{2}\boldsymbol{D}_K^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{D}_K, \\ \mathsf{H}_{\boldsymbol{\mu}_i\boldsymbol{\mu}_i}\,\underline{\ell} &= -\boldsymbol{Z}_i^T \text{diag}\{\mathcal{B}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\boldsymbol{Z}_i - \boldsymbol{\Sigma}^{-1}, \\ \mathsf{H}_{\boldsymbol{\mu}_i\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= -\tfrac{1}{2}\boldsymbol{Z}_i^T \text{diag}\{\mathcal{B}^{(3)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\mathcal{Q}(\boldsymbol{Z}_i)\boldsymbol{D}_K, \\ \mathsf{H}_{\text{vech}(\boldsymbol{\Lambda}_i)\text{vech}(\boldsymbol{\Lambda}_i)}\,\underline{\ell} &= -\tfrac{1}{4}\boldsymbol{D}_K^T[\mathcal{Q}(\boldsymbol{Z}_i)^T \text{diag}\{\mathcal{B}^{(4)}(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\}\mathcal{Q}(\boldsymbol{Z}_i) + 2(\boldsymbol{\Lambda}_i^{-1} \otimes \boldsymbol{\Lambda}_i^{-1})]\boldsymbol{D}_K. \end{aligned}$$

A.5 Newton-Raphson Scheme

We are a now in a position to describe an efficient Newton-Raphson scheme for solving the maximization problem (5) in Section 3. In particular, we make use of the block-diagonal structure in the $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ section of $\mathsf{H}_{\boldsymbol{\beta\beta}}\underline{\ell}$ to reduce the number of operations to $O(m)$. Define

$$\boldsymbol{s}_{\boldsymbol{\theta\xi}} \equiv \left(\boldsymbol{H}_{\boldsymbol{\theta\theta}} - \sum_{i=1}^m \boldsymbol{H}_{\boldsymbol{\theta\xi}_i}\boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{-1}\boldsymbol{H}_{\boldsymbol{\theta\xi}_i}^T\right)^{-1}\left(\boldsymbol{g}_{\boldsymbol{\theta}} - \sum_{i=1}^m \boldsymbol{H}_{\boldsymbol{\theta\xi}_i}\boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{-1}\boldsymbol{g}_{\boldsymbol{\xi}}\right).$$

and let $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\xi}_i^{(0)}$, $1 \le i \le m$, be starting values of the relevant parameter vectors and let a superscript of (t) denote the same vectors after t iterations of the Newton-Raphson algorithm. Then the updates are given by:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{s}_{\boldsymbol{\theta\xi}}^{(t)} \quad \text{and} \quad \boldsymbol{\xi}_i^{(t+1)} = \boldsymbol{\xi}_i^{(t)} - (\boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{(t)})^{-1}(\boldsymbol{g}_{\boldsymbol{\xi}_i}^{(t)} - \boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\theta}}^{(t)}\boldsymbol{s}_{\boldsymbol{\theta\xi}}^{(t)}), 1 \le i \le m. \tag{1}$$

Note that these updates involves inversion of 'small' matrices – i.e. those of dimension similar to $\boldsymbol{\beta}$ and the $\boldsymbol{u}_i$. The Newton-Raphson scheme then involves repeated application of (1) until convergence. If an update results in $\mathbf{\Sigma}$ being negative definite we switch the parameterization $\mathbf{\Sigma} = \boldsymbol{R}^T\boldsymbol{R}$ where

$$\boldsymbol{R} = \begin{bmatrix} e^{r_{11}} & r_{12} & \cdots & r_{1K} \\ 0 & e^{r_{22}} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{r_{KK}} \end{bmatrix}$$

and the r_{ij}s are unconstrained. This is a minimal unconstrained parameterization for $\mathbf{\Sigma}$ which simultaneously ensures identifiability of the r_{ij}s and positive definiteness of $\mathbf{\Sigma}$. An analogous reparameterization is made if any of the $\mathbf{\Lambda}_i$s are negative definite. The derivative and Hessian formula of Appendix A.3 and A.4 under this parameterization can be derived with a little modification. Finally, to increase the robustness of the Newton-Raphson algorithm we incorporated step-halving to ensure that the variational log-likelihood increased at each step of the algorithm.

A.6 Asymptotic Covariance Matrix

Results used in Appendix A.5 that take advantage of diagonal structure in $\mathsf{H}\underline{\ell}(\boldsymbol{\beta}, \mathbf{\Sigma}, \boldsymbol{\mu}, \mathbf{\Lambda})$ can also be used to obtain a streamlined expression for the asymptotic covariance matrix of the model parameters. These lead to $\widehat{\text{Asy.Cov}}(\widehat{\boldsymbol{\theta}}) = -\left[\boldsymbol{H}_{\boldsymbol{\theta\theta}} - \sum_{i=1}^m \boldsymbol{H}_{\boldsymbol{\theta\xi}_i}\boldsymbol{H}_{\boldsymbol{\xi}_i\boldsymbol{\xi}_i}^{-1}\boldsymbol{H}_{\boldsymbol{\theta\xi}_i}^T\right]^{-1}$ which allows standard errors to be computed in $O(m)$ operations.

Appendix B: Proofs

B.1 Proof of Theorem 1

The proof relies on straightforward algebra and the following lemmas:

LEMMA 1: [THEOREM 22 OF MAGNUS & NEUDECKER (1988)] *Let $\boldsymbol{A}$ be a positive definite matrix and $\boldsymbol{B}$ be a positive semidefinite matrix of the same dimensions as $\boldsymbol{A}$. Then $|\boldsymbol{A}+\boldsymbol{B}| \geq |\boldsymbol{A}|$ with equality if and only if $\boldsymbol{B} = \mathbf{0}$.*

LEMMA 2. [SPECIAL CASE OF THEOREM 7.7.6 OF HORN & JOHNSON (1985)] Let the symmetric matrix $\boldsymbol{A}$ be partitioned as

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}^T & \boldsymbol{A}_{22} \end{bmatrix}$$

where $\boldsymbol{A}_{11}$ is square and invertible. Then $\boldsymbol{A}$ is positive definite if any only if both $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22} - \boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$ are positive definite.

LEMMA 3. *Let $\mathbf{\Psi}$ be a symmetric, positive definite $mK \times mK$ matrix. Given the $K \times K$ blocks $\mathbf{\Psi}_i$, $1 \leq i \leq m$, down the main diagonal of $\mathbf{\Psi}$, the determinant $|\mathbf{\Psi}|$ is uniquely maximized by setting all other entries of $\mathbf{\Psi}$ equal to zero.*

Proof of Lemma 3. Consider the partition of $\mathbf{\Psi}$:

$$\mathbf{\Psi} = \begin{bmatrix} \widetilde{\mathbf{\Psi}} & \boldsymbol{C} \\ \boldsymbol{C}^T & \boldsymbol{D} \end{bmatrix}.$$

where $\widetilde{\boldsymbol{\Psi}}$ is an $m(K-1)\times m(K-1)$ matrix, $\boldsymbol{C}$ is an $m(K-1)\times K$ matrix and $\boldsymbol{D}$ is a $K\times K$ matrix. Then, from a standard result on determinants of partitioned matrices, $|\boldsymbol{\Psi}| = |\widetilde{\boldsymbol{\Psi}}||\boldsymbol{D}-\boldsymbol{C}^T\widetilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{C}|$. Since $\boldsymbol{\Psi}$ is positive definite then, from Lemma 2, the matrices $\widetilde{\boldsymbol{\Psi}}$, $\boldsymbol{D}-\boldsymbol{C}^T\widetilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{C}$ must also be positive definite. Also $\boldsymbol{\Psi}^{-1}$ is positive definite, which implies that $\boldsymbol{C}^T\widetilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{C}$ is positive semidefinite.

We shall prove the lemma by induction over m. The lemma holds trivially when $m=1$. By the induction hypothesis, we may assume that $|\widetilde{\boldsymbol{\Psi}}|$ is uniquely maximized by taking the off block-diagonal components of $\widetilde{\boldsymbol{\Psi}}$ to vanish. For $\widetilde{\boldsymbol{\Psi}}$ and $\boldsymbol{D}$ fixed, we can use Lemma 1 with $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{C}^T\widetilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{C}$ and $\boldsymbol{B}=\boldsymbol{C}^T\widetilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{C}$ to show that $|\boldsymbol{D}-\boldsymbol{C}^T\widetilde{\boldsymbol{\Psi}}\boldsymbol{C}|$ is uniquely maximized by taking $\boldsymbol{C}=\boldsymbol{0}$. The lemma then follows by induction.

□

The right-hand side of (6) is Section 3, with q set to the $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ density and $p(\boldsymbol{y}|\boldsymbol{u})$ given by (3) in Section 3, is

$$\begin{aligned}\underline{\ell}(\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\mu},\boldsymbol{\Lambda}) \;&=\; \tfrac{mK}{2}+\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\mu})-\boldsymbol{1}^T B(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\mu},\mathrm{dg}(\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T))+\boldsymbol{1}^T c(\boldsymbol{y})\\ &\quad -\tfrac{1}{2}\{\boldsymbol{\mu}^T\boldsymbol{G}^{-1}\boldsymbol{\mu}+\mathrm{tr}(\boldsymbol{G}^{-1}\boldsymbol{\Lambda})\}+\tfrac{1}{2}\log|\boldsymbol{G}^{-1}\boldsymbol{\Lambda}|.\end{aligned}$$

Now consider the special case of the grouped data GLMM (1) in Section 2. Applying the definitions of $\boldsymbol{y}_i$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ given in Section 2 and setting $\boldsymbol{Z}=\mathrm{blockdiag}_{1\le i\le m}(\boldsymbol{Z}_i)$ and $\boldsymbol{G}=\boldsymbol{I}\otimes\boldsymbol{\Sigma}$ we obtain

$$\begin{aligned}\underline{\ell}(\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\mu},\boldsymbol{\Lambda}) \;&=\; \textstyle\sum_{i=1}^m\Big[\boldsymbol{y}_i^T(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\boldsymbol{\mu}_i)-\boldsymbol{1}_i^T B(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\boldsymbol{\mu}_i,\mathrm{dg}(\boldsymbol{Z}_i\boldsymbol{\Lambda}_i\boldsymbol{Z}_i^T))+\boldsymbol{1}_i^T c(\boldsymbol{y}_i)\\ &\quad +\tfrac{1}{2}\{\log|\boldsymbol{\Sigma}^{-1}|-\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i-\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i)\}+\tfrac{K}{2}\Big]+\tfrac{1}{2}\log|\boldsymbol{\Lambda}|\end{aligned}$$

By Lemma 3, $|\boldsymbol{\Lambda}|$ is maximal for $\boldsymbol{\Lambda}=\mathrm{blockdiag}(\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_m)$. Hence, there is no loss from replacement of $\frac{1}{2}\log|\boldsymbol{\Lambda}|$ by $\frac{1}{2}\sum_{i=1}^m\log|\boldsymbol{\Lambda}_i|$, leading to the expression (4) in Section 3.

B.2 Consistency of Gaussian Variational Approximation for Simple Generalized Linear Mixed Models

Hall, Ormerod & Wand (2011) studied the theoretical properties of GVA in the Poisson case with $(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\boldsymbol{u}_i)_j=\beta_0+u_i+\beta_1X_{ij}$ (the design structure of Example 1 in 2) and $n_i=n$ for all $1\le i\le m$. Note that we have changed from using x_{ij}s to X_{ij}s so as to treat the design as random. Extending these theoretical properties for the general one-parameter exponential family case is not a simple matter. This is due to the fact the proofs in Hall *et al.* (2011) rely on, amongst other things, that $E(Y_{ij}|X_{ij},U_i)=\exp(\beta_0+U_i+\beta_1X_{ij})$ which allows separation of the e^{β_0}, e^{U_i} and $e^{\beta_1X_{ij}}$ through multiplication and the fact that the exponential function is closely connected to moment generating functions. Thus, in this section, instead of pursuing rigorous proofs for this extension, we provide heuristic arguments for the consistency for simple generalized linear mixed models with

$$p(Y_{ij}|X_{ij},U_i)=\exp\{Y_{ij}(\beta_0+U_i+\beta_1X_{ij})-b(\beta_0+U_i+\beta_1X_{ij})+c(Y_{ij})\}$$

for $1 \leq i \leq m$, $1 \leq j \leq n$ where n some fixed positive integer. For this case the variational lower bound on the log-likelihood is given by

$$\begin{aligned}\underline{\ell}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}) &= \sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}(\beta_0 + \beta_1 X_{ij} + \mu_i) - B^{(0)}(\beta_0 + \beta_1 X_{ij} + \mu_i, \lambda_i) + c(Y_{ij}) \\ &\quad -\frac{m}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{m}\log(\lambda_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\mu_i^2 + \lambda_i)\end{aligned}$$

where $B^{(r)}(\mu, \sigma^2) = \int_{-\infty}^{\infty} b^{(r)}(\mu + \sigma x)\phi(x)\,dx$. Next, let

$$(\underline{\widehat{\boldsymbol{\beta}}}, \underline{\widehat{\sigma}}^2, \underline{\widehat{\boldsymbol{\mu}}}, \underline{\boldsymbol{\lambda}}) = \underset{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}}{\text{argmax}}\, \underline{\ell}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}).$$

We are interested in the values of $(\underline{\widehat{\boldsymbol{\beta}}}, \underline{\widehat{\sigma}}^2, \underline{\widehat{\boldsymbol{\mu}}}, \underline{\boldsymbol{\lambda}})$ values as n and m approach ∞.

The heuristic proof for the consistency and rates of convergence for the estimators $\underline{\widehat{\boldsymbol{\beta}}}$ and $\underline{\widehat{\sigma}}^2$ involves the following three steps:

1. Simplifying GVA, using asymptotic arguments, to show that the GVA is similar to a fixed effects model.
2. Using likelihood theory the rates of convergence of the parameters in the fixed effects model are found.
3. The rates of convergence of the GVA estimates are then shown to be asymptotically close to the parameter estimates in the fixed effects model.

First we make the following assumptions:

(A1) For $1 \leq j \leq n$, the triples (X_{ij}, Y_{ij}, U_i) are independent and identically distributed as (X_i, Y_i, U_i), $1 \leq i \leq m$, say, which in turn is distributed as (X, Y, U);

(A2) the random variables X and U are independent;

(A3) the sets of variables $S_i = \{(X_{ij}, Y_{ij}, U_i) : 1 \leq j \leq n\}$, for $1 \leq i \leq m$, are independent and identically distributed;

(A4) each Y_{ij} , conditional on both X_{ij} and U_i, has the distribution

$$p(Y_{ij}|X_{ij}, U_i) = \exp\left\{Y_{ij}(\beta_0^0 + U_i + \beta_1^0 X_{ij}) - b(\beta_0^0 + U_i + \beta_1^0 X_{ij}) + c(Y_{ij})\right\}$$

where β_0^0 and β_1^0 denote the true values of β_0 and β_1 respectively;

(A5) each U_i is normal $N(0, (\sigma^2)^0)$, where $(\sigma^2)^0$ denotes the true value of σ^2, for some $(\sigma^2)^0 > 0$;

(A6) when choosing $(\underline{\widehat{\beta}}_0, \underline{\widehat{\beta}}_1, \underline{\widehat{\sigma}}^2, \underline{\widehat{\boldsymbol{\mu}}}, \underline{\widehat{\boldsymbol{\lambda}}})$ we search in a bounded rectangular region $[-C_1, C_1] \times [-C_1, C_1] \times [C_1^{-1}, C_1] \times [-C_2, C_2]^m \times [C_1^{-1}, C_1]^m$ where $C_1 > \max(|\beta_0^0|, |\beta_1^0|, (\sigma^2)^0, 1/(\sigma^2)^0)$ and C_2 is a sufficiently large (but finite) constant;

(A7) the random variable X is bounded;

(A8) and the functions

$$\psi_k(t_0, t_1) = E_X\left[X^k b^{(2)}(t_0 + t_1 X)\right]$$

for $k = 0, 1, 2$ are well defined for all real t_0, t_1.

The first (A1)-(A5) were used explicitly by Hall *et al.* (2011) whilst (A6) is an extension of (A11) of Hall *et al.* (2011). Assumptions (A7) and (A8) are imposed to simplify results presented here. These assumptions, in particular (A7), may be weakened, but at the expense of a longer proofs.

Let $\underline{\ell} \equiv \underline{\ell}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})$ then

$$\frac{\partial \underline{\ell}}{\partial \beta_0} = \sum_{i=1}^m \sum_{j=1}^n \left\{ Y_{ij} - B^{(1)}(\beta_0 + \beta_1 X_{ij} + \mu_i, \lambda_i) \right\}, \tag{2}$$

$$\frac{\partial \underline{\ell}}{\partial \beta_1} = \sum_{i=1}^m \sum_{j=1}^n \left\{ X_{ij} Y_{ij} - X_{ij} B^{(1)}(\beta_0 + \beta_1 X_{ij} + \mu_i, \lambda_i) \right\}, \tag{3}$$

$$\frac{\partial \underline{\ell}}{\partial \sigma^2} = -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^m (\mu_i^2 + \lambda_i), \tag{4}$$

and for $1 \le i \le m$ we have

$$\frac{\partial \underline{\ell}}{\partial \mu_i} = \sum_{j=1}^n \left\{ Y_{ij} - B^{(1)}(\beta_0 + \beta_1 X_{ij} + \mu_i, \lambda_i) \right\} - \frac{\mu_i}{\sigma^2}, \tag{5}$$

$$\text{and} \quad \frac{\partial \underline{\ell}}{\partial \lambda_i} = \frac{1}{2}\left[\lambda_i^{-1} - \sigma^{-2} - \sum_{j=1}^n B^{(2)}(\beta_0 + \beta_1 X_{ij} + \mu_i, \lambda_i) \right]. \tag{6}$$

First we note some properties of the GVA estimators which hold regardless of whether n or m diverge. Setting equation (4) to zero and solving we obtain

$$\widehat{\underline{\sigma}}^2 = \sum_{i=1}^m \frac{\widehat{\underline{\mu}}_i^2 + \widehat{\underline{\lambda}}_i}{m}.$$

Subtracting (2) set to zero and the sum of (5) of i from 1 to m set to zero and solving we obtain

$$\sum_{i=1}^m \widehat{\underline{\mu}}_i = 0.$$

Finally, setting equation (6) to zero we can deduce that $\widehat{\underline{\lambda}}_i$ satisfies

$$\widehat{\underline{\lambda}}_i = \left[\widehat{\underline{\sigma}}^{-2} + \sum_{j=1}^n B^{(2)}(\widehat{\underline{\beta}}_0 + \widehat{\underline{\beta}}_1 X_{ij} + \widehat{\underline{\mu}}_i, \widehat{\underline{\lambda}}_i) \right]^{-1}$$

from which it is obvious that $0 < \widehat{\underline{\lambda}}_i < \widehat{\underline{\sigma}}^2$ (since $\widehat{\underline{\sigma}}^2$ is bounded away from zero and $B^{(2)}(\mu, \lambda)$ is positive for all μ and $\lambda > 0$).

We now consider the properties of $\widehat{\underline{\lambda}}_i$ and $B^{(r)}(\mu, \widehat{\underline{\lambda}}_i)$ as n diverges. First, let

$$\gamma_i = \inf_{X,\beta_0,\beta_1,\sigma^2,\mu_i,\lambda_i} B^{(2)}(\beta_0 + \beta_1 X + \mu_i, \lambda_i).$$

Then γ_i is bounded away from zero since X, β_0, β_1, σ^2, μ_i and λ_i are bounded, using (A6) and (A7) respectively, and the fact that $B^{(2)}(\cdot, \cdot)$ is bounded away from zero for finite arguments. Hence,

$$\widehat{\underline{\lambda}}_i \leq \left[\widehat{\underline{\sigma}}^{-2} + n\gamma_i\right]^{-1} = O_p(n^{-1}),\ 1 \leq i \leq m,$$

which follows from $\widehat{\underline{\sigma}}^2$ being positive and bounded (Assumption 6).

Next, using the fact that for all μ, λ and r the derivative of $B^{(r)}(\mu, \lambda)$ with respect to λ satisfies $\partial B^{(r)}(\mu, \lambda)/\partial\lambda = B^{(r+2)}(\mu, \lambda)/2$, the Taylor series expansion of $B^{(r)}(\mu, \lambda_i)$ around $\lambda_i = 0$ is given by

$$B^{(r)}(\mu, \lambda_i) = \sum_{k=0}^{\infty} \frac{B^{(r+2k)}(\mu, 0)\lambda_i^k}{2^k k!} = \sum_{k=0}^{\infty} \frac{b^{(r+2k)}(\mu)\lambda_i^k}{2^k k!}$$

where the right hand side follows from the fact that $B^{(r)}(\mu, 0) = b^{(r)}(\mu)$ for all real μ and integers r. Hence, for all μ and r we have

$$B^{(r)}(\mu, \widehat{\underline{\lambda}}_i) = b^{(r)}(\mu) + \tfrac{1}{4}b^{(r+2)}(\mu)\widehat{\underline{\lambda}}_i + O(\widehat{\underline{\lambda}}_i^2) = b^{(r)}(\mu) + O_p(n^{-1}). \tag{7}$$

Using this equation we can eliminate the λ_is so we can deduce that

$$\begin{aligned}
\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij} &= \frac{1}{mn}\left[\sum_{i=1}^{m}\sum_{j=1}^{n} b^{(1)}(\widehat{\underline{\beta}}_0 + \widehat{\underline{\beta}}_1 X_{ij} + \widehat{\underline{\mu}}_i)\right] + O_p(n^{-1}) \qquad (8)\\
\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij} &= \frac{1}{mn}\left[\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}b^{(1)}(\widehat{\underline{\beta}}_0 + \widehat{\underline{\beta}}_1 X_{ij} + \widehat{\underline{\mu}}_i)\right] + O_p(n^{-1}) \qquad (9)\\
\widehat{\underline{\sigma}}^2 &= \sum_{i=1}^{m} \frac{\widehat{\underline{\mu}}_i^2}{m} + O_p(n^{-1}) \qquad (10)
\end{aligned}$$

and for $1 \leq i \leq m$ we have

$$\frac{1}{n}\sum_{j=1}^{n} Y_{ij} = \frac{\widehat{\underline{\mu}}_i}{n\widehat{\underline{\sigma}}^2} + \frac{1}{n}\left[\sum_{j=1}^{n} b^{(1)}(\widehat{\underline{\beta}}_0 + \widehat{\underline{\beta}}_1 X_{ij} + \widehat{\underline{\mu}}_i)\right] + O_p(n^{-1}). \tag{11}$$

B.2.1 An Analogous Fixed Effects Model

Looking at equations (9) and (11) with $\widetilde{\beta}_{0i} = \beta_0 + \mu_i$ we observe that, ignoring $O_p(n^{-1})$ terms, these are the estimating equations for a generalized linear model (containing only fixed effects) where

$$p(Y_{ij}|X_{ij}, U_i) = \exp\left\{Y_{ij}(\widetilde{\beta}_{0i} + \widetilde{\beta}_1\, X_{ij}) - b(\widetilde{\beta}_{0i} + \widetilde{\beta}_1\, X_{ij}) + c(Y_{ij})\right\}$$

conditional on X_{ij} where $\widetilde{\beta}_1$ is a slope parameter and $\widetilde{\beta}_{0i}^0 = \beta_0^0 + U_i$ for $1 \le i \le m$, $1 \le j \le n$. The log-likelihood for such a model is

$$\ell(\widetilde{\boldsymbol{\beta}}_0, \widetilde{\beta}_1) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left\{ Y_{ij}(\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) - b^{(0)}(\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) \right\}$$

where $\widetilde{\boldsymbol{\beta}}_0 = (\beta_{01}, \ldots, \beta_{0m})$. It is important to note that this is a non-standard context where the number of parameters P is $P = m + 1$ and the number of observations N is $N = mn$ so that P is growing proportionally to N with N diverging.

Asymptotic properties of maximum likelihood estimates for exponential families with diverging number of parameters has been considered by Portnoy (1988) and Kakade *et al.* (2010). In particular Portnoy (1988) showed that the maximum likelihood estimators are consistent provided $\frac{P}{N} = \frac{m+1}{mn} \to 0$ which occurs if we let n diverge. Thus, under appropriate conditions, we can apply standard asymptotic results.

The maximum likelihood estimators $\widehat{\widetilde{\boldsymbol{\beta}}}_0$ and $\widehat{\widetilde{\beta}}_1$ of $\widetilde{\boldsymbol{\beta}}_0$ and $\widetilde{\beta}_1$ respectively satisfy

$$\begin{aligned} \sum_{i=1}^{m} \sum_{j=1}^{n} \left\{ X_{ij} Y_{ij} - X_{ij} b^{(1)}(\widehat{\widetilde{\beta}}_{0i} + \widehat{\widetilde{\beta}}_1 X_{ij}) \right\} &= 0 & (12) \\ \text{and} \quad = \sum_{j=1}^{n} \left\{ Y_{ij} - b^{(1)}(\widehat{\widetilde{\beta}}_{0i} + \widehat{\widetilde{\beta}}_1 X_{ij}) \right\} &= 0, \quad 1 \le i \le m. & (13) \end{aligned}$$

The non-zero second derivatives of ℓ with respect to $\widetilde{\boldsymbol{\beta}}_0$ and $\widetilde{\beta}_1$ are given by

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \widetilde{\beta}_1^2} &= -\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2 b^{(2)}(\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}), & (14) \\ \frac{\partial^2 \ell}{\partial \widetilde{\beta}_1 \partial \widetilde{\beta}_{0i}} &= -\sum_{j=1}^{n} X_{ij} b^{(1)}(\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) & (15) \\ \text{and} \quad \frac{\partial \ell}{\partial \widetilde{\beta}_{0i} \partial \widetilde{\beta}_{0k}} &= \begin{cases} -\sum_{j=1}^{n} b^{(2)}(\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) & \text{if } i = k, \\ 0 & \text{otherwise}, \end{cases} & (16) \end{aligned}$$

for $1 \le i \le m$ and $1 \le k \le m$. The Fisher information matrix is given by

$$I(\widetilde{\beta}_1, \widetilde{\boldsymbol{\beta}}_0) = \begin{bmatrix} \sum_{i=1}^{m} \psi_2(\widetilde{\beta}_{0i}, \widetilde{\beta}_1) & \psi_1(\widetilde{\beta}_{01}, \widetilde{\beta}_1) & \psi_1(\widetilde{\beta}_{02}, \widetilde{\beta}_1) & \cdots & \psi_1(\widetilde{\beta}_{0m}, \widetilde{\beta}_1) \\ \psi_1(\widetilde{\beta}_{01}, \widetilde{\beta}_1) & \psi_0(\widetilde{\beta}_{01}, \widetilde{\beta}_1) & 0 & \cdots & 0 \\ \psi_1(\widetilde{\beta}_{02}, \widetilde{\beta}_1) & 0 & \psi_0(\widetilde{\beta}_{02}, \widetilde{\beta}_1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_1(\widetilde{\beta}_{0m}, \widetilde{\beta}_1) & 0 & \cdots & \cdots & \psi_0(\widetilde{\beta}_{0m}, \widetilde{\beta}_1) \end{bmatrix}.$$

where $\widetilde{\boldsymbol{\beta}}_0 = (\widetilde{\beta}_{01}, \ldots, \widetilde{\beta}_{0m})$. Using the partitioned matrix inversion formula

$$\left\{ \left[I(\widetilde{\beta}_1, \widetilde{\boldsymbol{\beta}}_0) \right]^{-1} \right\}_{11}^{1/2} = \left(n \sum_{i=1}^{m} \psi_2(\widetilde{\beta}_{0i}, \widetilde{\beta}_1) - \frac{\psi_1(\widetilde{\beta}_{0i}, \widetilde{\beta}_1)^2}{\psi_0(\widetilde{\beta}_{0i}, \widetilde{\beta}_1)} \right)^{-1/2} = O((mn)^{-1/2})$$

provided that $\sum_{i=1}^m \psi_2(\widetilde{\beta}_{0i}, \widetilde{\beta}_1) - \psi_1(\widetilde{\beta}_{0i}, \widetilde{\beta}_1)^2/\psi_0(\widetilde{\beta}_{0i}, \widetilde{\beta}_1) \neq 0$. This condition holds provided $\widetilde{\beta}_1$ and $\widetilde{\beta}_{0i}$, $1 \leq i \leq m$ are finite which is true by (A6). Similarly,

$$\begin{aligned} \left\{ \left[I(\widetilde{\beta}_1, \widetilde{\boldsymbol{\beta}}_0) \right]^{-1} \right\}_{i+1,i+1}^{1/2} &= n^{-1/2} \left[\psi_0(\widetilde{\beta}_{0i}, \widetilde{\beta}_1) - \frac{\psi_1(\widetilde{\beta}_{0i},\widetilde{\beta}_1)^2}{\sum_{j \neq i} \psi_2(\widetilde{\beta}_{0j},\widetilde{\beta}_1) - \frac{\psi_1(\widetilde{\beta}_{0j},\widetilde{\beta}_1)^2}{\psi_0(\widetilde{\beta}_{0j},\widetilde{\beta}_1)}} \right]^{-1/2} \\ &= O(n^{-1/2}) \end{aligned}$$

provided $\sum_{j \neq i} \psi_2(\widetilde{\beta}_{0j}, \widetilde{\beta}_1) - \psi_1(\widetilde{\beta}_{0j}, \widetilde{\beta}_1)^2/\psi_0(\widetilde{\beta}_{0j}, \widetilde{\beta}_1) \neq 0$ which holds using similar reasoning. Hence, we can deduce, using standard results, that

$$\widehat{\widetilde{\beta}}_{0i} = \widetilde{\beta}_{0i}^0 + O_p(m^{-1/2}), \;\; 1 \leq i \leq m, \qquad \text{and} \qquad \widehat{\widetilde{\beta}}_1 = \widetilde{\beta}_1^0 + O_p((mn)^{-1/2}) \tag{17}$$

where $\widehat{\boldsymbol{\beta}}_0 = (\widehat{\beta}_{01}, \ldots, \widehat{\beta}_{0m})$, $\boldsymbol{\beta}_0^0 = (\beta_{01}^0, \ldots, \beta_{0m}^0)$ are the true values of $\boldsymbol{\beta}_0^0$ and $\widetilde{\beta}_1^0$ is the true value of $\widetilde{\beta}_1$.

B.2.2 Heuristics for the Consistency of Gaussian Variational Approximations

We now give a heuristic argument for the consistency of Gaussian variational approximations for simple generalized linear mixed models. Equating (11) and (13) we can deduce

$$\frac{1}{n} \sum_{j=1}^{n} \left[b^{(1)}(\beta_0 + \mu_i + \beta_1 X_{ij}) - b^{(1)}(\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) \right] = O_p(n^{-1}). \tag{18}$$

Now, from the mean value theorem we know that for all z and z' there exists a $z^* \in (z, z')$ such that $(b^{(1)}(z) - b^{(1)}(z'))/(z - z') = b^{(2)}(z^*)$ due to the fact that $b^{(1)}$ is continuously differentiable. Rearranging we have

$$b^{(1)}(z) - b^{(1)}(z') = b^{(2)}(z^*)(z - z').$$

Hence, from (18), we have

$$\frac{1}{n} \sum_{j=1}^{n} b^{(2)}(z^*) \left[(\beta_0 + \mu_i + \beta_1 X_{ij}) - (\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) \right] = O_p(n^{-1})$$

for some $z^* \in (\beta_0 + \mu_i + \beta_1 X_{ij}, \widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij})$. Next, since β_0, β_1, the μ_is, each of the X_{ij}s, $\widetilde{\beta}_{0i}$ and $\widetilde{\beta}_1$ are bounded the value $b^{(2)}(z^*)$ must be bounded away from zero. Hence,

$$\begin{gathered} \frac{1}{n} \sum_{j=1}^{n} \left[(\beta_0 + \mu_i + \beta_1 X_{ij}) - (\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) \right] = O_p(n^{-1}), \; 1 \leq i \leq m, \quad \text{and} \\ \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} \left[(\beta_0 + \mu_i + \beta_1 X_{ij}) - (\widetilde{\beta}_{0i} + \widetilde{\beta}_1 X_{ij}) \right] = O_p(n^{-1}) \end{gathered} \tag{19}$$

due to the fact that the X_{ij}s are bounded. From (19) we have

$$\underline{\widehat{\beta}}_0 + \underline{\widehat{\mu}}_i = \widehat{\widetilde{\beta}}_{0i} + O_p(n^{-1}), \; 1 \leq i \leq m, \quad \text{and} \quad \underline{\widehat{\beta}}_1 = \widehat{\widetilde{\beta}}_1 + O_p(n^{-1}).$$

Hence,

$$\underline{\widehat{\beta}}_0+\underline{\widehat{\mu}}_i=\beta_0^0+U_i+O_p(m^{-1/2}+n^{-1}),\ 1\le i\le m,\quad \text{and}\quad \underline{\widehat{\beta}}_1=\beta_1^0+O_p((mn)^{-1/2}+n^{-1}). \tag{20}$$

Then averaging over the m equations on the left of (20) and using the fact that $\sum_{i=1}^m \underline{\widehat{\mu}}_i=0$ we have

$$\underline{\widehat{\beta}}_0=\beta_0^0+O_p(m^{-1/2}+n^{-1}).$$

Hence, using the above with (20) can deduce that $U_i=\underline{\widehat{\mu}}_i+O_p(n^{-1/2})$ and finally

$$\underline{\widehat{\sigma}}^2=\sum_{i=1}^m\frac{\underline{\widehat{\mu}}_i^2}{m}+O_p(n^{-1})=\sum_{i=1}^m\frac{U_i^2}{m}+O_p(n^{-1})=(\sigma^2)^0+O_p(m^{-1/2}+n^{-1}).$$

The rates of convergence for $\underline{\widehat{\beta}}_0$ and $\underline{\widehat{\sigma}}^2$ are the same as those in Hall *et al.* (2011) for the Poisson case. However, the rates of convergence for $\underline{\widehat{\beta}}_1$ are slightly better. This agrees with the rates of convergence obtained by Hall, Pham, Wand & Wang (2011).

Additional References

Abramowitz, M. & Stegun, I. (1972). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications.

Hall P., Pham T., Wand, M.P. & Wang, S.S.J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. Unpublished manuscript.

Horn, R.A. & Johnson, C.R. (1985). *Matrix Analysis*. Cambridge, UK: Cambridge University Press.

Kakade, S.M., Shamir, O., Sridharan, K. & Tewari, A. (2010). Learning exponential families in high-dimensions: strong convexity and sparsity. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, volume 9 of JMLR Workshop and Conference Proceedings*, 381–388.

Magnus, J.R. & Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Chichester: John Wiley & Sons.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics*, **16**, 356–366.

Smyth, G. (2009). `statmod 1.4.0` Statistical modeling. R package.
`http://cran.r-project.org`.